

Towards Calibrated, Sharp, and Interpretable Probabilistic Prediction

by

Zeyu Sun

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical and Computer Engineering)
in the University of Michigan
2024

Doctoral Committee:

Professor Alfred O. Hero III, Chair
Assistant Professor Yang Chen
Assistant Professor Qing Qu
Professor Clayton Scott

Zeyu Sun

zeyusun@umich.edu

ORCID iD: 0000-0003-3903-928X

© Zeyu Sun 2024

DEDICATION

To my parents.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Dr. Alfred Hero. I feel incredibly fortunate to have had him as my mentor. Without his constant encouragement, valuable guidance, and enduring patience, this dissertation would not have been possible.

I am also deeply thankful to the members of my thesis committee, Dr. Yang Chen, Dr. Qing Qu, and Dr. Clayton Scott, for their invaluable guidance throughout my Ph.D. studies and for their insightful feedback during the completion of this dissertation.

During my time at the University of Michigan, I had the privilege of working with many outstanding researchers in the Hero Group: Morteza Nosahd, Benoit Dufumier, Mayank Baranwal, Rohit Gupta, Mehmet Aktukmak, Yu Wang, Dogyoon Song, Byoungwook Jang, Yun Wei, Bob Malinas, Haonan Zhu, and Neophytos Charalambides. This journey has been all the more rewarding thanks to their support and collaboration throughout my studies.

I would also like to thank the many friends I've made along the way; their companionship has made my life more colorful and memorable.

Finally, I am forever grateful to my parents for their unconditional love and unwavering support. They always believed in me, even during times when I struggled to believe in myself.

TABLE OF CONTENTS

DEDICATION	ii
ACKNOWLEDGEMENTS	iii
LIST OF FIGURES	viii
LIST OF TABLES	xiii
ABSTRACT	xvii
CHAPTER	
1 Introduction	1
1.1 Probabilistic Forecasting	1
1.1.1 Forecast Verification	1
1.1.2 Data-Driven Probabilistic Forecast Pipeline	2
1.2 Solar Flare Forecasting	3
1.3 Outlines and Contributions	4
2 Solar Flare Prediction using Deep Learning Ensemble on Multimodal Data	8
2.1 Introduction	8
2.2 Data	11
2.2.1 Data sources	11
2.2.2 Data fusion	12
2.2.3 Sample extraction and labeling	14
2.2.4 Train/validation/test split	17
2.2.5 Random undersampling	18
2.2.6 Image resizing	19
2.2.7 Standardization	19
2.3 Methodology	20
2.3.1 Deep learning models	20
2.3.2 Stacking ensemble	21
2.3.3 Evaluation tools	25
2.3.4 Statistical performance comparisons	28
2.3.5 Interpretation of CNNs	29
2.4 Results	31
2.4.1 Data from another solar cycle improves prediction	31

2.4.2	LSTM performs better than CNN	36
2.4.3	Stacking LSTM and CNN leads to better prediction	38
2.4.4	CNN identifies the emergence of preflare features	42
2.5	Conclusions and discussion	48
3	Incorporating Polar Field Data in Flare Prediction	53
3.1	Introduction	53
3.2	Data Preparation	57
3.2.1	SPF Preprocessing	57
3.2.2	Observation Preprocessing	59
3.2.3	Train-Validation-Test Splitting	59
3.3	Prediction algorithms	60
3.3.1	Logistic Regression	60
3.3.2	Mixture of Logistic Regressions	61
3.3.3	Multi-layer Perceptron	63
3.3.4	Recurrent Neural Network	63
3.4	Experiments	64
3.4.1	Configurations	64
3.4.2	Tuning	64
3.4.3	Significance Assessment	65
3.4.4	Results	66
3.5	Conclusions	69
4	Multiway Covariance Models	71
4.1	Introduction	71
4.2	Background	72
4.2.1	Notation	72
4.2.2	Multiway Data	73
4.2.3	Multiway covariance representations for diffusion processes	74
4.3	Multiway Covariance Models	76
4.3.1	The Kronecker Product Covariance Model	77
4.3.2	Kronecker Covariance Decomposition and Kronecker PCA	79
4.3.3	Kronecker Graphical Lasso and Tensor Lasso	80
4.3.4	Tensor Graphical Lasso	82
4.3.5	Sylvester Graphical Lasso and SG-PALM	82
4.4	Illustrations and Performance Comparisons	84
4.4.1	Theoretical Guarantees	84
4.4.2	Numerical Experiments with Synthetic Data	85
4.4.3	Application to Predicting Evolution of Solar Active Regions	94
4.5	Conclusion and discussion	97
5	Julia Software for Kronecker Models	99
5.1	Introduction	99
5.2	Impact on covariance model development	100
5.3	Ongoing research and further development	104

6	Recalibration in Binary Classification	106
6.1	Introduction	106
6.1.1	Related work	107
6.1.2	Contributions	108
6.1.3	Organization	109
6.2	Preliminaries	109
6.2.1	Notation	109
6.2.2	Calibration and sharpness	110
6.3	Optimal recalibration	110
6.4	Recalibration via uniform-mass binning	112
6.4.1	Uniform-mass binning algorithm for recalibration	112
6.4.2	Theoretical guarantees	113
6.5	Recalibration under label shift	115
6.5.1	Revisiting the problem formulation	115
6.5.2	Two-stage recalibration under label shift	116
6.6	Numerical experiments	118
6.7	Discussion	121
6.8	Appendix: Proof of Proposition 6.1	123
6.9	Appendix: Proof of Theorem 6.1	123
6.9.1	High-probability certification of the conditions	124
6.9.2	Conditional upper bounds on reliability risk and sharpness risk	126
6.9.3	Completing the proof of Theorem 6.1	130
6.10	Appendix: Proof of Theorem 6.2	131
6.10.1	Useful lemmas	131
6.10.2	Completing the proof of Theorem 6.2	134
6.11	Appendix: Proof of the argument in Remark 6.6	137
6.12	Appendix: Details on the experiments	138
6.12.1	Verifying results for UMB	139
6.12.2	Comparing recalibration methods	139
6.12.3	Comparing recalibration schemes under label shift	142
7	Solar Flare Prediction under Event Rate Shift	144
7.1	Introduction	144
7.2	Methods	147
7.2.1	Mathematical setup	147
7.2.2	Forecast verification methods	148
7.2.3	Recalibration under event rate shift	153
7.3	Experiments	155
7.3.1	Data preparation	155
7.3.2	Experiment setup	157
7.3.3	Recalibration performance on Split 2	158
7.3.4	Recalibration performance on Split 1	162
7.3.5	Case study: Logistic regression	165
7.4	Conclusions and discussion	168

7.5	Appendix	168
7.5.1	Classification algorithms	168
7.5.2	Optimal probability thresholds	171
7.5.3	Proofs	175
8	Conclusion	177
8.1	Summary	177
8.2	Future Work	178
	BIBLIOGRAPHY	181

LIST OF FIGURES

FIGURE

2.1	Q-Q (quantile-quantile) plot of 50 matched magnetogram pairs of HARP and TARP from May 1 to October 28 in 2010. Active regions with pixels outside of $\pm 70^\circ$ from the central meridian are not used. For each pair, the co-temporal magnetograms are sampled at a rate of every 8 hours. The pixels within the intersection of the bounding boxes of active region pairs are used. Lighter colors indicate higher latitudes.	13
2.2	2D histograms of summary parameters <code>USFLUXL</code> , <code>MEANGL</code> , <code>R.VALUE</code> , and <code>AREA</code> between SHARP and SMARP. SHARP summary parameters are suffixed with <code>_HMI</code> and SMARP with <code>_MDI</code> . The orange lines in the diagonal blocks are the least square fit of SMARP summary parameters on the corresponding SHARP summary parameters, with coefficient k , intercept b , and Pearson correlation coefficient r displayed in the corner.	14
2.3	Demonstration of the sample extraction and labeling procedure of an active region. The dark orange dots represent flares that occurred in an active region, with the last flare exceeding the M1.0 threshold. The blue sample is labeled as the negative class because no flare of any class occurs in the observation and the prediction period. The gray sample is irrelevant to the task since all flares in the prediction period are weaker than M1.0. The orange sample is labeled as the positive class because the prediction period contains a flare of size exceeding M1.0.	15
2.4	Neural network architectures. (a) shows the LSTM architecture. (b) shows the CNN architecture.	22
2.5	An illustration of the relation between the reliability diagram and BSS.	28
2.6	Verification plots on SHARP test data to compare models trained on <code>FUSED_SHARP</code> and <code>SHARP_ONLY</code> . Shown in (a1)–(a3) are the reliability diagram, ROC, and SSP for LSTM. Shown in (b1)–(b3) are the same plots but for CNN. In each panel, the blue/orange curve is the test performance for the model trained on <code>FUSED_SHARP</code> / <code>SHARP_ONLY</code> . In each graph, solid curves and error bars (or shaded area) indicate respectively the means and the standard deviations calculated from 10 random experiments. In each reliability plot, the short horizontal bars indicate the number of samples in each probability bin, and the two curves are separated horizontally to prevent error bars from overlapping.	34
2.7	Same as Figure 2.6 but for SMARP test data to compare models trained on <code>FUSED_SMARP</code> and <code>SMARP_ONLY</code>	35

2.8	Verification plots of the LSTM and the CNN on FUSED_SHARP . Shown are the reliability diagram, ROC, and SSP, from left to right. This figure essentially extracts the blue curves (representing FUSED_SHARP) in both rows of Figure 2.6 and overlaps them together.	37
2.9	Test set TSS for base learners and meta-learners using different criteria.	39
2.10	Stacking weight α fitted using different criteria on different datasets. All 10 values of α in an experiment setting are shown as points next to the corresponding box.	40
2.11	(a1)–(a3): CNN predicted probability (y-axis) vs. LSTM predicted probability (x-axis) for the train, the validation, and the test set. The green solid line in (a2) and (a3) is the decision boundary of the ensemble with meta-learner fitted on the validation set. The violet dashed line in (a3) is analogous to the green line except that it is fitted on the test set, and hence can be seen as the oracle. (b): ACC as a function of α on the validation and the test set. The vertical green line shows the value of α that maximizes the validation ACC. The leftmost values of the ACC curves ($\alpha = 0$) correspond to the ACC of the CNN, and the rightmost values of these curves ($\alpha = 1$) correspond to the ACC of the LSTM.	41
2.12	Examples of synthetic bipole images and attribution maps.	43
2.13	CNN predictions of part of time series of in HARP 384, 1806, 5107, and 5982. The labels are shown as blue open boxes and predicted probabilities as green plus symbols. The point-in-time instance is labeled as positive if an M1.0+ flare occurred in the future 24 hours in that active region. GOES flare events during and 24 hours within the sample sequence are shown as short vertical bars, with y-coordinates indicating flare intensities (peak flux in W/m^2) on a log scale.	44
2.14	Attribution results of Deconvolution, Guided Backpropagation, DeepLIFT, and Grad-CAM on the last magnetogram in the sample sequences of HARP 384, 1806, 5107, and 5982. DeepLIFT chooses the first sample in the sequence as the reference. “LayerGradCam-4” means Grad-CAM with respect to the output of the fourth, or the second to last, convolutional layer. The interactive movie of heatmaps on all 9 samples in HARP 5982 using more attribution methods can be accessed at https://zeyusun.github.io/attribution/captum_movie_first.html	45
2.15	Highly attributed pixels in the last frame by Integrated Gradients on four select HARPs shown in rows. In (a), the left/right panel shows the first/last magnetogram in the sample sequence of HARP 384. The magnetograms are in the SHARP resolution, with ticks on the axes indicating pixels. Pixel values saturate at ± 500 Gs. The red/blue contours on the right panel (last frame) highlight the areas with strong positive/negative Integrated Gradients relative to the first frame. The same contours are mapped to the left panel (first frame) for contrast. The contours are drawn on the attribution map smoothed with a Gaussian kernel with a standard deviation of 3 pixels. Figures in (b), (c), and (d) are similar to (a) but for other HARPs. An animated version of this figure that serially presents the complete sample sequences of the four HARPs is available online.	47

2.16	Line-of-sight magnetic field (a) and solar EUV image (b) of HARP 5982 (NOAA AR 12423) at 23:10:17 on Sep 27, 2015. Images are taken from https://solarmonitor.org/ . Note that the image title of (b) should be “AIA 171 Å” instead of “AIA 174 Å”.	48
3.1	<i>Top</i> : The north and south polar strength and the monthly mean sunspot numbers plotted over 4 solar cycles. The shaded period is the time during which the HMI instrument has been generating data. Polar fields data are from the Wilcox Solar Observatory [273, 128]. Sunspot number data are from WDC-SILSO, Royal Observatory of Belgium, Brussels. <i>Middle</i> : Solar cycle variation of the north polar field strength (CAPN2), AR unsigned flux (USFLUXN), and flux-weighted AR latitude (USFLATN) obtained from HMI data. The unsigned flux in the northern hemisphere (USFLUXN) is positively correlated with the north polar fields (CAPN2), suggesting the inclusion of polar field data as additional features in a machine learning flare prediction model can potentially improve prediction accuracy. <i>Bottom</i> : Similar to the Middle panel but for the Southern Hemisphere of the sun.	56
3.2	Raw (blue curve) and Kalman filtered (red curve) polar field features. Polar field strengths at the north pole CAPN2, south pole CAPS2, and the solar cycle approximation, defined as the unsigned polar field difference $ CAPS2(t) - CAPN2(t) $, are shown in the left, middle, and right-hand panels respectively. The data source switched from MDI to HMI in 2010, which is accompanied by a visually obvious reduction in noise. The HMI data source has lower noise due to its high resolution.	58
3.3	The classification accuracy curves of MOE with and without using polar fields data w.r.t. to the sunspot number when the splitting method is instance-based. Cycle approximation with filtered data is used as input. The curves are moving averaged over 90 days. The incorporation of the data substantially improves classification accuracy in some portions of the solar cycle.	68
4.1	Multiway data and patterned covariance.	74
4.2	Poisson-AR(1) (top) and 2D Convection-diffusion (bottom) state variables of size 64×64 at three different time steps, which are chosen such that t is sufficiently large for the processes to be in steady state and that the sampling times are adjusted to the temporal dynamics of each process. Here, models parameters used for simulations are: $a = -0.5$ for Poisson-AR, $\theta = 0.05$ and $\epsilon = 0.01$ for convection-diffusion process; the driving noise is generated from white Gaussian with $\sigma_w = 0.1$ in both cases.	89
4.3	Visualizations of the inverse covariance structures for Poisson-AR(1) dynamics and its estimates. Here, white entries indicate zeros in the inverse covariance matrices. The zoomed-in plots show two temporal blocks (each of size 64×64) of spatial inverse correlation structures with the diagonal elements removed for clearer visualization. SG-PALM and the associated Sylvester graphical model produce the richest structures.	90

4.4	Visualizations of the inverse covariance structures for the Convection-Diffusion dynamics and its estimates. Here, white entries indicate zeros in the inverse covariance matrices. The zoomed-in plots show two temporal blocks (each of size 64×64) of spatial inverse correlation structures with the diagonal elements removed for clearer visualization. SG-PALM and the associated Sylvester graphical model produce the richest structures.	91
4.5	Visualizations of the middle 128 rows and columns of the covariance structures for Poisson-AR(1) and Convection-Diffusion dynamics and their estimates, which show two temporal blocks of spatial correlation structures, each of size 64×64 , with the diagonal elements removed for clearer visualization of the pattern. Here, white entries indicate zeros in the covariance matrices. Since the covariances are not sparse in general, all matrices are thresholded for clearer inspections of patterns.	92
4.6	Examples of one-hour ahead prediction of the first two AIA channels of last frames of 13-frame videos, leading to B- (first two rows) and M/X-class (last two rows) flares, produced by the SG-PALM, Tlasso, TeraLasso, IndLasso algorithms, comparing to the ground truth image (far left column). The column ordering from left to right is in the order of increasing normalized mean squared error (NMSE) - see Table 4.5. Note that in general linear forward predictors tend to underestimate the contrast ratio of the images. The proposed SG-PALM produced the best-quality images in terms of both the spatial structures and contrast ratios. This figure is excerpted from Wang and Hero [305] (cf. Figure 3 in the reference).	96
5.1	Visualization of an estimated multiway sparse precision matrix (left) and accuracy of multiway PEnKF algorithm (right), using the proposed software.	102
6.1	Medians (solid lines) and 10-90 percentile ranges (shaded areas) of quadrature estimates of population risks over 10 realizations and theoretical risk upper bounds ($\delta = 0.1$) (dashed lines) for various n and B . (a)-(c) The empirical rates, $R^{\text{cal}} = O(n^{-0.99}B^{0.98})$ and $R^{\text{sha}} = O(B^{-1.83})$, align with theoretically predicted rates, $\tilde{O}(B/n)$ and $O(B^{-2})$, in Thm. 6.1. (d) The empirically observed R and our upper bound exhibit similar trends as a function of B	119
6.2	The optimal number of bins B for different sample sizes n plotted in linear scale (<i>left</i>) and log scale (<i>middle</i>), and the population risk for various combinations of n and B , with the optimal B marked by black dots (<i>right</i>). Note that the risk surface is relatively smooth around its minimum, suggesting the robustness of the optimal B	119
6.3	Optimal recalibration function h^* and recalibration function estimates by Platt Scaling [229], the hybrid method [162], UWB, and UMB when the parametric assumption is (a) correct and (b) misspecified. UMB traces the h^* in both cases, whereas the hybrid method traces Platt scaling, exhibiting an intrinsic bias from h^* in (b)	120
6.4	Stochastic dependence among $P, f, \tilde{S}, S, \mathcal{B}$, and $\hat{h}$	126
6.5	Risks vs. number of bins B	141

6.6	<i>Left:</i> calibration curves of for COMPOSITE $\hat{h}_Q$, SOURCE $\hat{h}_P$, TARGET $\hat{h}_Q^{\text{target}}$, and LABEL-SHIFT $\hat{g}$. <i>Right:</i> the marginal density of $Z = f(X)$ under Q	143
7.1	Brier score decomposition for forecast Z	151
7.2	Monthly event rates vs mean sun spot numbers from May 2010 to May 2022. The sun spot numbers is an indicator of the solar activity and the phase in a solar cycle. The event rate correlates well with the trend of the sunspot number. Source of sunspot number data: WDC-SILSO, Royal Observatory of Belgium, Brussels.	157
7.3	The probability calibration on the test set (Split 2) with event rate $s' = 0.009$. The blue solid (oranged dashed) curve shows the recalibration for classifiers trained on unbalanced (rebalanced) training data where $s = 0.041$ ($s = 1$). . . .	159
7.4	Recalibrated performance of models on Split 2 (2016.5 – 2022.5) with different imbalance treatment in training: no treatment (up), loss is reweighted (middle), and data is rebalanced (bottom).	160
7.5	Skill bias diagram. The mean TSS and mean FB on Split 2 are plotted for each algorithm with unadjusted predictions (shown as open squares) and recalibrated predictions (shown as solid circles). The contours show the HSS on the skill-bias diagram with the Split 2 event rate $s = 0.009$. The left panel corresponds to models trained on the unbalanced data, whereas the right panel corresponds to models trained on the rebalanced data. The FB is displayed in log scale.	161
7.6	The probability calibration on the test set (Split 1) with event rate $s' = 0.041$. The blue solid (oranged dashed) curve shows the recalibration for classifiers trained on unbalanced (rebalanced) training data where $s = 0.009$ ($s = 1$). . . .	162
7.7	Recalibrated performance of models on Split 1 (2010.5 – 2016.5) with different imbalance treatment in training: no treatment (up), loss is reweighted (middle), and data is rebalanced (bottom).	163
7.8	Skill bias diagram similar to Figure 7.5, except that models are trained on Split 2 and tested on Split 1.	164
7.9	Graphical tools for a logistic regression model trained with rebalanced data and tested on Split 2. The first row shows skill score profiles (SSP), reliability diagram (RD), ROC, and Skill-bias diagrams for the raw predicted probabilities. The second row shows the same figures for the recalibrated probabilities. The red cross on the ROC and Skill-bias corresponds to the 0.5 threshold.	167
7.10	The factor α as a function of forecast rate r when the base rate $s = 0.15$. When $r = s$, the forecast frequency is unbiased and $\text{HSS} = \text{TSS}$. When $r > s$, there is a tendency to overforecast and $\text{HSS} < \text{TSS}$. When $r < s$, there is a tendency to underforecast and $\text{HSS} > \text{TSS}$	174

LIST OF TABLES

TABLE

1.1	Our contributions to the probabilistic forecasting pipeline. The roman numerals highlighted with different background colors (I , II , and III) before a topic indicates the part of the thesis it belongs to.	3
2.1	Active region summary parameters used in this study. The line-of-sight magnetic field is denoted by B_L	12
2.2	<i>Up</i> : Samples sizes of all possible flare activity evolution types, with missing/inconsistent data removed. The flare activity in the observation period of a sample can be QUIET (the active region is flare-quiet), WEAK (only flares of size smaller than M1.0 occur), or STRONG (there is at least one large flare of size M1.0 or above). The prediction period can be classified the same way. The entries denote the sample counts in SMARP/SHARP data sets. <i>Down</i> : The associations of flare activity evolution types and the class labels for the “strong-vs-quiet” flare prediction task. Positive samples are denoted as + and negative samples as -. Samples with the evolution type signifying a decaying flare activity (denoted as a blank space) or leading to only small flares (denoted as ?) are not relevant to our task.	16
2.3	Sample sequences extracted from SMARP and SHARP	17
2.4	A contingency table consisting of TP (true positive), FP (false positive), FN (true negative), and TN (true negative).	25
2.5	Sample sizes of a random realization of the four datasets	32
2.6	Test set performance of the LSTM and the CNN on 24-hour “strong-vs-quiet” flare prediction. The two datasets within each comparison group share common test sets. The $1-\sigma$ error is calculated from 10 random experiments. Bold fonts indicate the experiments in which the mean of the metric on the fused dataset is higher than that on the single dataset within the group. Note that comparisons should not be made between groups because the metrics are evaluated on different test sets.	33
2.7	Paired t -tests for significant improvement of test set performance on the fused datasets as measured by different metrics. The alternative hypothesis H_1 claims that metric S on the fused dataset (FUSED_SHARP or FUSED_SMARP) is greater than the respective single dataset (SHARP_ONLY or SMARP_ONLY), which is tested against the null hypothesis H_0 claiming otherwise. The bold font p -values are less than 0.05 and considered to be significant.	35

2.8	Paired t -tests for significant improvement of the LSTM over the CNN in terms of different metrics S on the test set of the four datasets. The alternative hypothesis H_1 claims $S_{\text{LSTM}} > S_{\text{CNN}}$. The bold font p -values are less than 0.01 and considered to be significant.	37
3.1	SHARP dataset statistics after preprocessing step. See Table 3 in the appendix for the definitions of the keywords.	58
3.2	Table of statistically significant (p-value $\leq 5\%$) performance improvements due to adding polar field features to SHARP features, The hss2 skill score argument (p) means only polar fields data, (s) means only SHARP parameters, and (s+p) means both data sources are used as input to the algorithms. (diff) is the mean hss2 score percentage difference when the polar field is incorporated in addition to SHARP parameters. The highlight colors in the table specify the common configurations where MOE achieves statistically significant performance improvements due to adding the polar field. The full tables including the statistically non-significant configurations are provided in the supplementary.	66
3.3	List of SHARP Parameters and Brief Descriptions	70
4.1	Overview of the models and the algorithms reviewed in Section 4.3. Here, Σ and Ω denote the covariance and the inverse covariance (precision), respectively. For the sake of brevity, the model in the second column is described for $K = 2$	77
4.2	Comparison of known theoretical guarantees on sample complexity (statistical error) and computational complexity. Here, $M = \max\{d_1, d_2, N\}$, $m_k = \prod_{i \neq k} d_i$ is the co-dimension of the k -th mode, $d = \prod_{k=1}^K d_i$, and s_k characterizes the sparsity of each of the inverse covariance Kronecker factors $s_k = \{(i, j) : i \neq j, [\Omega_k]_{i,j} \neq 0\} $, s is the sparsity of the full inverse covariance $s = \{(i, j) : i \neq j, \Omega_{i,j} \neq 0\} $ and $s = \sum_{k=1}^K m_k s_k$ if Ω satisfies the Kronecker sum model.	86
4.3	Comparisons of performances measured by $\log\left(\ \hat{\Sigma} - \Sigma\ _F \backslash \ \Sigma\ _F\right)$ for KPCA as well as $\log\left(\ \hat{\Omega} - \Omega\ _F \backslash \ \Omega\ _F\right)$ and the Mathews Correlation Coefficient (MCC) for SG-PALM, Tlasso, TeraLasso, Glasso. The MCC is a measure of the quality of sparsity recovery considered as a binary classification problem, where ± 1 indicates perfect agreement or disagreement between the truth and the estimation. Here the Frobenius norm errors are included in the first row under each generating type while the MCCs are in the second row. Note that the best performers under each type/criteria are highlighted.	93
4.4	Average runtime and its standard error (in seconds) of estimating spatio-temporal (inverse) covariance matrices of size $d \times 50$, where d is varying, using various algorithms. Comparisons under various problem sizes (i.e., different d and N) are shown. Note the sparse multiway precision models (SG-PALM, KGlasso, TeraLasso) are comparably fast and are all faster than Glasso (for large problems) and KPCA.	93

4.5	Comparison of the SG-PALM, Tlasso, TeraLasso, IndLasso performances measured by NRMSE in predicting the last frame of 13-frame video sequences leading to B- and M/X-class solar flares. The NRMSEs are computed by averaging across testing samples and AIA channels for each pixel. SG-PALM achieves the best overall NRMSEs across pixels. B flares are generally easier to predict due to both a larger number of samples in the training set and typically smoother transitions from frame to frame within a video.	96
6.1	90%-quantiles of the risks of Platt scaling [229], the hybrid method [162], uniform-width binning (UWB) [109], and uniform-mass binning (UMB) over 100 random calibration datasets drawn from $Z \sim \text{Uniform}[0, 1]$, and (a) $Y \mid Z \sim \text{Bernoulli}(h_{\text{beta}}(z; 4, 4, 0))$, or (b) $Y \mid Z \sim \text{Bernoulli}(h_{\text{beta}}(z; 0.1, 4, 0))$. While Platt scaling and the hybrid method achieve lower R under the correct parametric assumption, UWB and UMB may outperform when the parametric assumption fails.	120
6.2	Risks under label shift from $\mathcal{D}(0.5)$ to $\mathcal{D}(0.1)$, with $n_P = 10^3$ and $n_Q = 10^2$. Standard deviations are computed from 10 random realizations. LABEL-SHIFT, only applying an injective $\hat{g}$, achieves $R^{\text{sha}} = 0$ but incurs high R^{cal} . SOURCE, recalibrated with $B = n_P^{1/3}$ on $\mathcal{D}_P$, incurs high R^{cal} . TARGET, recalibrated with $B = n_Q^{1/3}$ on $\mathcal{D}_Q$, incurs low R^{cal} . Our proposed COMPOSITE achieves the lowest nonzero R^{sha} and lowest R^{cal}	121
7.1	Probabilistic metrics. Suppose N probability forecasts are binned into B groups. The sample size, center predicted probability, and the observed frequency of the i -th bin are denoted as N , $\langle f_i \rangle$, and $\langle o_i \rangle$, respectively. The average observed frequency (i.e., climatology) is computed as $\bar{o} = (\sum_{i=1}^B N_i \langle o_i \rangle) / N$	151
7.2	A contingency table of $n + p$ forecast-observation pairs categorized into TP (true positive), FP (false positive), FN (false negative), and TN (true negative). The number of in each category is denoted as a, b, c, d , respectively.	153
7.3	SHARP summary parameters that are used as active region features in this chapter. See Bobra and Couvidat [30] for details.	156
7.4	Statistics of the two periods.	157
7.5	Mean adjusted (Adj) performance on Split 2 and p -values from one-tailed paired t -tests for significant improvement over the uncalibrated models. For each imbalance treatment and each machine learning algorithm, 5 models are independently tuned, trained and tested, producing 5 metric values computed from the uncalibrated predicted probabilities, and 5 metric values computed from the recalibrated probabilities. A one-tailed paired t -test is performed to test if the recalibrated metric is significantly larger than the raw metric. p -value smaller than 0.05 is treated as significant. p -values smaller than 0.001 is reported as 0.001.	159
7.6	Similar to Table 7.5, except that the models are trained on Split 2 and tested on Split 1.	164
7.7	Logistic regression coefficients (include the intercept) ordered by absolute value from top to bottom and then from left to right.	166

7.8	Metrics of the selected Logistic regression model. The bold fonts shows the value is better (closer to 1 for FB and higher for other metrics).	166
-----	--	-----

ABSTRACT

Predicting the future has long been a fundamental aspiration of humankind. In the current era of big data, the capacity to collect vast amounts of information presents a substantial challenge in effectively leveraging this data, given its increasing diversity, dynamism, and scale. In addition, data-driven methods, especially the evermore powerful deep learning models, are lacking in terms of interpretability and uncertainty quantification.

This dissertation focuses on the theory and the practice that improve the calibration, sharpness, and interpretability in probabilistic forecasting, with a particular focus on applications in solar flare forecasting and active region evolution prediction. In terms of data, we devise methods that fuse historical data, combine multimodal data, and construct physically meaningful features, to unleash the potential of deep neural networks in solar flare forecasting. In terms of models, we apply modern methods to estimate the high-dimensional multiway covariance, and use them to predict the evolution of the high-dimensional multi-channel spatio-temporal process observed in solar active regions. In terms of post-processing, we develop statistical theories and efficient algorithms in binary recalibration problems with potential distribution shifts, and apply them in solar flare prediction regulated by solar cycles. In terms of post-hoc interpretation, we develop attribution tools for astrophysicists to identify the precursors of solar flares discovered by convolutional neural networks.

With these theoretical and practical advances, we provide refreshing understanding of the statistical theory that underpins the general probabilistic prediction, and practical predictive methodology and techniques that utilize massive and complex data.

CHAPTER 1

Introduction

1.1 Probabilistic Forecasting

Forecasting, i.e., predicting, the future is one thing that we humans have been trying to do since the earliest times. In some simple cases, the future of a system is fully determined, like the movement of an ideal pendulum in a vacuum. In other cases, the future is random or chaotic by nature, like the outcome of tossing a die, or the weather. Regardless of the existence of deterministic or stochastic mechanisms, in our attempt to predictively model any system stochastic measurement errors are unavoidable. Therefore, forecasting problems are inherently probabilistic. Thus, the events to be predicted are best modeled as random variables drawn from a (possibility unknown) distribution and the prediction procedure is a function mapping past observables to the event space.

1.1.1 Forecast Verification

So what is considered as a good prediction procedure? This question has inspired forecast verification, the process of assessing the quality of forecasts. Forecast verification has been developed in meteorology [320], finance and economics [15, 54], medical diagnosis [223], and statistics [97, 96]. There have not been a gold standard for what constitutes a good forecasts, but some of the desired attributes for prediction procedures are as follows [320]:

- Accuracy: the average correctness of between individual forecasts relative to the events they predict.
- Robustness: the ability of a forecast procedure to consistently achieve high accuracy over different conditions, e.g., distribution shifts.
- Calibration: the property that the true occurrence rate of events is equal to the predicted rate of events.

- Sharpness (or refinement): the variance of the predictor function about its mean.

Another important attribute of forecasts is interpretability. In the era of artificial intelligence (AI), the forecasts and decisions are often made by machines based on “black box” mechanisms. Although these AI methods exceed human performances in many applications, the decisions are not always interpretable. However, interpretability is important because: (1) scientists want to know why the machine made a particular decision to enhance scientific understanding or discovery, (2) without an explanation of the decision and understanding, it is difficult to predict how well the AI black box will generalize and how sensitive it may be to deliberate adversarial attack, and (3) interpretability of the AI black box is crucial for improving AI technology.

For deep learning methods such as neural networks, the most successful and widely-used method to interpret or explain the machine’s decisions is to use attribution methods. Attribution methods reveal the role of particular features or training instances in determining the the predictions.

Increasingly over time, the focus in forecasting has shifted to modeling events as a spatio-temporal process, as described by [94]:

“A pressing need is to go beyond the univariate, real-valued case, which we review, to the multivariate case, such as in temporal, spatial, and spatiotemporal scenarios and trajectories.”

In this thesis, we consider a particular spatio-temporal prediction problem: solar active region evolution.

1.1.2 Data-Driven Probabilistic Forecast Pipeline

Machine learning systems are applied to many different applications. The training and evaluating of machine learning models can be summarized as the following pipeline:

Data $\rightarrow$ Model $\rightarrow$ Post-processing $\rightarrow$ Evaluation

The contribution of this thesis to this probabilistic forecasting pipeline is summarized in Table 1.1. These contributions are organized into 7 chapters, divided into three thematic areas of parts:

- Part **I** contains Chapter 2 and 3, where the focus is on machine learning strategies for solar flare predictions, addressing challenges in multisource and multimodal learning using neural networks, stacking ensemble, and attribution methods.

	Data	Model	Post-processing	Deployment
Calibration	III Calibration data	III Uniform-mass binning	III Recalibration	III Label Shift Correction
Sharpness	I Multisource multimodal data	I Neural networks	I Stacking	I Skill scores
Interpretability		II Kronecker covariance models	I Attribution methods	

Table 1.1: Our contributions to the probabilistic forecasting pipeline. The roman numerals highlighted with different background colors (**I**, **II**, and **III**) before a topic indicates the part of the thesis it belongs to.

- Part **II** contains Chapter 4 and 5, where the focus is on scalable and interpretable modeling of multiway high-dimensional spatio-temporal processes using Kronecker covariance models.
- Part **III** contains Chapter 6 and 7, where the focus is on post-hoc recalibration of probabilistic predictions produced by machine learning models.

1.2 Solar Flare Forecasting

The central application domain of this thesis is probabilistic forecasting of solar flares. Solar flares are explosions in localized magnetic active regions on the sun, featuring a sudden brightening across the electromagnetic spectrum. They are powerful enough to affect terrestrial satellite communications, power grids, and cause aurora. Strong solar flares can have significant terrestrial impacts, including interfering with radio communications, causing satellite damage, disrupting power grids, and creating auroras in lower latitudes. Solar active regions involve complex magnetic configurations that contain precursors to the solar flare. These precursors can be useful for solar flare prediction. Physical modeling of the magnetohydrodynamics in solar active region has been a long-term effort [244, 51]. However, physical modeling is not sufficient to accurately predict solar flares. In this thesis, we explore statistical methods to model solar active regions, and data-driven methods to predict solar flares.

Solar Active Regions Modeling On the Sun, there are some regions that are more magnetically active than others. Many powerful solar events occur in these regions. Thus, modeling the evolution of solar active regions is of interest. Physical modeling of the mag-

netohydrodynamics for solar active regions has been of interest for many years [244, 51], as have statistical and data-driven modeling approaches [134, 118, 341].

Data-Driven Solar Flare Forecasting Although the underlying physical mechanism of solar flares remains poorly understood, significant progress has been made over the past decade, largely due to the unprecedented abundance of high-quality data collected by several newly-launched space observatories. Many machine learning approaches including several deep learning models have been proposed utilizing physics-based or machine-learned features extracted from image data, but few have dealt with the full time series of AR images. However, the temporal evolution of active regions undoubtedly contains information for flare production.

1.3 Outlines and Contributions

This dissertation studies theoretical underpinning and practical techniques in probabilistic forecasting, with a focus on improving calibration, sharpness, and interpretability. Each chapter in the thesis provides a comprehensive and self-contained exposition of the topic pursued. Consequently, certain chapters may share similar introductory material to appropriately set the context for their specific focus. Below we list the high level goals and contributions of each chapter.

Chapter 2 explores utilising data from multiple sources and merging predictions from different modalities of the data. Specifically, we consider the flare prediction problem that distinguishes flare-imminent active regions, which produce an M- or X-class flare in the future 24 hours, from quiet active regions, which do not produce any flare within ± 24 hours. Using line-of-sight magnetograms and parameters of active regions in two data products covering Solar Cycle 23 and 24, we train and evaluate two deep learning algorithms—convolutional neural network (CNN) and long-short term memory (LSTM) network—and their stacking ensembles. The decisions of the CNN are explained using visual attribution methods. The chapter presents the following three main findings. (1) Training the LSTM on data from two solar cycles achieves statistically significantly higher True Skill Scores (TSS) as compared to training the LSTM on data from a single solar cycle. (2) On data from Solar Cycle 23, a stacking ensemble that combines predictions from LSTM and CNN using the TSS criterion achieves statistically significantly higher TSS than the best individual model. (3) A visual attribution method called Integrated Gradients [272] is able to attribute the CNN’s predictions of flares to the emerging magnetic flux in the active region. The Integrated Gradients attribution also reveals a limitation of CNN as a flare prediction method using line-of-sight

magnetograms: it treats the polarity artifact of line-of-sight magnetograms as positive evidence of flares. This chapter is based on [270] which was published in the *Astrophysical Journal*.

Chapter 3 exploits a novel polar field feature to improve solar flare prediction. We show that incorporating data associated with the sun’s north and south polar field strengths can improve solar flare prediction performance, under a variety of machine learning models. When used to supplement local data from active regions on the photospheric magnetic field of the sun, the polar field data provides global information to the predictor. While such global features have been previously proposed for predicting the next solar cycle’s intensity, in this chapter we propose using them to help classify individual solar flares. We conduct experiments using Helioseismic and Magnetic Imager (HMI) data employing four different machine learning algorithms that can exploit polar field information. Additionally, we propose a novel probabilistic mixture of experts model that can simply and effectively incorporate polar field data and provide on-par prediction performance with state-of-the-art solar flare prediction algorithms such as the Recurrent Neural Network (RNN). Our experimental results indicate the usefulness of the polar field data for solar flare prediction, which can improve Heidke Skill Score (HSS2) by as much as 10.1%. This chapter is based on [7] which was published in *Frontiers in Astronomy and Space Sciences*. My contributions in this chapter include constructing the polar field features (Equation (3.1) and (3.2)), revising the paper, and creating and interpreting Figure 3.1.

Chapter 4 explores the general class of Kronecker-structured covariance models for multiway data. Many applications produce multiway data of exceedingly high dimension. Modeling such multi-way data is important in multichannel signal and video processing where sensors produce multi-indexed data, e.g. over spatial, frequency, and temporal dimensions. We address the challenges of covariance representation of multiway data and review some of the progress in statistical modeling of multiway covariance over the past two decades, focusing on tensor-valued covariance models and their inference. Several new insights are developed for applying Kronecker models to physical processes governed by partial differential equations. We also illustrate the application of Kronecker-structured covariance models in predicting the evolution of solar active regions over time. This chapter is based on [310] which was published in the *Statistical Surveys*. My contributions in this chapter include surveying the Kronecker product covariance models, running experiments, and drafting and revising the paper. In particular, my focus was on surveying, drafting, and revising Kronecker covariance decomposition, Kronecker PCA, and Tensor graphical Lasso; revising the introduction of spatio-temporal dynamics including the Poisson-AR(1) process and the convection-diffusion process; drafting and revising applications to predicting evolution of solar active regions and

future works on Kronecker-structured autoencoders.

Chapter 5 describes an implementation of Kronecker-structured covariance models in the Julia programming language. The increasing availability of high dimensional multi-indexed data has created an opportunity for the development of graphical modeling software that bridges the gap between second order statistical and computational models. In this chapter, we introduce **TensorGraphicalModels**, a suite of Julia tools for statistical estimation of high-dimensional multiway (tensor-variate) covariance and precision matrices, with applications to ensemble Kalman filtering. The Julia package implements several state-of-the-art multiway covariance and precision matrix estimators. These are illustrated in the context of a physics-driven forecasting example with an integrated implementation of a tensor ensemble Kalman filter (EnKF). This chapter is based on [309] which was published in *Software Impacts*. My contributions in this chapter include constructing the solar flare datasets, implementing the Julia code for TeraLasso, drafting the ongoing research on the inference of synchronic global maps of the solar photospheric magnetic field.

Chapter 6 explores the theory of optimal probabilistic recalibration in binary classification. Recalibrating probabilistic classifiers is vital for enhancing the reliability and accuracy of predictive models. Despite the development of numerous recalibration algorithms, until our work there has been a lack of comprehensive theory that integrates calibration and sharpness (which is essential for maintaining predictive power). In this chapter, we introduce the concept of minimum-risk recalibration within the framework of mean-squared-error (MSE) decomposition, offering a principled approach for evaluating and recalibrating probabilistic classifiers. Using this framework, we analyze the uniform-mass binning (UMB) recalibration method and establish a finite-sample risk upper bound of order $\tilde{O}(B/n+1/B^2)$ where B is the number of bins and n is the sample size. By balancing calibration and sharpness, we further determine that the optimal number of bins for UMB scales with $n^{1/3}$, resulting in a risk bound of approximately $O(n^{-2/3})$. Additionally, we tackle the challenge of label shift by proposing a two-stage approach that adjusts the recalibration function using limited labeled data from the target domain. Our results show that transferring a calibrated classifier requires significantly fewer target samples compared to recalibrating from scratch. We validate our theoretical findings through numerical simulations, which confirm the tightness of the proposed bounds, the optimal number of bins, and the effectiveness of label shift adaptation. This chapter is based on [271] which appears in the proceedings of the *Thirty-seventh Conference on Neural Information Processing Systems*.

Chapter 7 applies the optimal recalibration and label shift theory to probabilistic solar flare forecasting. Solar flare forecasting can be prone to prediction bias, producing poorly-calibrated and unreliable predictions. A major cause of prediction bias is the varying event

rates between the training and the testing phase, which either comes from data collection (e.g., collecting the training set and the test set from different phases in a solar cycle) or data processing (e.g., resampling the training data to tackle the class imbalance). Much of the research on machine learning methods for solar flare forecasting do not directly address the prediction bias caused by shifting event rates. In this chapter, we propose a simple yet effective calibration technique when the event rate in the test set is known. We evaluated our calibration technique on various machine learning methods (logistic regression, quadratic discriminant analysis, decision trees, gradient boosting trees, and multilayer perceptrons) that are trained and evaluated on Space-Weather HMI Active Region Patches (SHARPs) from 2010 to 2022, roughly covering Solar Cycle 24. The experimental results show that, under the event rate shift, the proposed calibration generally improves the reliability, discrimination, and various skills as measured by Brier Skill Scores (BSS) and Heidke Skill Scores (HSS). While the proposed calibration decreases True Skill Scores (TSS) which is insensitive to the event rate on the target domain, it preserves the peak TSS maximized by varying the probability threshold. We provide a decision theoretical explanation for the degradation of TSS. Our method is quite general and can be applied to a wide variety of probabilistic flare forecast methods to improve their calibration and discrimination.

CHAPTER 2

Solar Flare Prediction using Deep Learning Ensemble on Multimodal Data

2.1 Introduction

Solar flares are abrupt electromagnetic explosions occurring in magnetically active regions on the solar surface. Intense solar flares are frequently followed by coronal mass ejections and solar energetic particles, which may disturb or disable satellites, terrestrial communication systems, and power grids. Predicting such strong flares from solar observations is therefore of particular significance and has been one of the primary tasks in space weather research.

Flare prediction can be posed as a classification problem, asking for a binary decision on whether the sun will produce a flare above some level in a future time window. Since strong solar flares mostly occur in active regions, it is common to first produce predictions for each active region on the solar disk. In this chapter, we consider a “strong-vs-quiet” flare prediction problem, distinguishing active regions that will produce an M- or X- class flare in the future 24 hours, from those that stay flare quiescent within 24 hours before—and after—the forecast issuance time.

Over the past decade, a great amount of flare prediction studies have been conducted on a data product named Space-Weather HMI Active Region Patches [SHARPs, 31]. The SHARP database is derived from full-disk observations of the Helioseismic and Magnetic Imager [HMI, 251] aboard the *Solar Dynamics Observatory* (SDO), containing maps and summary parameters of automatically tracked active regions from May 2010 to the present day, covering much of Solar Cycle 24. Despite the fact that SHARP is one of the most recent and highest quality datasets of its kind, it only contains a limited number of strong events, as Solar Cycle 24 is the weakest solar cycle in a century. Recently, a new data product, Space-Weather MDI Active Region Patches [SMARPs, 32], was developed as an effort to extend backward the SHARP database to include active region observations in Solar Cycle 23, a much stronger solar cycle with significantly more flaring events. In fact, Solar Cycle

23 is the longest solar cycle (147 months) in the past 150 years¹. The SMARP database was derived from the Michelson Doppler Imager [MDI, 249] aboard the *Solar and Heliospheric Observatory* (SoHO), which observed the sun from 1996 to 2010. Compared to its successor HMI, MDI’s measurement of the solar surface magnetic field is only restricted to the line-of-sight component, with lower spatial resolution, lower signal-to-noise ratio, and shorter cadence. As such, SMARP does not contain as much information as SHARP, and its data quality is not as high. Nonetheless, SMARP’s coverage of a stronger solar cycle and its partial compatibility with SHARP make it a valuable data product to use with SHARP, especially for statistical studies in which a large sample size or a long time span is desired.

Many machine learning methods for flare prediction have been proposed in recent years. They roughly fall into three categories in terms of how flare pertinent features are extracted from data. The first category uses *explicit* parameterization of observational data that are considered relevant to flare production, e.g., SHARP parameters that characterize the photospheric magnetic field. Much of the effort in data-driven flare forecasting has been made in this category, exploring a wide range of machine learning algorithms including discriminant analysis [169], regularized linear regression [145], support vector machine [335, 30, 217, 83], k-nearest neighbors [217], extremely random trees [217], random forests [185, 83], multi-layer perceptrons (MLP) [83], residual networks [218, 219], long short-term memory (LSTM) networks [49, 188], etc. The second category learns features from images using fixed transformations, e.g., random filters [145], Gabor filters [145], wavelet transforms [115]. The third category, only popularized more recently, *implicitly* learns flare indicative signatures directly from active region magnetic field maps. This category features mainly convolutional neural networks (CNNs) [135, 180]. Note that the three categories are not mutually exclusive. For example, methods in the second category typically also depend on explicitly constructed features [e.g. 145] as the information within transformation coefficients is often limited. In this study, two representative deep learning methods, LSTM and CNN, are considered. LSTM uses times series of active region summary parameters derived from line-of-sight magnetograms, whereas CNN uses static point-in-time magnetograms.

With so many machine learning algorithms developed for flare forecasting, one might expect an improved performance by combining different methods. This expectation seems even more reasonable if component methods in the combination use different data to provide complementary information. This is the idea behind ensemble learning, a learning paradigm that capitalizes on different models to achieve better performance than is achievable by any of the models alone. During the past few decades, the rapidly-evolving field of ensemble learning has achieved great success in many areas, which has attracted the attention of the space

¹Source: <https://ntrs.nasa.gov/api/citations/20130013068/downloads/20130013068.pdf>

weather community [213]. In this chapter, using CNN and LSTM models independently trained on active region magnetograms and parameter sequences, respectively, we consider a particular type of ensemble method called stacking [322]. Ideas similar to our stacking ensemble method have previously appeared in solar flare forecasting, most notably Guerra et al. [108, 106]. In their work, full-disk probabilistic forecasts are linearly combined with weights selected by maximizing a potentially non-convex performance metric (e.g., the True Skill Score, the Heidke Skill Score). In contrast, our stacking ensemble linearly combines two state-of-the-art machine learning classifiers (LSTM and CNN). To select the combination weights, we consider a convex cross-entropy loss function in addition to other performance metrics.

Deep learning models are often considered to be “black-box” due to lack of interpretability. Recently the machine learning community has developed empirical tools that aim to better interpret the decisions of the deep neural network (DNN). Among these tools is the class of attribution methods [e.g. 261, 254, 256, 272] that attribute a score to gauge the contribution of each input feature for a given input sample. Attribution methods such as the occlusion method and Grad-CAM have been previously used to interpret CNNs in flare prediction applications [24, 329]. In this work, we evaluate additional attribution methods (Deconvolution, Guided Backpropagation, DeepLIFT, and Integrated Gradients) on the interpretation of CNNs trained to predict flares. In particular, we show that Integrated Gradient attribution maps, which have the same resolution as the input image, lead to insights on the important magnetic features that inform the CNN’s decisions on flare prediction.

The contributions of this chapter are as follows:

1. We demonstrate the value of combining SMARP and SHARP to improve flare prediction performance.
2. We compare the flare prediction performance of LSTM and CNN on an equal footing, i.e., on the same temporally evolving active region dataset.
3. We demonstrate that stacking the LSTM and CNN can significantly improve flare class prediction in certain settings.
4. We provide visual explanations of the CNN predictor using visual attribution methods including Deconvolution, Guided Backpropagation, Integrated Gradients, DeepLIFT, and Grad-CAM. We demonstrate the potential of these methods in identifying flare indicative signatures, interpreting CNN’s decisions, revealing model limitations, and suggesting model modifications.

The rest of the chapter is organized as follows. Section 2.2 introduces the data sources and how they are processed into machine-learning-ready datasets. Section 2.3 describes the flare prediction methods, stacking ensemble, and visual attribution methods. Section 2.4 presents and compares the flare prediction performance on the datasets. Section 2.5 concludes the chapter by presenting the lessons learned from the experiments.

2.2 Data

2.2.1 Data sources

Observational data of active regions of Solar Cycle 23 and 24 are extracted from the SMARP and the SHARP data product, respectively. Both SMARP and SHARP contain automatically-tracked active region cutouts of full-disk line-of-sight magnetograms, referred to as Track Active Region Patches (TARPs) and HMI Active Region Patches (HARPs), respectively. They also contain summary parameters that characterize physical properties of active regions. We consider definitive SMARP and SHARP records in Cylindrical Equal-Area (CEA) coordinates hosted in Joint Science Operations Center². We query SMARP records from 1996 April 23 to 2010 October 28 and SHARP records from 2010 May 1 to 2020 December 1, both at a cadence of 96 minutes. Only good quality SMARP and SHARP records within $\pm 70^\circ$ of the central meridian matching at least one NOAA active region are considered. For active region summary parameters, we use four metadata keywords that are common to SMARP and SHARP, i.e., `USFLUXL`, `MEANGBL`, `R.VALUE`, and `AREA`. Definitions of these summary parameters are listed in Table 2.1. For images, we use photospheric line-of-sight magnetic field maps, or magnetograms, from the two data products.

Observational data samples are labeled using the GOES catalog of X-ray solar flare events. Based on the peak magnitude of 1–8 Å soft X-ray flux measured by *Geostationary Operational Environmental Satellites* (GOES), solar flare events are classified into five increasingly intense classes: A, B, C, M, and X, sometimes appended with a number that indicates a finer scale classification. M- and X- class flares are referred to as strong flares throughout the chapter. We only consider GOES solar flare events with at least one associated NOAA active region that can be used to cross-reference the `NOAA_ARS` keyword in SHARP (or SMARP) databases to associate the flare with a HARP (or TARP). The GOES event records are queried using the SunPy package [284] from the beginning of 1996 to the end of 2020, covering the period of the SMARP and SHARP observations used in this chapter. Of note, although the GOES catalog is widely considered as the “go-to” record database in solar flare forecasting, it is

²See <http://jsoc.stanford.edu>.

Table 2.1: Active region summary parameters used in this study. The line-of-sight magnetic field is denoted by B_L .

Keyword	Description	Pixels	Formula	Unit
USFLUXL	Total line-of-sight unsigned flux	Pixels in the TARP/HARP region	$\sum B_L dA$	Maxwell
MEANGBL	Mean gradient of the line-of-sight field	Pixels in the TARP/HARP region	$\frac{1}{N} \sum \sqrt{\left(\frac{\partial B_L}{\partial x}\right)^2 + \left(\frac{\partial B_L}{\partial y}\right)^2}$	Gauss/pixel
R.VALUE	R , or a measure of the unsigned flux near polarity inversion lines [253]	Pixels near polarity inversion lines	$\log(\sum B_L dA)$	Maxwell
AREA	De-projected area of patch on sphere in micro-hemisphere	Pixels in the TARP/HARP region	$\sum dA$	mH

not error-free. There are cases in which flares, even the strong ones, are not assigned to any active region [173]. Furthermore, small-sized flares could be buried under the background radiation, a phenomenon frequently observed for A- and B-class flares, especially after a strong flare occurs. Moreover, there are 61 event records annotated with an unknown GOES event class, most of them in the year 1996. These 61 event records are excluded in this study.

2.2.2 Data fusion

The challenge of combining the disparate SHARP and SMARP data is mitigated by the fact that there is a short overlapping time period over which they were jointly collected (May 1 to October 28 of 2010). We used this common time period to evaluate the dissimilarities between the SHARP/SMARP data and to develop methods for data alignment. As explained below, our analysis of the data over the common time period led us to adopt a simple method for fusing the SHARP and SMARP data: (1) we downsampled the SHARP magnetograms to match the resolution of the SMARP magnetograms; (2) we separately transformed the SHARP and SMARP summary parameters by Z-score (translation-scale) standardization.

We first discuss the fusion of the SHARP and SMARP magnetograms. SHARP magnetograms inherit the HMI resolution of about $0.5''$ per pixel, whereas SMARP magnetograms inherit the MDI resolution of about $2''$ per pixel. To compare HMI and MDI magnetograms, Liu et al. [189] reduced HMI spatial resolution to match MDI's by convolving a two-dimensional Gaussian function with an FWHM of 4.7 HMI pixels and truncated at 15 HMI pixels. Then, the HMI pixels enclosed in each MDI pixel are averaged to generate an MDI proxy pixel. Subsequently, a pixel value transformation $\text{MDI} = -0.18 + 1.40 \times \text{HMI}$ is applied. In this work, we adopted a simpler approach by subsampling SHARP magnetograms

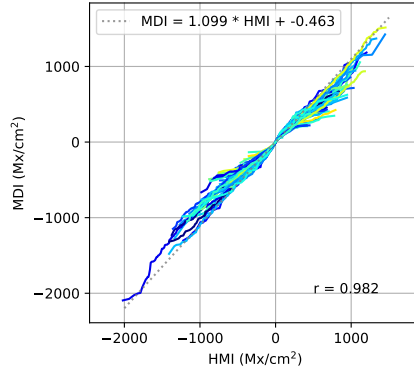


Figure 2.1: Q-Q (quantile-quantile) plot of 50 matched magnetogram pairs of HARP and TARP from May 1 to October 28 in 2010. Active regions with pixels outside of $\pm 70^\circ$ from the central meridian are not used. For each pair, the co-temporal magnetograms are sampled at a rate of every 8 hours. The pixels within the intersection of the bounding boxes of active region pairs are used. Lighter colors indicate higher latitudes.

4 times in both dimensions to match the resolution of SMARP magnetograms. Unlike Liu et al. [189], we approximated pixel value transformation with an identity map, as pixel value distributions of SHARP and SMARP magnetograms in the overlap period are very similar (Figure 2.1). Our approximation agrees well with Riley et al. [241], who found a multiplicative conversion factor of 0.99 ± 0.13 between MDI and HMI using histogram equating. The discrepancy between our multiplicative conversion factor (1.099) and that of Liu et al. [189] (1.40) may be because they considered full-disk magnetograms whereas we focus on active regions. In addition, they considered only 12 pairs in June–August 2010, whereas we considered every possible matching in May–October 2010. Moreover, they performed pixel-to-pixel match of full-disk magnetograms, whereas we use histogram-based methods on active regions because more precise models for aligning coordinates between CEA-projected SHARP and SMARP are not yet available [32]. Furthermore, they considered pixels within 0.866 solar radius of Sun’s center, whereas we considered pixels within $\pm 70^\circ$ from the central meridian.

We next discuss the fusion of SHARP and SMARP summary parameters. Although designed to represent the same physical quantity, summary parameters with identical keywords in SHARP and SMARP are calculated from two pipelines with different source data, and the differences between them cannot be neglected. Bobra et al. [32] investigated these differences by comparing the marginal and the pairwise joint distribution of co-temporal SMARP and SHARP summary parameters for 51 NOAA active regions over the overlap period of MDI and HMI [32, Figure 3]. Motivated by these findings, we investigated the linear associations between SHARP and SMARP using a univariate linear regression analysis. Specifically, SMARP parameters were regressed on their SHARP counterparts. As shown in

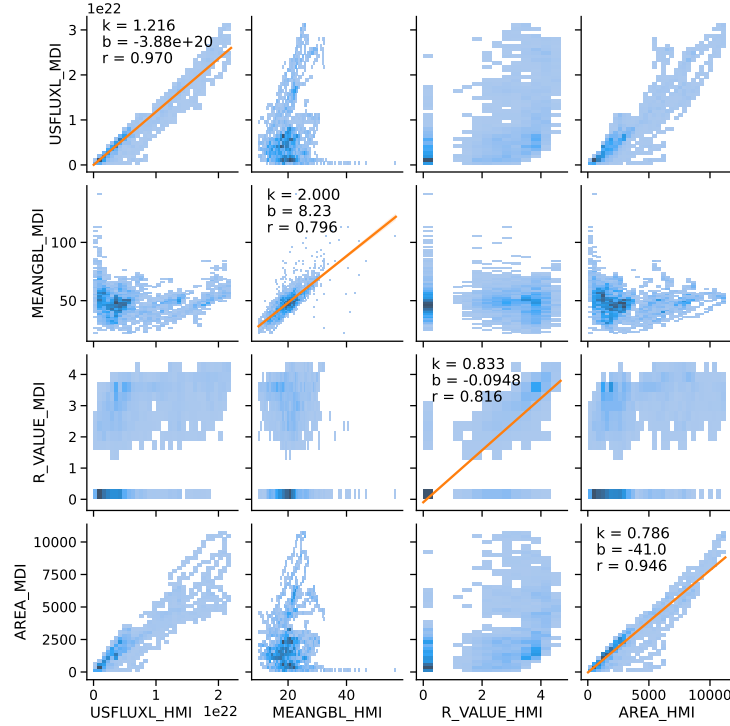


Figure 2.2: 2D histograms of summary parameters USFLUXL, MEANGBL, R_VALUE, and AREA between SHARP and SMARP. SHARP summary parameters are suffixed with _HMI and SMARP with _MDI. The orange lines in the diagonal blocks are the least square fit of SMARP summary parameters on the corresponding SHARP summary parameters, with coefficient k , intercept b , and Pearson correlation coefficient r displayed in the corner.

Figure 2.2, USFLUXL is the most correlated parameter between SHARP and SMARP, with Pearson correlation coefficient $r = 0.970$, whereas MEANGBL is the least correlated parameter, with $r = 0.796$. Note that applying linear transformations to the SHARP summary parameters would have no effect once Z-score standardization was performed. This is because Z-scores are invariant to univariate linear transformation. Therefore, in practice, the linear transformation on SHARP summary parameters is not performed.

2.2.3 Sample extraction and labeling

We focus our joint SMARP and SHARP analysis on a particular task of interest, which we refer to as “strong-vs-quiet” flare prediction: based on a sequence of observations of an evolving active region, the objective is to discriminate active regions that will generate strong flares in the near future, from active regions having no flare activity whatsoever. To construct a dataset for this task, we extract data samples using a sliding window approach similar to Angryk et al. [13]. Specifically, samples are extracted from a 24-hour time window

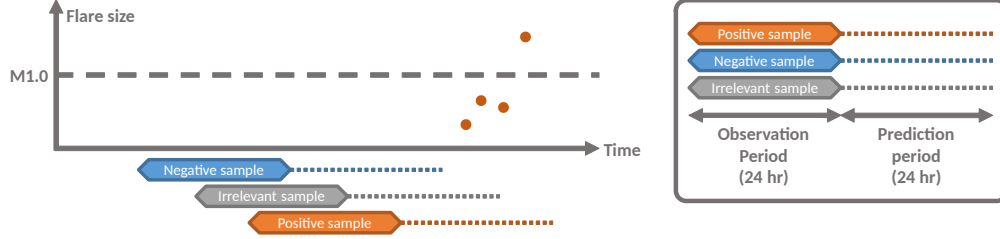


Figure 2.3: Demonstration of the sample extraction and labeling procedure of an active region. The dark orange dots represent flares that occurred in an active region, with the last flare exceeding the M1.0 threshold. The blue sample is labeled as the negative class because no flare of any class occurs in the observation and the prediction period. The gray sample is irrelevant to the task since all flares in the prediction period are weaker than M1.0. The orange sample is labeled as the positive class because the prediction period contains a flare of size exceeding M1.0.

that slides through an active region sequence with step size of 96 minutes (i.e., one 24-hour subsequence starts 96 minutes after the starting point of its previous subsequence). The 24-hour time window that a sample covers is called the *observation period*, and the 24-hour time window following immediately after the observation period is called the *prediction period*. Then we retain the samples that either: (1) exhibit an M- or X-class flare in the prediction period (assigned to the positive class); or (2) have no flare of any class in both observation and the prediction period (assigned to the negative class). Table 2.2 lists the sample sizes of all the possible flare activity evolution types and the associations between evolution types and class labels. Figure 2.3 shows examples of extracted and labeled samples.

We note that two evolution types are excluded in this study. Evolution types denoted by blank spaces in Table 2.2 indicate a decay in flare activity—a process different from the onset or the continuation of the flare activity. They are unrelated to our task and also less studied in the literature. Including them in the dataset brings an unnecessary source of heterogeneity and makes learning a reliable predictor substantially more difficult. Evolution types denoted by question marks have only weak flares in the prediction period. They are excluded to enhance the contrast between the two classes, to avoid the concerns on the detection of weak flares (many B- and C-class flares are obscured by background radiation), and to avoid the controversy on the granularity of labels (for instance, an M1.0 class flare and a C9.9 class flare relieve a similar amount of energy but are categorized differently). Possible limitations of our sample selection rules are discussed in Section 2.5.

After extracting and labeling active region samples, we discarded the samples having inconsistent or missing data. We consider a point-in-time record in a sample sequence as a “bad record” if (1) the magnetogram contains Not-a-Number (NaN) pixels, (2) the

Table 2.2: *Up*: Samples sizes of all possible flare activity evolution types, with missing/inconsistent data removed. The flare activity in the observation period of a sample can be **QUIET** (the active region is flare-quiet), **WEAK** (only flares of size smaller than M1.0 occur), or **STRONG** (there is at least one large flare of size M1.0 or above). The prediction period can be classified the same way. The entries denote the sample counts in SMARP/SHARP data sets. *Down*: The associations of flare activity evolution types and the class labels for the “strong-vs-quiet” flare prediction task. Positive samples are denoted as + and negative samples as -. Samples with the evolution type signifying a decaying flare activity (denoted as a blank space) or leading to only small flares (denoted as ?) are not relevant to our task.

Observation	Prediction		
	QUIET	WEAK	STRONG
QUIET	130695 / 66349	12341 / 10715	932 / 296
WEAK	12688 / 11110	12033 / 14891	1915 / 1366
STRONG	1071 / 282	1723 / 1371	1754 / 1187

Observation	Prediction		
	QUIET	WEAK	STRONG
QUIET	-	?	+
WEAK		?	+
STRONG			+

Table 2.3: Sample sequences extracted from SMARP and SHARP

	Positive (M1.0+)	Negative (Quiet)	Event Rate
SMARP	4601	130695	0.0340
SHARP	2849	66349	0.0412

magnetogram has either height or width deviating more than 2 pixels from the median dimension in the sample sequence, or (3) any of the summary parameters is NaN. A sample is discarded if it contains more than 2 bad records or the last record is a bad record. The validity of the last record is enforced because the CNN uses only the last record in the sample sequence.

The numbers of SMARP and SHARP samples output by the above pipeline are shown in Table 2.3. The count of negative samples is observed to dominate in both SMARP and SHARP. To address the issue of significant class imbalance, we randomly undersample the negative samples to equalize the positive and negative classes, which will be described in more detail in Section 2.2.5.

2.2.4 Train/validation/test split

A common practice in machine learning is to divide the data samples into three disjoint subsets, as known as splits: a training set on which the model is fitted, a validation set on which hyperparameters are selected, and a test set on which the model is evaluated for generalization performance. The ability of a trained machine learning algorithm to generalize to unseen samples hinges on the distributional similarity among splits. Therefore, it is important that splits be sufficiently similar in distribution.

Due to the temporal coherence of an active region in its lifetime, a random split of data samples will have samples coming from one active region categorized into different splits. Such correlation constitutes an undesirable information leakage among splits. For instance, information leaking from the training set into the test set will likely result in an overly optimistic estimate of the generalization performance. Much of the flare prediction literature deals with this issue by taking a chronological split, e.g., a year-based split [e.g. 30, 49]. Unfortunately, it is observed that the splits may not share the same distribution due to solar cycle dependency [301]. Some other works take an active-region-based split, where data samples from the same active region must belong to the same split [e.g. 108, 40, 346, 180]. Compared to splitting by years, this approach has the advantage that active regions in each split are randomly dispersed in different phases of a solar cycle, removing the bias introduced

by artificially specifying splits. This distributional consistency between splits comes at the price of an additional source of information leakage due to sympathetic flaring in co-temporal active regions.

2.2.5 Random undersampling

As shown in Table 2.3, both SMARP and SHARP exhibit prominent class imbalance, with positive (minority class) samples significantly outnumbered by negative (majority class) samples. In flare forecasting, class imbalance has been recognized as a major challenge, both in forecast verification [323, 144] and in data-driven methodology [30, 4]. A data-driven forecasting method needs to be calibrated to handle class imbalance properly, in order to effectively detect the events of interest, as opposed to being overwhelmed by the sheer volume of the negative samples in the training set.

Methods to tackle class imbalance can be categorized into three types: data-level methods, algorithm-level methods, and a hybrid of the two [155, 142]. Data-level methods rebalance the class distribution by oversampling the minority class and/or undersampling the majority class—both have been used in flare forecasting [e.g. 239, 332]. Classifiers trained on rebalanced datasets, without being biased towards the majority class, are more likely to effectively detect the event of interest. Such classifiers are also generally more robust to variations in class imbalance than classifiers trained on the original imbalanced data [326]. Algorithm-level methods modify the learners to alleviate their bias towards the majority groups. The most popular algorithm-level approach—also widely used in flare forecasting [e.g. 30, 218, 188]—is cost-sensitive learning, which assigns a higher penalty to the misclassification of samples from the minority class to boost their importance [155]. The penalty weights for different classes are usually set to be inversely proportional to their samples sizes [218, 188, 4]. Other algorithm-level approaches include imbalanced learning algorithms, one-class learning, and ensemble methods [9], but they are rarely used in flare forecasting. Recent work by Ahmadzadeh et al. [4] provided a thorough investigation of class imbalance in flare forecasting by presenting an empirical evaluation of multiple approaches.

We handle the class imbalance problem using random undersampling: we randomly remove samples from the majority class until the number of positive and negative samples are equalized. By training on rebalanced training and validation sets, we obtain a predictor that is more robust to shifts in climatological flare rates and learns more resilient pre-flare features. Following Zheng et al. [346] and Deng et al. [67], we also perform random undersampling on test sets to preserve the class proportion consistency among splits. By testing on rebalanced test sets, we evaluate the generalization performance of the predictor under

the *same* climatological rate as it is trained with. We note there are also studies that do not rebalance the test set in order to evaluate the performance under a realistic event rate [53, 4]. However, the bias caused by the class-balance change between the test and the training set is often neglected. We discuss such bias in addition to possible corrective methods in Section 2.5. Applying such corrections will be left to future work that directly addresses operational applications.

2.2.6 Image resizing

The CNN requires all input images to be of the same size, but the active region cutouts are of different sizes and aspect ratios. Resizing (via interpolation), zero padding, and cropping are among mostly used methods to convert different-sized images into a uniform size. Jonas et al. [145] cropped and padded input images to a square aspect ratio and then downsampled them to 256×256 pixels. This has the advantage of preserving the aspect ratio. However, since many active regions are east-west elongated, cropping may exclude part of active regions and padding may introduce artificial signals. In this work, we resize all active region magnetograms to 128×128 pixels using bilinear interpolation, similar to Huang et al. [135] and Li et al. [180].

2.2.7 Standardization

Magnetogram pixel values and summary parameters are different physical quantities expressed in different units and ranges. Unlike physical modeling, many machine learning algorithms are invariant to scaling the input; they only care about the relative feature amplitudes. Moreover, drastically different ranges of features may hurt the convergence and stability of many algorithms. Therefore, the data of different scales are typically transformed into the same range via a process called standardization. In particular, Z-score standardization transforms the input data by removing the mean and then dividing by the standard deviation.

In this work, we apply the Z-score standardization to the image data using the mean and standard deviation of the magnetogram pixels in SHARP. This is because the pixel values between SMARP and SHARP are similar. We apply the Z-score standardization to SMARP and SHARP summary parameters separately. That is, the mean and standard deviation are calculated for SHARP and SMARP separately, and data in one dataset is standardized using the mean and the standard deviation in that dataset. The transformation is “global” [4] in that it is calculated regardless of the splits. Empirical evaluation in Ahmadzadeh et al. [4] showed a global standardization to be better than the local standardization, i.e., the

mean and standard deviation are calculated only for the training split. We note that, with this standardization, the linear transformation converting SHARP summary parameters to SMARP proxy data is no longer needed; any coefficients and bias will have no effect after standardization.

2.3 Methodology

In this section, we introduce the two deep learning models, LSTM and CNN, that we use for flare prediction. Then we describe the stacking ensemble approach that combines the two models. Subsequently, we describe the forecast verification methods (skill scores and graphical tools) we used. Then we discuss how we use the paired t -test to compare empirical performance between algorithms and settings with statistical confidence. Lastly, we introduce the visual attribution methods used to interpret the decisions generated by the CNN.

2.3.1 Deep learning models

We use two deep neural network models, LSTM and CNN, to predict strong flares from active region observation. LSTMs use 24-hour-long time series of summary parameters before the prediction period begins, whereas CNNs use the static point-in-time magnetogram right before the prediction period begins. Both networks output the probability that an input sample belongs to the positive class, i.e., the probability that the active region will produce a strong flare in the next 24 hours, rather than continue to be flare-quiescent.

The long short-term memory (LSTM) network was introduced by Hochreiter and Schmidhuber [126] as a type of recurrent neural networks that learns from sequential data for classifying text and speech. A common LSTM unit is composed of a cell, an input gate, an output gate, and a forget gate. In solar flare prediction, LSTMs have been applied to prediction using SHARP parameter series [49, 188]. The architecture of the LSTM used in this chapter is adapted from Chen et al. [49], shown in Figure 2.4(a). Two LSTM layers, each with 64 hidden states, are stacked. The last output of the second LSTM layer, a 64-dimensional vector, is sent to a linear layer with 2 outputs. The softmax is applied to this 2-dimensional output to get the predicted probabilities of the positive and the negative class.

The convolutional neural network (CNN) is a neural network architecture that learns from images. CNNs have been applied to solar flare forecasting by Huang et al. [135] and Li et al. [180]. We use the architecture proposed by Li et al. [180], illustrated in Figure 2.4(b), which was itself inspired by the VGG network [258] and the Alexnet network [157].

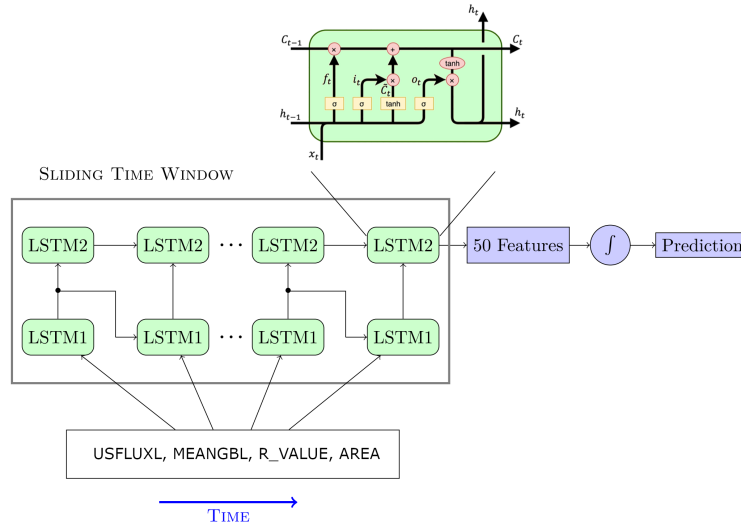
The first two convolutional layers have kernels of size 11×11 , designed to learn low-level and concrete features. The three following convolutional layers have kernels of size 3×3 , designed to learn more high-level, abstract concepts. Batch normalization is used after all convolutional and linear layers to speed convergence. ReLU nonlinearity is applied to only convolutional layers. The batch normalization outputs of the two linear layers are randomly dropped out with a probability of 0.5 in training to reduce overfitting. The 2-dimensional output is passed to softmax to generate a probability assignment between the positive and the negative class. More details of this architecture can be found in Li et al. [180].

The procedures used to train the LSTM and the CNN are similar. For both models, the Adam optimizer [151] is used to minimize the cross-entropy loss with learning rate 10^{-3} and batch size 64. Both models are evaluated on the validation set after each epoch of training. To prevent overfitting, the training is early-stopped if no improvement on the validation True Skill Score (or TSS, explained later in Section 2.3.3) is observed for a certain number of epochs, called the *patience*, before early stopping. The LSTM is trained for at most 20 epochs with a patience of 5 epochs, whereas the CNN is trained at most 20 epochs with a patience of 10 epochs. After training, the LSTM or the CNN with the best validation TSS among the checkpoints of all epochs is selected and evaluated on the test set to estimate its generalization performance.

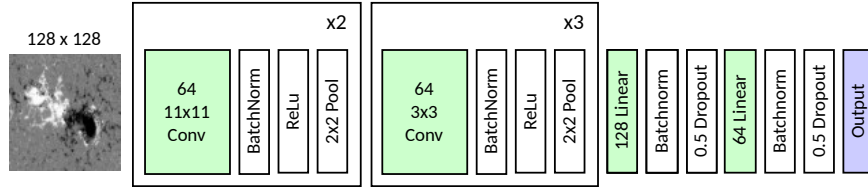
2.3.2 Stacking ensemble

In ensemble learning, the most common approach to combining individual models—called *base learners*—is averaging their outputs, possibly with non-equal weights, to produce a final output. Another approach is stacking. Different from output averaging, stacking uses cross-validation training to learn the best combination—called the *meta-learner*—of the outputs of the base learners. The meta-learner is often chosen to be global and smooth [322], such as linear models [35, 164, 286] and decision trees [287, 78]. Training a stacking ensemble consists of two stages. First, the base learners are fitted on the training set. Then, the predicted probabilities by all base learners on the validation set, as well as their labels, are collected into the so-called *level-one* data, on which the meta-learner is fitted. Cross-validation is frequently used in place of a simple train-validation split to significantly increase the sample size of the level-one dataset. Either way, it is important that the level-one data are out-of-sample for base learners, otherwise the meta-learner will inevitably prefer models that overfit the training data over ones that make more realistic decisions [321].

An early application of stacking for solar flare prediction problems was presented by Džeroski and Ženko [78]. These authors proposed a decision-tree-based stacking method,



(a) LSTM architecture



(b) CNN architecture

Figure 2.4: Neural network architectures. (a) shows the LSTM architecture. (b) shows the CNN architecture.

demonstrating it on the *UCI Repository of machine learning databases* [75], including a dataset with 1389 flare instances, each characterized by 10 categorical attributes. In the space weather community, Guerra et al. [108] proposed stacking for flare prediction. They linearly stacked four full-disk probabilistic forecasting methods, with the weights maximizing HSS under the constraint that they are non-negative and sum to 1. They found that the stacking ensemble performed similarly to an equally weighted model. Guerra et al. [106] continued in this direction adopting stacking for more forecasting methods and a larger data sample. They also considered an unconstrained linear combination with a climatological frequency term. They found most ensembles perform better than a bagging model that essentially averages the members’ predictions. However, these authors overlooked the nonconvexity of the objective in training the meta-learner. Furthermore, their conclusions about the superiority of the stacking ensemble over the equally weighted model were limited to the evaluation of in-sample error, which is unlikely to generalize. We will discuss these issues as we introduce our proposed stacking ensemble.

We formulate the stacking ensemble as a linear combination of LSTM and CNN. Given a sample x_i , the stacking ensemble outputs the probability that the sample belongs to the positive class

$$r_i = \alpha p_i + (1 - \alpha)q_i, \quad 0 \leq \alpha \leq 1, \quad (2.1)$$

where p_i and q_i are the predicted probability by LSTM and CNN, respectively, and α is the meta-learner parameter. We note that, in order to be most effective, stacking methods require base learners that are diverse and complement each other. Examples include base learners trained on different types of data, each providing an alternative view of the same phenomenon. The magnetograms and summary parameters in SHARP/SMARP provide such diverse multiviews. These multiviews are processed by CNN and LSTM, respectively, to generate two predicted probabilities, which are then fused into a single prediction by the aforementioned stacking procedure.

The meta-learning of the stacking ensemble essentially means finding the optimal combination weight α . To that end, we minimize a loss function that penalizes the difference between the prediction r_i and the binary label $y_i \in \{0, 1\}$ for samples in the validation set. A natural choice is to maximize the accuracy or skill scores, or equivalently, to minimize the loss which is the negation of these metrics. The downside of these loss functions is that they may not be convex or differentiable. This is not problematic when there are only a few base learners, as is the case with this work and Guerra et al. [108]; a grid search can be applied to find the weights that minimize the loss. However, as the number of base learners increases,

the grid search quickly becomes infeasible, and iterative algorithms have to be used—many of them require convexity and differentiability of the loss function for guaranteed convergence. In machine learning, convexity and smoothness ensure the uniqueness of the minimizer and guarantee faster convergence rates for iterative algorithms [220, 33]. Guerra et al. [106] found that for some optimization metrics, the resulting weights were sensitive to the initialization of the solver. This is likely the consequence of the nonconvexity of the loss function. Their proposed solution was to run the solver with multiple initializations and take the average.

To circumvent convergence problems, machine learning researchers often use loss functions that are convex and differentiable. One example is the negative log-likelihood function for the logistic regression model, whose minimum corresponds to the maximum likelihood estimator (MLE). Within the meta-learning framework specified by Equation (2.1), the negative log-likelihood loss function is

$$L(\alpha) = -\log \prod_{i=1}^n r_i^{y_i} (1 - r_i)^{1-y_i} \quad (2.2)$$

$$= \sum_{i=1}^n \underbrace{(-y_i \log r_i - (1 - y_i) \log(1 - r_i))}_{L_i}. \quad (2.3)$$

The negative log-likelihood objective can also be interpreted as the binary cross-entropy loss, a divergence measure between the distributions of ground truth labels and predicted probabilities. This loss function can be decomposed into the summation of instance-wise loss L_i , having gradient and the Hessian

$$L'_i(\alpha) = \left(-\frac{y_i}{r_i} + \frac{1 - y_i}{1 - r_i} \right) (p_i - q_i), \quad (2.4)$$

$$L''_i(\alpha) = \left(\frac{y_i}{r_i^2} + \frac{1 - y_i}{(1 - r_i)^2} \right) (p_i - q_i)^2 \geq 0. \quad (2.5)$$

Minimizing L on $\alpha \in [0, 1]$ is a convex optimization problem and we use grid search to find the minimizer. In general cases with more than two base learners, iterative algorithms like projected gradient descent or Newton’s method will be more efficient. We point out that convex loss functions are widely adopted in the literature of stacking, such as least square estimate [35], regularized linear regression [164], multi-response linear regression [286], and hinge loss [255].

		Predicted		Total
		Negative	Positive	
True	Negative	TN	FP	n
	Positive	FN	TP	p
Total		n'	p'	$p + n$

Table 2.4: A contingency table consisting of TP (true positive), FP (false positive), FN (true negative), and TN (true negative).

2.3.3 Evaluation tools

The prediction probabilities output by CNN and LSTM can be turned into binary decisions by thresholding and the algorithm performance can be represented as a contingency table (or confusion matrix), as shown in Table 2.4. The contingency table contains the most complete information for categorical prediction. However, a single numerical metric is often needed to summarize the table for model selection. Accuracy and skill scores are examples of such contingency table based metrics that are adopted in space weather forecasting.

We start our discussion on metrics with accuracy (ACC), also known as rate correct, the simplest metric that is widely used in all sorts of domains. In terms of the contingency table, accuracy is defined as

$$A = \frac{d + a}{n + p}. \quad (2.6)$$

For a highly imbalanced classification problem like solar flare prediction, accuracy is generally not considered a useful metric, since a no-skill classifier that assigns the majority label to all samples will be correct most of the time. Therefore, a plethora of skill scores are devised to overcome this issue.

A skill score provides a normalized measure of the improvement against a specific reference method. In its most general form, a skill score can be expressed as

$$\text{Skill} = \frac{A_{\text{forecast}} - A_{\text{reference}}}{A_{\text{perfect}} - A_{\text{reference}}}, \quad (2.7)$$

where A_{forecast} , $A_{\text{reference}}$, and A_{perfect} are the accuracy of the forecast to be evaluated, the reference forecast, and the perfect forecast, respectively. A higher skill score indicates better performance, with the maximum value 1 corresponding to the perfect performance, 0 corresponding to no improvement over the reference, and negative values corresponding to performance worse than the reference. Below, we review some of the mostly used skills scores

in flare forecasting. For a more complete discussion, we refer readers to Woodcock [323] and Wilks [320].

The Heidke Skill Score (HSS), also known as Cohen’s kappa coefficient due to Cohen [55], uses a random forecast independent from the flare occurrences as a reference. The expected number of correct forecasts made by the random predictor, denoted by E , can be calculated using the law of total expectation as

$$E = \frac{p}{p+n} \times p' + \frac{n}{p+n} \times n'. \quad (2.8)$$

The accuracy of the random predictor can then be expressed as

$$A_{\text{reference}} = \frac{E}{p+n}. \quad (2.9)$$

Defined using this reference accuracy, HSS has the form

$$\text{HSS} = \frac{a+d-E}{p+n-E} = \frac{2[(a \times d) - (c \times b)]}{pn' + p'n}. \quad (2.10)$$

HSS quantifies the forecast improvements over a random prediction. Since the random reference forecast is dependent on the event rate (climatology) $p/(p+n)$, HSS has to be used with discretion in comparing methods when the event rate varies.

The True Skill Score (TSS), also known as Hanssen & Kuiper’s Skill Score (H&KSS) or Peirce Skill Score. It is the difference between the probability of detection (POD) and the false alarm rate (FAR):

$$\text{TSS} = \underbrace{\frac{a}{p}}_{\text{POD}} - \underbrace{\frac{b}{n}}_{\text{FAR}}. \quad (2.11)$$

TSS falls into the general skill score definition with a reference accuracy [21]

$$A_{\text{reference}} = \frac{c(d-b)^2 + b(a+c)^2}{(n+p)[c(d-b) + b(a+c)]}, \quad (2.12)$$

constructed such that both the random and unskilled predictors score 0. A nice property of TSS is its invariance to the class imbalance ratio, and hence is suggested by Bloomfield et al. [29] to be the standard measure for comparing flare forecasts.

We note that, on a balanced dataset for which the event rate is 0.5, it can be shown that $\text{TSS} = \text{HSS} = 1 - 2(1 - \text{ACC})$. The trend and the paired t -test results for TSS apply to ACC and HSS due to perfect correlation. Therefore, we mainly focus on the discussion on TSS,

list ACC as a complement metric, and omit HSS as it is equal to TSS in our setting. For probabilistic forecasts, the aforementioned metrics (ACC, HSS, and TSS) depend upon the threshold applied to the predicted probability. A common practice is to apply a threshold of 0.5, which is considered to be “random” by many researchers. In contrast, the following two metrics, BSS and AUC, are irrelevant to the threshold, and they need information (i.e., predicted probabilities) beyond the mere contingency table.

The Brier Skill Score (BSS) is a skill score evaluating the quality of a probability forecast. It is of a nature different from those of HSS and TSS, in that it directly uses probabilistic predictions without thresholding them. The BSS also admits the general skill score formulation, with the accuracy replaced by the Brier Score (BS), defined as the mean squared error between the probability predictions f_i ’s and binary outcomes o_i ’s:

$$\text{BS} = \frac{1}{n} \sum_{i=1}^n (f_i - o_i)^2. \quad (2.13)$$

With a reference forecast that consistently predicts the average event frequency $\bar{o}$ (also known as climatology), the BSS is given by

$$\text{BSS} = 1 - \frac{\text{BS}_{\text{forecast}}}{\text{BS}_{\text{reference}}}. \quad (2.14)$$

It is sometimes of interest to decompose BS into three components of reliability, resolution, and uncertainty [207, 201]. BSS is frequently accompanied by the reliability diagram discussed below, providing more complete information about the performance of probabilistic predictions.

For completeness we briefly discuss the Area Under Curve (AUC), defined as the area under the receiver operating characteristic (ROC) curve. The ROC curve depicts how the probability of detection changes with the false alarm rate by varying the classification threshold. A higher AUC implies a higher average detection probability over all false alarm rates. Unlike dichotomous metrics like TSS and HSS, AUC summarizes detection performance over all possible false positive rates and, in particular, does not depend on the threshold selected to convert probabilistic forecasts into binary decisions. Consequently, the AUC is not as useful in flare prediction, especially when stringent false positive control is exercised [262].

The above skill scores provide one way to directly compare flare prediction models. In addition to such metrics, flare forecasts are often evaluated using graphical tools for diagnostics and comparison. Apart from ROC curves discussed above, reliability diagrams (RD) and skill score profiles (SSP) are also commonly used in forecast verification. All three of them are applicable to forecasts that predict probabilities or continuous scores (e.g., logits)

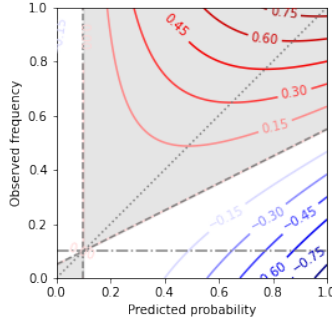


Figure 2.5: An illustration of the relation between the reliability diagram and BSS.

that can be converted to probabilities. We briefly discuss RD and SSP below.

The reliability diagram, also known as the calibration curve, measures how well a probabilistic forecast agrees with the observation. The predicted probabilities are binned into groups and the observed event rate within each group is plotted. If the predicted probability agrees well with the observed rate, the points will be close to the diagonal of the plot (the line of perfect reliability). Such a forecast is called reliable. Any forecast that produces predictions independent of flare activity has all its points close to the horizontal line at the event rate. BSS provides a metric that accounts for both reliability and resolution. Figure 2.5 shows an example of the plane on which the reliability diagram is drawn. The climatology rate is set to be 0.1. The overall BSS can be seen as a histogram weighted average of the contributions of the points on the reliability diagram. The contours are equal contribution lines. The points in the shaded area contribute positively to BSS. The dashed line with slope 1/2 is called the “no skill” line, the points on which have zero contribution to the overall BSS.

A skill score profile plot shows how skill scores change as a function of the probability threshold. A method with high and flat profile is usually desired, as such a method achieves high skill scores and the performance is robust to the changes of the threshold.

2.3.4 Statistical performance comparisons

We use a one-sided paired t -test to assess the comparative performance of a pair of prediction algorithms, called algorithm 1 and 2, that are tested on the same test data. Specifically, two competing hypotheses are formulated: the null hypothesis (H_0) that algorithms 1 and 2 have identical performance and the alternative hypothesis (H_1) that algorithm 2 is better than algorithm 1. Suppose we have n pairs of empirical performance $\{(x_i, y_i)\}_{i=1}^n$ achieved by algorithm 1 and 2 on n test samples. The paired t -statistic is $t = \sqrt{n}(\bar{y} - \bar{x})/\sigma$ where $\bar{y}$ and

$\bar{x}$ are the sample means of $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$, respectively, and σ^2 is the sample variance of the differences $\{y_i - x_i\}_{i=1}^n$. Under H_0 , t follows a Student- t distribution with $n - 1$ degrees of freedom [25]. The p -value associated with the test statistic t is defined as the area under the Student- t density to the right of the value t . Small p -values provide strong evidence in favor of H_1 , i.e., that algorithm 2 is better than algorithm 1. In this chapter, we use the paired t -test to examine the following hypotheses:

- Training a predictor (LSTM or CNN) on data from two solar cycles (SMARP and SHARP) improves upon training a predictor on data from a single solar cycle (only SMARP or only SHARP) (Section 2.4.1).
- LSTM achieves better performance than CNN (Section 2.4.2).
- The LSTM-CNN stacking ensemble achieves better performance than the better model between LSTM and CNN (Section 2.4.3).

2.3.5 Interpretation of CNNs

Deep learning methods are widely applied to many domains such as computer vision, natural language processing, speech processing, robotics, and games [see, e.g. 121, 257, 70]. As of today, deep learning algorithms largely remain black box methods, raising concerns of lack of interpretability, transparency, accountability, and reliability. Interpretability is of particular importance when deep learning is used in scientific discovery. Over the years, many tools for interpreting the functioning of deep neural networks have been proposed, revealing aspects of their underlying decision process.

One way to interpret a black-box model, often referred to as “attribution”, is to see how different parts of the input contribute to the model’s output. An attribution method generates a vector of the same size as the input, with each element indicating how much the corresponding element in the input contributes to the model decision for that input. In the context of CNNs, the attribution vector is a heatmap of the same size as the input image.

A multitude of attribution methods have been proposed for CNNs in the task of image classification. One type of approach is perturbation-based methods, among which occlusion [339] is well known. Occlusion masks the input image with a gray patch at different locations and sees how much the prediction score of the ground truth class drops. The prediction score drop varies with location, forming a heatmap, with large values indicating the positions of the features important to the CNN’s correct prediction. One drawback of the occlusion method is that it is computationally expensive. Another drawback is that the attribution

depends on the size and shape of the patch, which need to be tuned to obtain sensible results. Therefore, this type of approach is not used in our work.

Another type of approach uses gradient-based methods, the basic idea being that the gradient of the predicted score of a certain class with respect to the input reveals the contribution of each dimension of the input. Saliency map [259], one of the earliest gradient-based methods, is simply the absolute value of the gradients. The intuition is that the magnitude of the derivative indicates which pixels need to be changed the least to affect the class score the most [259]. Deconvolution Network [339] and Guided Backpropagation [261] attribution methods modify the backpropagation rule. Integrated Gradients [272] integrate the gradients along the path from a reference image to the target image. Formally, the integrated gradient along the i -th dimension for an input x and a baseline x' is

$$L_i^c(x; x') = (x_i - x'_i) \times \int_{\alpha=0}^1 \frac{\partial F_c(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha, \quad (2.15)$$

where $F_c(x)$ is the model output for class c with input x . One desirable property of Integrated Gradients, known as completeness, is that the pixels in the attribution map add up to the difference of prediction scores of the target and the reference image, i.e., $F(x) - F(x')$. DeepLIFT [256] and its gradient-based interpretation [12] can be seen as the gradient with modified partial derivatives of non-linear activations with respect to their inputs. Grad-CAM (Gradient-weighted Class Activation Mapping) [254] accredits decision-relevant signatures by generating a saliency map, highlighting pixels in the input image that increase the confidence of the network’s decision for a particular class. More formally, the Grad-CAM heatmap L^c for class c with respect to a particular convolutional layer is given by the positive part of the weighted sums of the layer’s activation maps A_k , i.e.,

$$L^c = \text{ReLU} \left(\sum_k \alpha_k^c A^k \right), \quad (2.16)$$

with weights α_k^c given by the spatial average of partial derivatives of the class-specific score y^c with respect to the class activation map as

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}, \quad (2.17)$$

where Z is a normalization constant. Intuitively, a class activation map is weighted more if the pixels therein make the CNN more confident in its decision that the input belongs to class c .

In solar flare prediction, Bhattacharjee et al. [24] applied the occlusion method and found that CNNs pay attention to polarity inversion regions. Yi et al. [329] applied Grad-CAM to CNNs and found that polarity inversion lines in full-disk MDI and HMI magnetograms are highlighted as an important feature for flare prediction. In this chapter, using a variety of attribution methods, we observe similar trends for the CNN trained on SHARP and SMARP data.

2.4 Results

We aim to answer the following four scientific questions: (1) Can additional data from another solar cycle benefit the performance of deep learning methods for solar flare prediction? (2) Do features implicitly learned by CNN work better than handcrafted physical parameters used by LSTM? (3) Can we combine the two deep learning methods to obtain a better prediction? (4) What preflare signatures can the CNN detect from the magnetogram of an active region?

To summarize, we report the following findings: (1) Additional training data from HMI collected in Solar Cycle 24 improve the predictive performance of both LSTM and CNN when tested on Solar Cycle 23. (2) LSTM (using summary parameters) generally outperforms CNN (directly using magnetograms) in flare prediction. (3) Stacking CNN and LSTM generally leads to better prediction performance. (4) Visual attribution methods help us interpret the decision of CNN by identifying preflare features. This section presents the empirical results that lead to these findings.

2.4.1 Data from another solar cycle improves prediction

A major goal of this chapter is to examine the utility of using SMARP and SHARP together. We set an experimental group and a control group and contrast their 24-hour “strong-vs-quiet” flare prediction performance. The control group consists of models that train, validate, and test exclusively on SHARP data. We refer to this type of dataset as **SHARP_ONLY**. Compared to the control group, models in the experimental group have the training set enriched by SMARP data, while the validation and the test set are kept the same. We call this type of dataset **FUSED_SHARP**. The only difference between **SHARP_ONLY** and **FUSED_SHARP** is that models using **FUSED_SHARP** have access to data from a previous solar cycle in the training phase. Symmetrically, we design **SMARP_ONLY** and **FUSED_SMARP** to examine the utility that SHARP brought to SMARP. Specifically, the four types of datasets are generated as follows:

1. Dataset **SHARP_ONLY**: 20% of all the HARP’s are randomly selected to form a test set.

Table 2.5: Sample sizes of a random realization of the four datasets

	Train		Validation		Test	
	Positive	Negative	Positive	Negative	Positive	Negative
SHARP_ONLY	1774	1774	665	665	410	410
FUSED_SHARP	6375	6375	665	665	410	410
SMARP_ONLY	2849	2849	860	860	892	892
FUSED_SMARP	5698	5698	860	860	892	892

20% of the remaining HARPs are randomly selected to form a validation set. The rest of the HARPs belong to the training set. In each split, negative samples are randomly selected to match the number of positive samples.

2. Dataset **FUSED_SHARP**: The test set and the validation set stay the same, respectively, with those in **SHARP_ONLY**. The remaining HARPs are combined with all TARPs to form the training set. In each split, negative samples are randomly selected to match the number of positive samples.
3. Dataset **SMARP_ONLY**: 20% of all the TARPs are randomly selected to form a test set. 20% of the remaining TARPs are randomly selected to form a validation set. The rest of the TARPs belong to the training set. In each split, negative samples are randomly selected to match the number of positive samples.
4. Dataset **FUSED_SMARP**: The test set and the validation set stay the same, respectively, with those in **SMARP_ONLY**. The remaining TARPs are combined with all HARPs to form the training set. In each split, negative samples are randomly selected to match the number of positive samples.

Since train/validation/test split and undersampling are both random, repeating these two steps with different seeds enables uncertainty quantification to the evaluation results. The tally of samples produced by one particular random seed is shown in Table 2.5. On each of the four types of datasets, LSTMs and CNNs are fitted on the training set, validated on the validation set, and evaluated on the test set.

Table 2.6 shows the results of the “strong-vs-quiet” active region prediction using the LSTM and the CNN. For LSTMs, a consistent improvement on the fused datasets (**FUSED_SHARP** and **FUSED_SMARP**) is observed in terms of the mean of all metrics. This aligns with the fact that more data are typically desired to improve the generalization performance of deep learning models because they are overparameterized and can easily overfit on small

Table 2.6: Test set performance of the LSTM and the CNN on 24-hour “strong-vs-quiet” flare prediction. The two datasets within each comparison group share common test sets. The $1\text{-}\sigma$ error is calculated from 10 random experiments. Bold fonts indicate the experiments in which the mean of the metric on the fused dataset is higher than that on the single dataset within the group. Note that comparisons should not be made between groups because the metrics are evaluated on different test sets.

	Dataset Model	Group 1		Group 2	
		FUSED_SHARP	SHARP_ONLY	FUSED_SMARP	SMARP_ONLY
ACC	CNN	0.906+/-0.036	0.922+/-0.017	0.901+/-0.028	0.877+/-0.031
	LSTM	0.950+/-0.012	0.942+/-0.016	0.905+/-0.025	0.900+/-0.024
AUC	CNN	0.980+/-0.009	0.981+/-0.006	0.963+/-0.017	0.950+/-0.020
	LSTM	0.990+/-0.004	0.986+/-0.004	0.966+/-0.015	0.963+/-0.015
TSS	CNN	0.812+/-0.071	0.843+/-0.034	0.802+/-0.056	0.754+/-0.061
	LSTM	0.900+/-0.023	0.884+/-0.032	0.810+/-0.050	0.800+/-0.049
BSS	CNN	0.649+/-0.152	0.714+/-0.064	0.628+/-0.114	0.520+/-0.121
	LSTM	0.799+/-0.036	0.775+/-0.047	0.626+/-0.107	0.586+/-0.108

datasets. For CNNs, an improvement is observed on FUSED_SMARP over SMARP_ONLY, but not on FUSED_SHARP over SHARP_ONLY. This indicates that the lower image quality in SMARP has a negative effect on CNN’s performance.

The statistical significance of the improvement on the fused datasets is tested using a one-sided paired t -test with significance level 95%. Table 2.7 shows the t -statistics and the associated p -values of the paired t -tests. The bold font p -values are less than 0.05 and considered to be significant. For LSTMs, the fused datasets are better than the single datasets in a statistically significant way in almost all settings. The only exception is BSS on FUSED_SHARP, whose p -value is only slightly larger than 0.05. For CNNs, across all metrics, statistically significant improvement is observed for FUSED_SMARP over SMARP_ONLY, but not for FUSED_SHARP over SHARP_ONLY. This indicates that adding SHARP magnetograms into SMARP during training helps the CNN to better predict flares, but not the other way around. One potential reason is SMARP magnetograms have a lower signal-to-noise ratio than SHARP magnetograms, which may have negatively affected the CNN. The LSTM, on the other hand, uses the active region summary parameters, which could suppress the effect of noise during summarizing magnetograms, providing information in a sufficiently good quality that does not offset the improvement induced by the increased training sample size.

Aside from the numerical metrics, we provide graphical evaluation results for Group 1

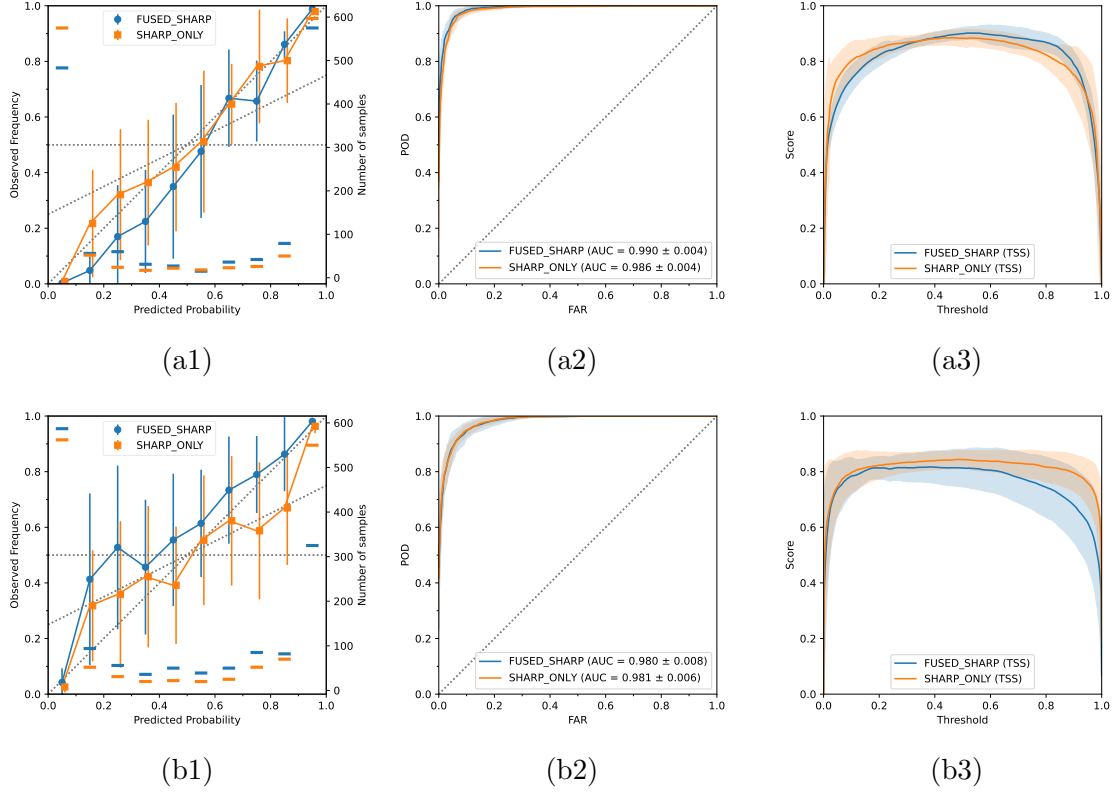


Figure 2.6: Verification plots on SHARP test data to compare models trained on FUSED_SHARP and SHARP_ONLY. Shown in (a1)–(a3) are the reliability diagram, ROC, and SSP for LSTM. Shown in (b1)–(b3) are the same plots but for CNN. In each panel, the blue/orange curve is the test performance for the model trained on FUSED_SHARP/SHARP_ONLY. In each graph, solid curves and error bars (or shaded area) indicate respectively the means and the standard deviations calculated from 10 random experiments. In each reliability plot, the short horizontal bars indicate the number of samples in each probability bin, and the two curves are separated horizontally to prevent error bars from overlapping.

Table 2.7: Paired t -tests for significant improvement of test set performance on the fused datasets as measured by different metrics. The alternative hypothesis H_1 claims that metric S on the fused dataset (FUSED_SHARP or FUSED_SMARP) is greater than the respective single dataset (SHARP_ONLY or SMARP_ONLY), which is tested against the null hypothesis H_0 claiming otherwise. The bold font p -values are less than 0.05 and considered to be significant.

Metric S	Model	H_1		$S_{\text{FUSED_SHARP}} > S_{\text{SHARP_ONLY}}$		$S_{\text{FUSED_SMARP}} > S_{\text{SMARP_ONLY}}$	
		p -value	t	p -value	t	p -value	t
ACC	CNN	0.885787	-1.292359	0.001862	3.881137	0.001862	3.881137
	LSTM	0.016544	2.514074	0.026797	2.219666	0.026797	2.219666
AUC	CNN	0.589845	-0.233881	0.001399	4.070352	0.001399	4.070352
	LSTM	0.000459	4.842485	0.033930	2.074572	0.033930	2.074572
TSS	CNN	0.885787	-1.292357	0.001862	3.881135	0.001862	3.881135
	LSTM	0.016544	2.514079	0.026796	2.219673	0.026796	2.219673
BSS	CNN	0.889419	-1.314583	0.000482	4.806837	0.000482	4.806837
	LSTM	0.054812	1.775082	0.000099	6.014784	0.000099	6.014784

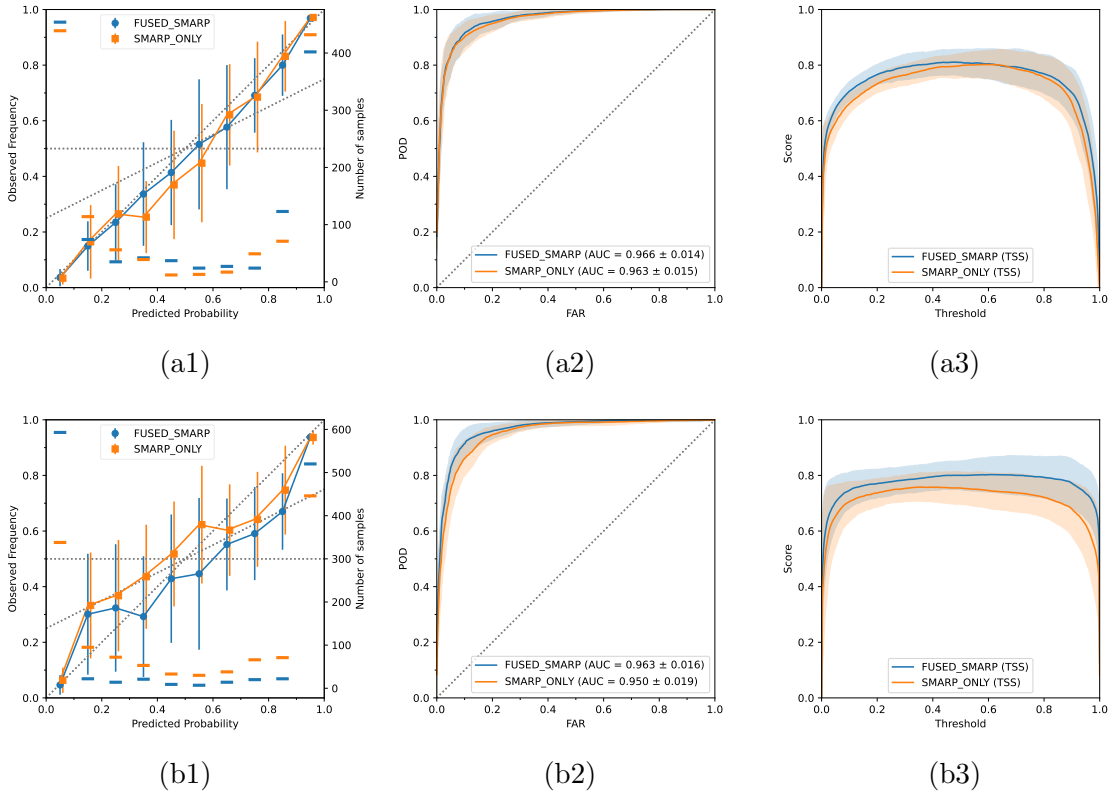


Figure 2.7: Same as Figure 2.6 but for SMARP test data to compare models trained on FUSED_SMARP and SMARP_ONLY.

(`FUSED_SHARP` and `SHARP_ONLY`) in Figure 2.6, and Group 2 (`FUSED_SMARP` and `SMARP_ONLY`) in Figure 2.7. A trend of over-forecasting for high probabilities and under-forecasting for low probabilities is observed in some cases but such effect is minor considering the size of the error bars. In reliability diagrams, all models have points closer to the diagonal, indicating high reliability. In ROC plots, it is observed that the LSTM achieves higher AUC on the fused datasets (`FUSED_SHARP` and `FUSED_SMARP`) than on the single datasets (`SHARP_ONLY` and `SMARP_ONLY`). For the CNN, similar improvement is also observed in the comparison of `FUSED_SMARP` and `FUSED_SHARP`, whereas the ROCs are almost indistinguishable for `FUSED_SHARP` and `SHARP_ONLY`. In skill score profiles, the TSS for LSTM trained on fused datasets are at the same level as that trained on single datasets. For the CNN, on the other hand, `FUSED_SHARP` displays a disadvantage against `SHARP_ONLY`, whereas `FUSED_SMARP` displays an advantage over `SMARP_ONLY`. This verifies the observations made from metrics. In all cases, the skill score profiles are high and relatively flat, indicating the robustness of the performance to the change of thresholds within a wide range of the varying threshold.

2.4.2 LSTM performs better than CNN

This section provides forecast verification to the LSTM and the CNN. We use the same evaluation results for 10 experiments in each setting mentioned in Section 2.4.1, but present them in a way that makes it easier to compare the LSTM and the CNN. We note the differences between our verification set-up and that in an operational setting:

1. In terms of data, the test set of our sort has lots of samples removed based on their active regions, observational data, and flare activities. About 1/5 of tracked active region time series in the evaluation period (May 2010–December 2020) are selected. Within each active region series, only samples with good quality observation and certain flaring patterns are selected (detailed in Section 2.2.3). Negative samples (flare-quiet active regions) are significantly downsampled to match the number of positive samples (strong-flare-imminent active regions). In contrast, operational forecasts do not discard any sample unless absolutely necessary.
2. In terms of outcomes, the forecast of our sort is independent for individual active regions, with the prediction result available every 96 minutes (i.e., MDI cadence) for valid active regions. In contrast, the end goal of an operational forecast is a full-disk forecast. For operational forecasts built upon active region based forecasts, the predictions for all active regions on the solar disk are aggregated to compute the full-disk prediction. In addition, operational forecasts are typically issued at a lower frequency (e.g., every 6 hours), but in a consistent manner.

Table 2.8: Paired t -tests for significant improvement of the LSTM over the CNN in terms of different metrics S on the test set of the four datasets. The alternative hypothesis H_1 claims $S_{\text{LSTM}} > S_{\text{CNN}}$. The bold font p -values are less than 0.01 and considered to be significant.

Dataset	FUSED_SHARP		SHARP_ONLY		FUSED_SMARP		SMARP_ONLY	
	p -value	t	p -value	t	p -value	t	p -value	t
Metric S								
ACC	0.001442	4.050296	0.007142	3.028403	0.234866	0.754672	0.001079	4.245351
AUC	0.003757	3.429557	0.002527	3.682754	0.227978	0.779031	0.000743	4.501005
TSS	0.001442	4.050297	0.007142	3.028405	0.234865	0.754673	0.001079	4.245350
BSS	0.005296	3.213872	0.002645	3.653351	0.531965	-0.082481	0.005781	3.159315

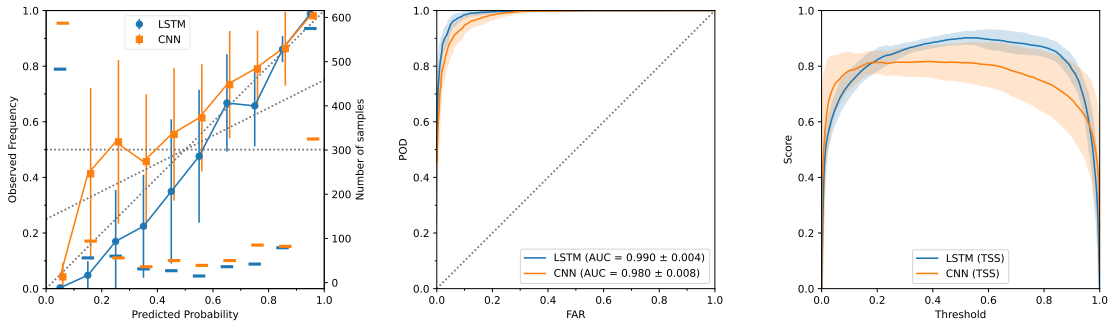


Figure 2.8: Verification plots of the LSTM and the CNN on FUSED_SHARP. Shown are the reliability diagram, ROC, and SSP, from left to right. This figure essentially extracts the blue curves (representing FUSED_SHARP) in both rows of Figure 2.6 and overlaps them together.

The verification results in this section should be interpreted with the above differences in mind.

It can be seen from Table 2.6 that the LSTM generally scores higher than the CNN in terms of mean performance. We performed paired t -test to validate this observation. The results in Table 2.8 confirm that the LSTM scores significantly higher ($p < 0.01$) than the CNN across all metrics on all datasets except for FUSED_SMARP. On FUSED_SMARP, although we cannot claim statistical significance, the LSTM's performance is slightly better or at the same level as the CNN as is observed from Table 2.6.

We only present the graphical verification results for both models trained and tested FUSED_SHARP, given that SHARP is widely used and validated by a wealth of studies. For the results on other datasets, the visualization can be obtained by simply rearranging the same results shown in Figure 2.6 and 2.7.

The reliability diagram in Figure 2.8 shows that the probabilistic prediction given by the LSTM is closer to the diagonal than the CNN, and hence more reliable. The CNN exhibits a trend of under-forecasting especially when the predicted probability is less than 0.5. The

histogram of predicted probability shows that probabilistic forecast by the LSTM is “more confident”, or has higher resolution, than LSTM, with most of the predicted probabilities close to 0 or 1.

The ROC in Figure 2.8 shows a clear advantage of the LSTM over the CNN, in the sense that it achieves a higher probability of detection with the same false alarm rate. This trend is also manifested in terms of AUC.

The SSP in Figure 2.8 shows LSTM achieves higher TSS on average for all thresholds within 0.2–0.9. It is also observed that the TSS for LSTM is maximized by a threshold very close to the climatological rate on the test set (which is 0.5 in our case), a necessary condition for a reliable predictor [158].

2.4.3 Stacking LSTM and CNN leads to better prediction

In this chapter, we only consider stacking methods to combine the CNN and the LSTM hoping for better predictive performance. We evaluate the test set performance of stacking methods using four different criteria:

- **CROSS_ENTROPY**: weights are optimized to minimize cross-entropy loss on the validation set.
- **BSS**: weights are optimized to maximize BSS on the validation set.
- **AUC**: weights are optimized to maximize AUC on the validation set.
- **TSS**: weights are optimized to maximize TSS on the validation set.

Among these criteria, cross-entropy and negative BSS are known to be convex; TSS is neither convex nor concave; we observe AUC to be concave but we do not have proof other than empirical evidence. Criteria **HSS** and **ACC** are excluded from the evaluation since their stacking weights are the same as that of TSS due to the perfect correlation mentioned in Section 2.3.3.

To provide baseline performances, we include the evaluation results for the two base learners, **LSTM** and **CNN**. In addition to the above stacking methods, we consider two other meta-learning schemes:

- **AVG** outputs the average of predicted probabilities of two base learners.
- **BEST** [78] selects the base learner that performs the best on the validation set and applies it to the test set.



Figure 2.9: Test set TSS for base learners and meta-learners using different criteria.

Splitting and undersampling are randomly performed 10 times on each of the four datasets FUSED_SHARP, FUSED_SMARP, SHARP_ONLY, and SMARP_ONLY. The test set TSS of the 10 random experiments for each criterion on each dataset are summarized as box plots in Figure 2.9. The optimal stacking weights for the four stacking ensembles are summarized in Figure 2.10.

Figure 2.9 shows that stacking methods perform slightly better than the BEST meta-learner, especially on FUSED_SMARP and SMARP_ONLY. Of note, the wide error bars are partially due to the randomness originating from data sampling. To fairly compare the methods, we perform paired t -tests with significance level 0.05. It turned out stacking is significantly better than BEST in the following three settings: BSS on FUSED_SMARP ($p = 0.048$), AUC on SMARP_ONLY ($p = 0.025$), and TSS on SMARP_ONLY ($p = 0.013$).

We also note in Figure 2.9 that BEST unsurprisingly achieves better performance than AVG but is slightly worse than the better performing base learner LSTM, most noticeably on FUSED_SHARP. In fact, BEST decided that CNN is the better model in 3 out of 10 experiments on FUSED_SHARP. This is not unexpected because the “best” model on the validation set is not necessarily the best on the test set.

From Figure 2.10, we can see that α is greater than 0.5 in most experiments, with the median falling between 0.55 and 0.9 in all settings. This suggests that stacking ensembles generally depend more on the LSTM than on the CNN. The variance of α is large in some settings, especially for the AUC on FUSED_SMARP. The variance of convex criteria (CROSS_ENTROPY and BSS) is not smaller than that of nonconvex criteria (TSS), indicating that the local minima of non-convex loss functions is not the major source of variance. We suspect the major source of the variance comes from the data sampling bias among experiments, which is, in turn, a collective consequence of the insufficient sample size, heterogeneity across active regions, and possibly a small amount of information leakage because the validation set is used both in the validation of base learners and the training of the meta-learner.

We inspect one experiment of stacking with criterion ACC and the results are presented



Figure 2.10: Stacking weight α fitted using different criteria on different datasets. All 10 values of α in an experiment setting are shown as points next to the corresponding box.

in Figure 2.11. Figure 2.11 (a1)–(a3) show the predicted probabilities by the LSTM and the CNN of each instance in the training, the validation, and the test set. The points are colored by their labels, with red representing the positive class and blue representing the negative class. The green solid line in (a2) and (a3) shows the decision boundary by the meta-learner with α fitted on the validation set to maximize ACC. The points (p, q) on the upper right side of the boundary are classified as positive because they satisfy $r = \alpha p + (1 - \alpha)q > 0.5$. In this experiment, the fitted $\alpha = 0.586$, suggesting the stacking ensemble relies almost equally on the CNN and the LSTM. The violet dashed line in (a3) is the decision boundary with α fitted on the test set, and hence can be seen as the oracle. It can be observed that the distribution of predicted probabilities on the validation set (a2) and the test set (a3) are similar. The distribution of predicted probabilities on the training data in (a1), on the other hand, looks completely different, with the CNN achieving almost perfect separation. In fact, the CNN overfitted on the in-sample data, as indicated by a significantly lower positive recall rate in (a2) and (a3). This validates the decision that meta-learners should not be fitted on the predicted probabilities of the same data used to train the base learners.

Figure 2.11(b) exhibits the stacking optimization process for the same experiment, in which the ACC is calculated on the validation set (a2) and the test set (a3) by scanning over a fine grid of $\alpha \in [0, 1]$ with resolution 0.001. Although the validation ACC (blue curve) is not concave with respect to α , it does exhibit a maximum at $\alpha = 0.586$ as indicated by the vertical green line. The stacking ensemble’s test set performance over α is shown as the red curve. These curves indicate that stacking the LSTM and the CNN indeed results in a small improvement of performance relative to implementing either of them alone, corresponding to the values of ACC at $\alpha = 1$ or $\alpha = 0$.

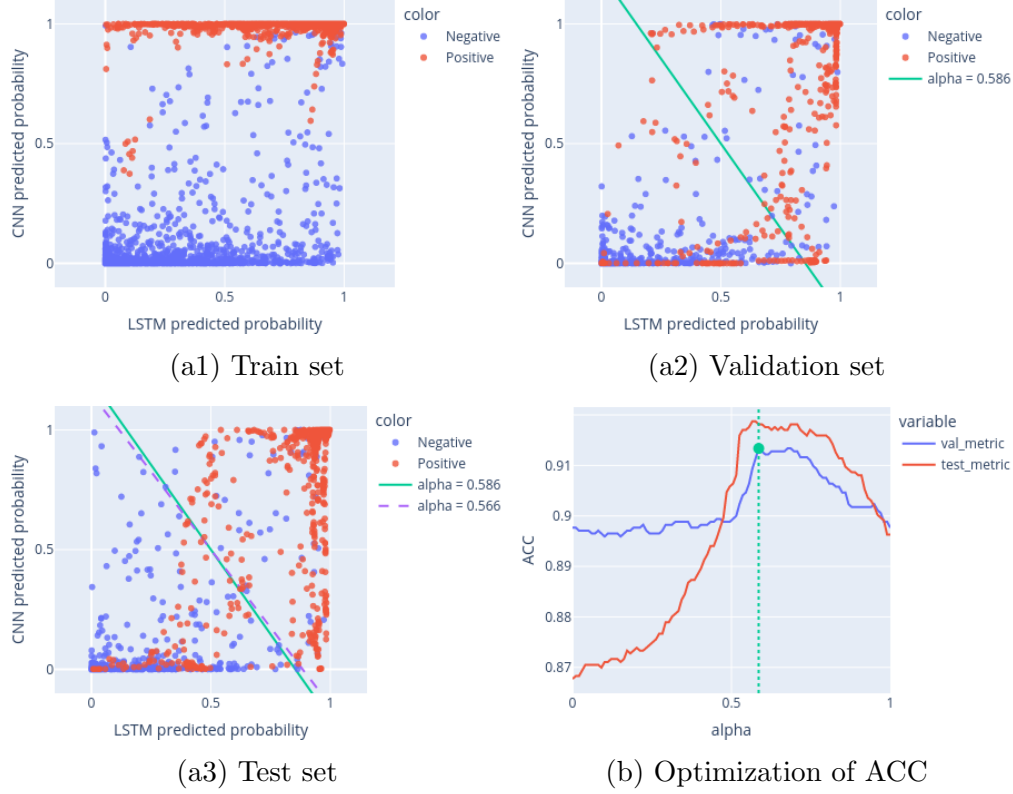


Figure 2.11: (a1)–(a3): CNN predicted probability (y-axis) vs. LSTM predicted probability (x-axis) for the train, the validation, and the test set. The green solid line in (a2) and (a3) is the decision boundary of the ensemble with meta-learner fitted on the validation set. The violet dashed line in (a3) is analogous to the green line except that it is fitted on the test set, and hence can be seen as the oracle. (b): ACC as a function of α on the validation and the test set. The vertical green line shows the value of α that maximizes the validation ACC. The leftmost values of the ACC curves ($\alpha = 0$) correspond to the ACC of the CNN, and the rightmost values of these curves ($\alpha = 1$) correspond to the ACC of the LSTM.

2.4.4 CNN identifies the emergence of preflare features

We use visual attribution methods to extract flare-indicative characteristics of magnetograms from trained CNNs. First, we use synthetic images to examine patterns that contribute to a positive decision of CNNs. The results of synthetic images help us understand better the attribution maps of real magnetograms. Then, we apply visual attribution methods to image sequences of selected active regions that transition from a flare-quiet state to a flare-imminent state. Setting the baseline to the first image in the sequence gives a time-varying attribution map that tracks magnetic field variations that contribute to the change in the predicted probability.

2.4.4.1 Synthetic image

To assist our understanding of attribution maps obtained by different methods, we first turned to synthetic magnetograms. We take the bipolar magnetic region (BMR) model in Yeates [327], represented as line-of-sight magnetic field B as a function of Heliographical location (s, ϕ) , where s denotes sine-latitude and ϕ denotes Carrington longitude. B is parameterized by amplitude B_0 , polarity separation ρ (in radian), tilt angle γ (in radian) with respect to the equator, and size factor a fixed to be 0.56 to match the axial dipole moment of SHARP [327]. The untilted BMR centered at origin has the form

$$B(s, \phi) = -B_0 \frac{\phi}{\rho} \exp \left[-\frac{\phi^2 + 2 \arcsin^2(s)}{(a\rho)^2} \right]. \quad (2.18)$$

We sweep a grid of B_0 , ρ , and tilt angle γ to generate a BMR dataset. Of particular interest are synthetic BMRs considered to be flare-imminent by CNNs. Figure 2.12 shows some examples of them and their attribution results, from which patterns of positive predictions can be summarized. Guided Backpropagation heatmaps have both poles highlighted with the signs matching the polarities. Integrated Gradients produces heatmaps that are more concentrated to polarity centers and attribute more credits to the negative polarities. DeepLIFT produces similar heatmaps to those by Integrated Gradients. Grad-CAM's results are not as interpretable as the above methods. They seem to avoid the polarities and highlight the background and sometimes the polarity inversion lines.

2.4.4.2 The emergence of preflare signatures in the active region evolution

We focus on the attribution results on SHARP as opposed to SMARP because the former has magnetograms of higher resolution and lower noise level. We choose the CNNs that are trained on SHARP_ONLY as opposed to FUSED_SHARP because the former is observed to

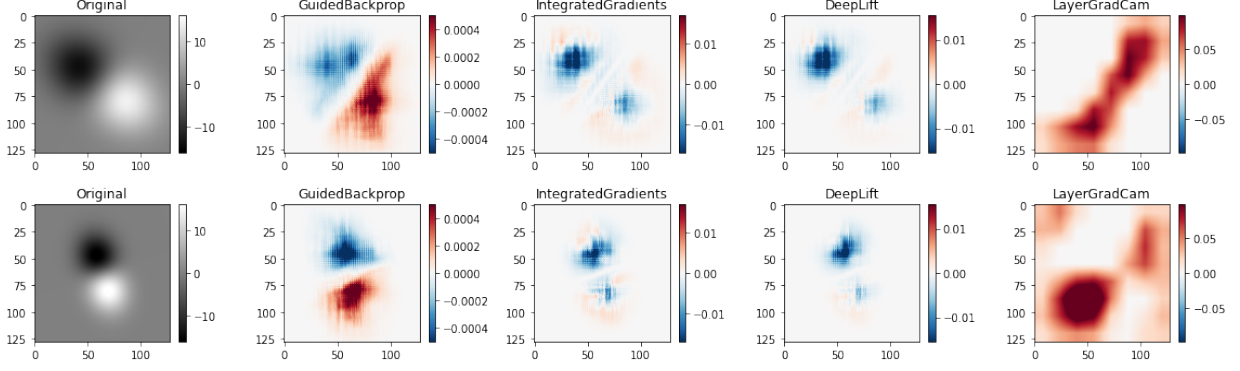


Figure 2.12: Examples of synthetic bipole images and attribution maps.

generalize better according to Section 2.2. To get results that reflect the generalization performance as opposed to training artifacts, we need to make sure that active regions being investigated are out-of-sample. To evaluate any active region of interest in SHARP, we perform 5-fold cross-validation on **SHARP_ONLY**, so that every active region is associated with a CNN that has never seen the active region in training. In addition, we do not enforce the flare-based sample selection rules and random undersampling, so that the evolution of attribution maps can be evaluated more coherently. As case studies, we select four HARP sequences that transition from a flare-quiescent state to a flare-imminent state. Figure 2.13 shows the labels and predicted probabilities of the four sample sequences. The attribution methods are performed on each HARP sequence in a frame-by-frame manner.

Figure 2.14 shows the last image of the four HARP sample sequences. The attribution maps of the same size as the input of the CNN (128×128 pixels) are upsampled to the original resolution of the SHARP magnetogram using the `resize` method of the Python package `skimage.transform` with 2nd-order spline interpolation. The attribution maps of DeepLIFT and Integrated Gradients are similar. As such, only the results of the former are shown. The results for Integrated Gradients can be accessed online with the link shown in the caption.

In Figure 2.14, the attribution maps of Guided Backpropagation are observed to be more concentrated in strong fields compared to that of Deconvolution. The reference image of DeepLIFT and Integrated Gradients are chosen as the first sample in each sequence. From these two methods, the change of the prediction scores is attributed to the change of magnetic configuration of the last frame relative to the first frame, with red pixels indicating positive contribution and blue pixels indicating negative contribution. Since the predicted event probability of the last frame is higher than the first frame for all HARPs (Figure 2.13), the red pixels outweigh the blue pixels in the attribution maps of DeepLIFT and Integrated

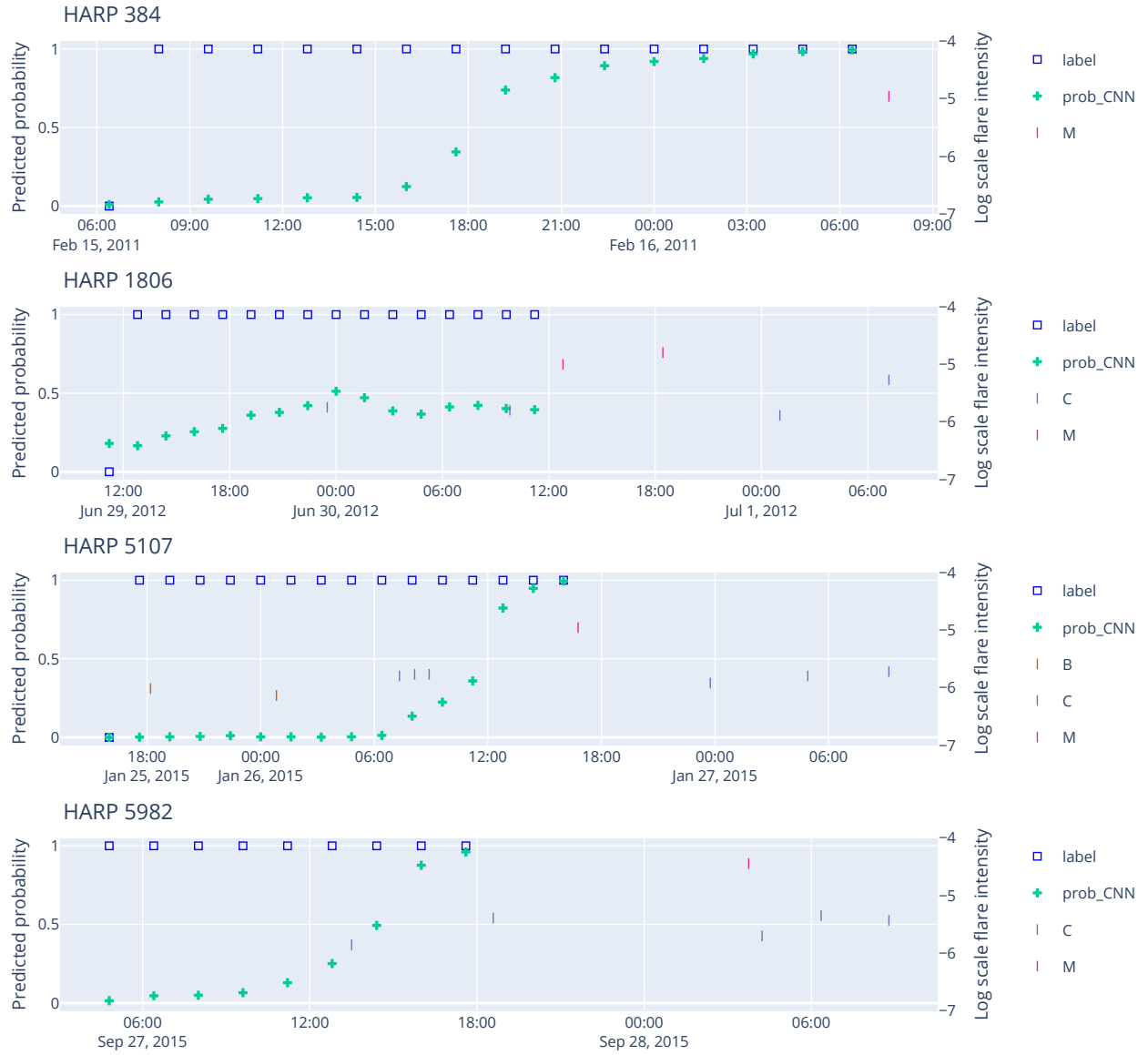


Figure 2.13: CNN predictions of part of time series of in HARP 384, 1806, 5107, and 5982. The labels are shown as blue open boxes and predicted probabilities as green plus symbols. The point-in-time instance is labeled as positive if an M1.0+ flare occurred in the future 24 hours in that active region. GOES flare events during and 24 hours within the sample sequence are shown as short vertical bars, with y-coordinates indicating flare intensities (peak flux in W/m^2) on a log scale.

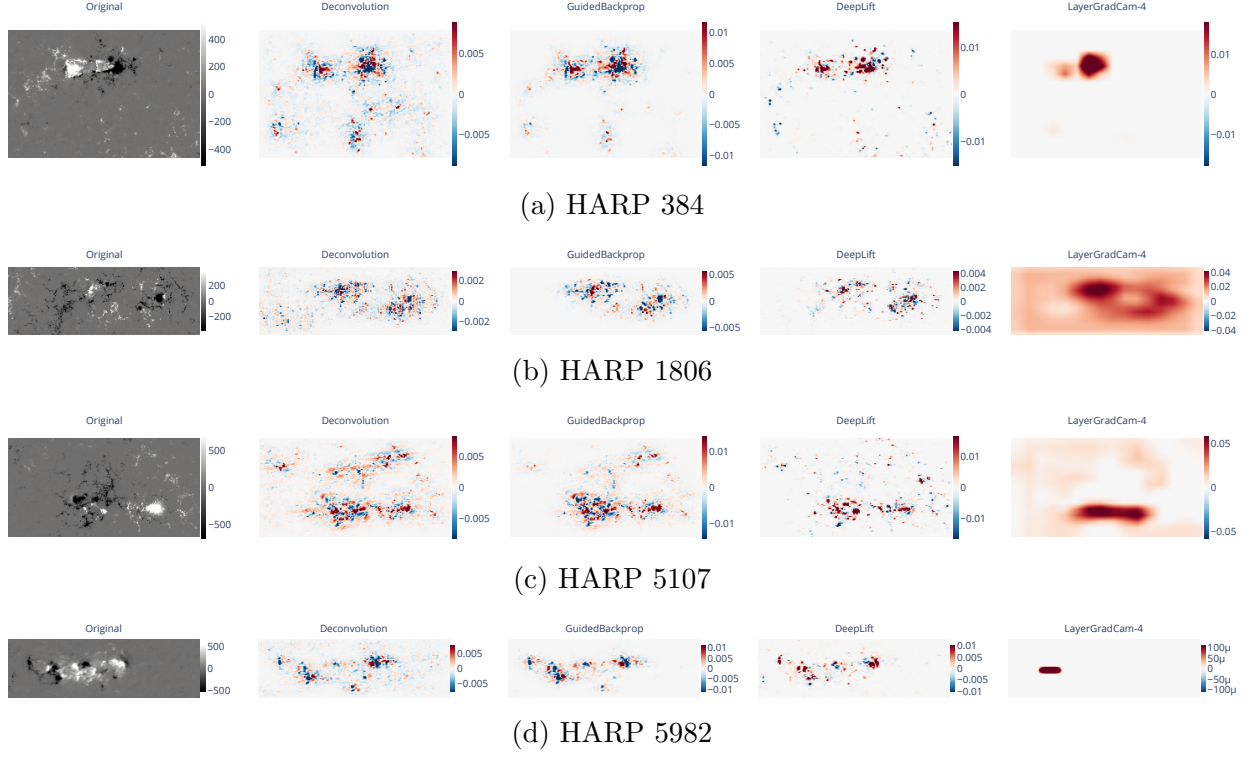


Figure 2.14: Attribution results of Deconvolution, Guided Backpropagation, DeepLIFT, and Grad-CAM on the last magnetogram in the sample sequences of HARP 384, 1806, 5107, and 5982. DeepLIFT chooses the first sample in the sequence as the reference. “LayerGradCam-4” means Grad-CAM with respect to the output of the fourth, or the second to last, convolutional layer. The interactive movie of heatmaps on all 9 samples in HARP 5982 using more attribution methods can be accessed at https://zeyusun.github.io/attribution/captum_movie_first.html.

Gradients. The Grad-CAM results roughly reveal the position of the strong fields and polarity inversion lines.

From the visual attribution map, the CNN’s prediction of a flaring active region can be accredited to the elements in the magnetogram. Figure 2.15 shows the contour plots of attribution maps generated by Integrated Gradients overlaid on magnetograms of the four HARP series. The contours enclose areas with large absolute values of Integrated Gradients in the last frame of each series, with red/blue contours indicating the region contributing positively/negatively to the increase in predicted probability. A general pattern is that the flux is emerging in red contours and canceling in blue contours. From the attribution maps, we can explain the increase in prediction scores as the consequence of the emerging flux outweighing the canceling flux.

The visual attribution maps can not only be used to identify preflare signatures in an

active region; comparing them with our knowledge of flaring active regions can provide insights to diagnose, and potentially improve, the machine learning method used to predict flares. Here we provide an example. A known artifact in magnetograms is the fake polarity inversion line (PIL) caused by the projection effect when the magnetic vector’s inclination relative to the line-of-sight surpasses 90° [172]. In Figure 2.15(d), the emerging polarity inversion line in the penumbra of the leading polarity (on the right/west part of the active region) is picked up as a preflare signature by the largest red contour. However, HARP 5982 is on the limb of the solar disk at the time (Figure 2.16), and the emerging PIL is caused by the highly inclined magnetic field in the penumbra as the flux rope is elevating from the surface. This shows that the CNN trained to associate magnetograms and flaring activities is not able to discern the polarity artifact by itself. This also suggests that the model could be potentially improved if we feed the location information to the CNN to help it correct such artifact. A similar PIL artifact is also observed in the following polarity of HARP 5107 in Figure 2.15(c). Since this artifact does not change much during the observation interval, it does not contribute as much to the change of the prediction score.

We remark the attribution maps obtained by Integrated Gradients are better in terms of resolution and interpretability than what were used in Bhattacharjee et al. [24] and Yi et al. [329]. The occlusion method in Bhattacharjee et al. [24] was shown to highlight the area between the opposite polarities, providing only crude attribution. This is because the size of the occlusion mask is usually chosen to be big enough to cover the informative regions. The result of Grad-CAM, being the attribution to a convolutional layer as opposed to the input, also suffers from the low-resolution issue. Both the Grad-CAM results in Figure 2.14 and in [329] are able to highlight active regions, but the resolution is not high enough to reveal any structural information within the active region at the level of magnetic elements. Guided Backpropagation in Yi et al. [329] is able to identify polarity inversion lines. However, it has been observed (and theoretically assessed) that Guided Backpropagation and Deconvolution behave similarly to an edge detector, i.e., they are activated by strong gradients in the image and insensitive to network decisions [e.g. 216, 3]. In contrast, the method of Integrated Gradients needs a baseline, which aligns with the natural way in which the human interprets an observation: by assigning credit or blame to a certain cause, we implicitly consider the absence of the cause [272]. In addition, Integrated Gradients essentially “decomposes” the change of the network’s prediction score to pixels in the input image, leading to a high-resolution attribution map.

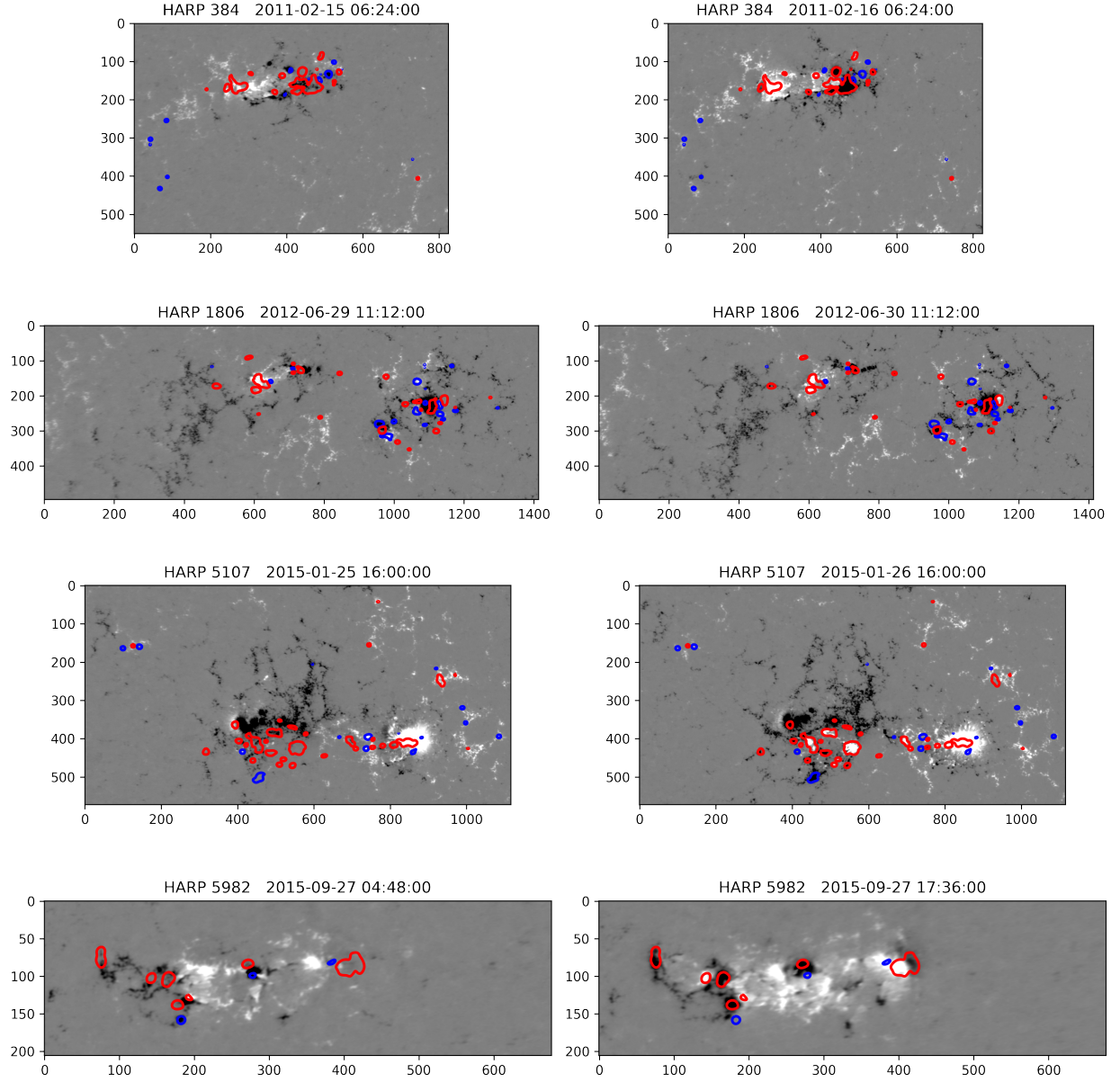


Figure 2.15: Highly attributed pixels in the last frame by Integrated Gradients on four select HARPs shown in rows. In (a), the left/right panel shows the first/last magnetogram in the sample sequence of HARP 384. The magnetograms are in the SHARP resolution, with ticks on the axes indicating pixels. Pixel values saturate at ± 500 Gs. The red/blue contours on the right panel (last frame) highlight the areas with strong positive/negative Integrated Gradients relative to the first frame. The same contours are mapped to the left panel (first frame) for contrast. The contours are drawn on the attribution map smoothed with a Gaussian kernel with a standard deviation of 3 pixels. Figures in (b), (c), and (d) are similar to (a) but for other HARPs. An animated version of this figure that serially presents the complete sample sequences of the four HARPs is available online.

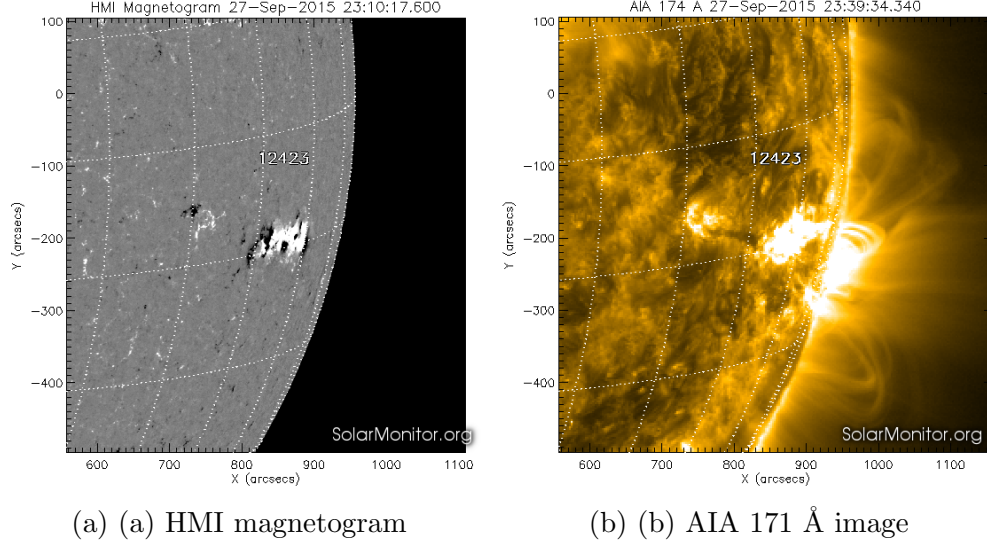


Figure 2.16: Line-of-sight magnetic field (a) and solar EUV image (b) of HARP 5982 (NOAA AR 12423) at 23:10:17 on Sep 27, 2015. Images are taken from <https://solarmonitor.org/>. Note that the image title of (b) should be “AIA 171 Å” instead of “AIA 174 Å”.

2.5 Conclusions and discussion

In this chapter, we used two solar cycles of active region observational data from SMARP [32] and SHARP to examine the improvement in flare predictive performance of two deep learning models, namely the LSTM and the CNN, when trained on the fused datasets. When tested on SMARP, both models showed significant improvement. When tested on SHARP, LSTM showed significant improvement. The results of the controlled comparative studies indicate such an improvement is due to the significantly increased sample size from the other solar cycle. Then, in our setting of flare prediction, we verified the performance of the LSTM and the CNN using skill scores, reliability diagrams, ROC, and skill score profiles. The comparison showed that the LSTM is generally a better model than the CNN. After that, we explored the possibility of combining the LSTM and the CNN for a better prediction performance in the framework of a meta-learning paradigm called stacking. The results showed that in some settings, the stacking model outperforms the best member in the ensemble. Lastly, we applied visual attribution methods to CNNs. The results demonstrate the utility of visual attribution methods in identifying flare-related signatures in active regions, including the flux emergence and new polarity inversion lines. The attribution map on one particular region on the limb of the solar disk revealed one limitation of the CNN and suggested potential modifications for improvement.

The questions raised in Section 2.4 are arguably broad and general. We have taken one

particular path to partially address each question. To inspire future studies, we provide additional comments and discussions related to these questions.

Task-based sample selection In this work, we studied the task of distinguishing M- and X-class flare producing regions from flare-quiet regions. This “strong-vs-quiet” task focuses on the increase or the continuation of flare activity and does not require distinguishing between flares of closer energy levels like M- and C-class flares. The samples that indicate a decay in flare activity and the samples that only lead to weak flares are therefore excluded. This ensures good baseline classification performance against which our proposed predictors could be reliably compared. Our findings may not extend to other task definitions, e.g., all-clear forecasts, where flare-quiet active regions that evolve from a flare-active state are of interest [21, 138]; and operationally-evaluated forecasts, where flare activities in the prediction period are considered as prescience and can not be used to select samples. The principal challenge to extending our analysis to these tasks is that weak- and no-flare activity annotations have higher uncertainty due to background radiation and other factors [201]. The higher levels of “label noise” when including weak flares as negative samples would make learning a reliable predictor substantially more difficult. A possible solution, and topic of future work, is to predict the continuous flare intensity level instead of the GOES flare activity class.

Evaluation under a realistic event rate As mentioned in Section 2.2.5, we rebalance the training and the validation set to prevent the predictor biasing towards the majority non-event class simply due to its volume, and we rebalance the test set to evaluate the predictor’s generalization ability under the same climatological rate as it is trained. Evaluating the performance under a realistic event rate requires more work other than simply applying the predictor on a test set that is not rebalanced: predictors trained on the balanced dataset will bias towards the minority class on a test set under a realistic rare event rate, producing an undesirably high false alarm rate. One possible solution to correct such bias is to treat the class proportions as priors and apply the Bayes rule [79]. This method requires an accurate estimate of the true event rate of the testing period.

The importance and challenges of data fusion Fusing data from multiple sources to produce more consistent, accurate, and useful information is a universal problem in astronomy. Although the astrophysics community is funding projects like DKIST [242] and the Vera Rubin Observatory [137], both of which will take 25–50 TB of data a day, astronomers cannot study long-term trends without including historical or old data sets (or waiting a

decade for these instruments to take enough data). In this work, we took a straightforward approach to add the new data in the training set with minimal calibration, and train the models as usual. Based on our experiments, we have shown that this simple approach can result in improvement. We anticipate that, with more accurate cross-calibration between the SMARP and SHARP, the benefit of combining them may be better than demonstrated here. There are several possible ways to improve upon the fusion method:

- To simulate the effect of unresolved structures in SMARP magnetograms, Gaussian blur can be applied to the higher quality SHARP magnetogram. This approach is used in comparing full-disk line-of-sight magnetograms of HMI and MDI in Liu et al. [189], in which the parameters of the Gaussian filters are tuned to minimize the root mean squared difference between them.
- Point spread functions can be estimated for MDI and HMI magnetograms separately, and deconvolution can be performed to remove stray light that is instrument-specific [199, 328].
- Magnetogram fusion can be performed in the other direction: super-resolving magnetograms in SMARP to mimic those in SHARP. Such an approach has been recently explored using deep neural networks [93, 146]. The improved overall image quality of super-resolved SMARP magnetograms could capture higher resolution magnetic field distributions and hence improve the accuracy of the active region summary parameters in SMARP.
- For the active region summary parameters, we took a “post-facto” correction approach by correcting the parameters of the same name in the two data products via linear regression. Alternatively, with fused magnetograms available, one can also re-compute the parameters on those transformed image data. This approach avoids the linear assumption and leads to parameters more consistent with the manipulated magnetograms, with the caveat that the manipulated magnetograms also suffer from the loss of information. More concretely, the effects of spatial resolution on the inferred magnetic field and derived quantities have been examined by Leka and Barnes [170], who found that, to preserve the underlying character of the magnetic field, post-facto binning can be employed with some confidence, albeit less so for derived quantities like vertical current density. In short, a universal and accurate fusing strategy that accounts for the instrumental spatial resolution is still hindered by our ignorance of the ground truth magnetic field structure, and the benefits and drawbacks of different fusing methods have to be evaluated case by case.

Machine learning with multi-source data Learning from multi-source data is also a prevalent topic in machine learning. In our work, machine learning models are trained as usual with new data added to the training set. An alternative approach would use transfer learning: train on the additional data first, then switch to the original data for fine-tuning. In heliophysics, this idea is recently explored by Covas [57] in the prediction of the solar surface longitudinally averaged radial magnetic field distribution, using historical data from 1874 to 1975 in addition to newer data obtained by SoHO and SDO.

Performance comparison between the LSTM and the CNN The active region summary parameters used by the LSTM are derived from magnetograms. In that sense, the data used by the CNN contains complete information of the data used by the LSTM. However, our experiments show that the LSTM generally has better performance. There are many potential reasons that the CNN does not perform better than, or as well as the LSTM: (1) the CNN takes in uniformly sized magnetograms whose size and aspect ratio are distorted. (2) the CNN only uses the image of the last frame in the sequence, whereas the LSTM uses all the data in the sequence; (3) the CNN learns the features by itself, whereas the LSTM uses hand-crafted parameterizations that are known to be relevant to flaring activity; (4) the CNN uses subsampled images with information loss, whereas the LSTM uses parameters derived from full resolution images; (5) the CNN has more parameters and is more prone to overfitting (which reflects on the lower training loss but not validation loss of the CNN in many experiments).

On the comparison of flare forecast methods Many flare forecast studies quote the skill scores directly from other studies for comparison. Even though the forecast goal is somewhat standard and used in many studies (e.g. to predict whether there will be an M1.0+ class flare occurring in the future 24 hours), to conclude the superiority of one method against another, both methods have to be evaluated on the same dataset. However, it is not trivial to come up with such a “common ground” for methods to compete because research codes are not usually publicly available, and because different opinions exist on the ways the data should be processed. The difficulty in methodical comparison spawns the effort in fairly comparing existing forecasts (e.g., the “All-Clear” workshops [21, 173]) and developing common datasets (e.g., SWAN-SF benchmark dataset [13]) or platforms (e.g., the FLARECAST project [92]). To advocate credible comparisons, we do not quote skill scores from other studies. Instead, we follow the above studies and make our code publicly available to facilitate future comparisons.

On stacking ensemble In our experiments, stacking CNN and LSTM performs similarly to the “select best” strategy but not significantly better in most settings. However, another stacking study Guerra et al. [106] used a larger number of base learners and showed that most ensembles achieved a better skill score (between 5% to 15%) than any of the members alone. This suggests that improved performance may be obtained by training a larger number of base learners on the SMARP/SHARP data studied here.

Choice of the baseline in interpretability methods Some visual attribution methods require reference input, such as Integrated Gradients and DeepLIFT. One naive choice is an image with all values equal to zero. Images of this sort imply a lack of patterns. These are the baselines mostly used for interpretation in computer vision tasks like object detection. In our case, the images are magnetic field component measurements, which can take on positive or negative values and a wide dynamic range, unlike normal images in real life. We choose the first image in the sequence as the reference, so that the visual attribution methods can attribute the change of prediction scores to the change of magnetic field configuration, which is of actual interest. There are other choices of baselines. One example is input images with Gaussian noise. Using this type of reference may reveal the sensitivity of the network’s prediction to local changes. Furthermore, integration may benefit from going beyond simply linearly interpolating the reference and the input on the original image space, i.e., the 2D cartesian plane. For example, one could consider applying attribution methods to the path of time series of magnetograms. The Integrated Gradients calculated with this approach would integrate temporal dependency of each point-in-time in the sequence, exploiting more information about the evolution of active regions.

CHAPTER 3

Incorporating Polar Field Data in Flare Prediction

3.1 Introduction

Solar flares rapidly release magnetic energy stored in the Sun’s corona, which heats large areas of the surrounding atmosphere. The terrestrial impacts of flares can be serious, especially when they are high-intensity. Although weak flares have a limited impact on mankind, strong flares can disturb satellite communication and GPS-navigation systems, in addition to causing dangerous ionizing radiation exposure to astronauts orbiting the earth. Hence, timely prediction of solar flares, especially the strong ones, is desirable so that precautions can be taken. However, strong flares occur rarely as compared to weak flares, and the underlying mechanisms and the factors that trigger and determine the strength of the flares are not well understood. This makes the prediction of solar flares a complex task.

As the coronal field is anchored in the photosphere, the onset of flares has been linked to changes in the photospheric magnetic field [301, 268]. Consequently, it is reasonable to expect that photospheric magnetic fields in active regions (AR) may be useful indicators of flare-triggering mechanisms. The Helioseismic and Magnetic Imager (HMI) [250, 127] on Solar Dynamics Observatory (SDO) [226] spacecraft has observed the photospheric magnetic field of the sun since 2010, taking high-resolution images with 1 arcsec spatial resolution and 720s cadence. From these magnetic field images, useful features are computed, known as Space-weather HMI Active Region Patches (SHARP) parameters, which are representative of the magnetic field’s scale, energy and complexity [31]. This time-series data has been successfully used for solar flare prediction for Solar Cycle 24 using machine learning algorithms [49, 145, 30, 217, 83, 185, 41, 154, 40]. See also [173] for a comprehensive review.

Several supervised machine learning models have been proposed to predict solar flares from HMI observations. NOAA Geo-stationary Operational Environmental Satellites (GOES) flare list [89] provides peak energy levels for each flare on a logarithmic scale separated into

five flare intensity classes, labeled A, B, C, M, X, with A the weakest and X the strongest intensities. These can be used as categorical response variables in machine learning models for flare prediction. Such models use input features, such as HMI SHARP parameters, to apply off-the-shelf machine learning algorithms to predict the labels of the flares, including discriminant analysis [167], linear regression [145], support vector machine [30, 217, 83], k-nearest neighbor [217], random forests [185, 83], multilayer perceptrons [83], and recurrent neural networks [49]. Recently, the magnetogram images of the active regions have been used as features in deep learning models such as convolutional neural networks (CNN) [[135], for prediction of flare intensity class. Ensemble models have also shown promise as a way of combining predictions of multiple models that use different data sources [107, 270].

Hale’s Polarity laws [116] for sunspots apprise that the polarities of the sunspots (distinguished in pairs with a leading/following counterpart from east to west) have a strong tendency to have the opposite pattern in the northern and southern hemispheres of the sun in a particular cycle. Moreover, the polarities of the spots of the present cycle are the opposite of the polarities in the previous cycle. Additionally, polar field reversals occur systematically at approximately the time of the solar cycle maximum [18], with peak polar field values occurring at the time of the sunspot cycle minimum. Thus, the polar fields are 180 degrees out of phase with the sunspot cycle. See Fig. 3.1 for plots over time of the polar fields along with curves of sunspot number, unsigned AR flux, and AR latitude. Numerous works show how the polar field is reformed each cycle from a small fraction of active region flux, which is transported by meridional flows from mid-latitudes to the poles [120, 303, 206].

It has been demonstrated that the strength of the sun’s polar fields can be used to predict the next solar cycle’s sunspot intensity based on the understanding of the sun’s magnetic dynamo [292, 71, 125]¹. Weak polar fields at the end of Cycle 23 motivated the authors of the 2005 paper [274] to forecast a weak Solar Cycle 24. Indeed, Solar Cycle 24 has been the weakest solar cycle in a century. To graphically illustrate the strong association between polar fields and sunspot activity, we the polar fields and monthly mean sunspot number over four solar cycles in the top panel of Fig. 3.1. It can be observed the magnitude of the difference between the two polar fields at the solar minimum is a good indicator of the intensity of the subsequent cycle.

The relation between polar fields and AR magnetic fields is more intricate, involving complex physical mechanisms. This is better observed with higher temporal resolution and in an AR-localized manner. The middle and bottom panels of Fig. 3.1 shows how polar fields interact with total unsigned flux and latitudinal distribution of ARs in the northern and the southern hemisphere, respectively, during the period that HMI has been in operation

¹<https://ntrs.nasa.gov/api/citations/20130013068/downloads/20130013068.pdf>

(the shaded area in the top panel of Fig. 3.1). The data are collected from the data product `hmi.meanpf_720s`, available at the Joint Science Operations Center (JSOC). The north and south polar field strengths are represented by CAPN2 and CAPS2, respectively, which correspond to the mean radial field within latitudes N60-N90 and the mean radial field within latitudes S60-S90. The unsigned AR flux can be well approximated by the unsigned flux in between S40 and N40, since flux is concentrated in ARs which are mostly located in those latitudes. Specifically, for the northern hemisphere, we compute the unsigned AR flux (USFLUXN) and unsigned-flux-weighted latitude (USFLATN) via

$$\text{USFLUXN} = \text{FLUXN_P} - \text{FLUXN_N}, \quad (3.1)$$

$$\text{USFLATN} = \frac{\text{FLATN_P} \times \text{FLUXN_P} - \text{FLATN_N} \times \text{FLUXN_N}}{\text{FLUXN_P} - \text{FLUXN_N}}, \quad (3.2)$$

where FLUXN_P (FLUXN_N) is positive (negative) flux between N0-N40 above a fixed threshold, and FLATN_P (FLATN_N) is positive (negative) flux weighted latitude between N0-N40. Similar quantities are computed for the southern hemisphere as well. The active regions' locations follow the well-observed butterfly diagram of sunspots, starting approximately ± 40 degrees latitude at the beginning of a solar cycle, and gradually drifting towards the equator as the cycle progresses. At the cycle minimum, there is an overlap of the old (near the equator) and new (near ± 40 degrees) cycle spots. These trends can be observed from the lower two panels of Fig. 3.1. It is worth noting that the rapid transition in AR latitude at the depth of solar minimum provides a strong signature of the solar cycle when the total flux and polar fields show the lowest variation with time. These strong and complex relations between polar fields and AR fields suggest supplementing SHARP features with HMI polar field strength may improve the prediction of AR-localized flare intensity levels. This is our main motivator for applying machine learning flare prediction approaches to the combined data.

In summary, we validate the usefulness of polar field features in predicting solar flares. The motivation is to assess whether combining SHARP keywords with polar field features can improve solar flare prediction performance or not. To this end, we design an experiment in which there are multiple possible ways of incorporating polar field data, four different machine learning models, and different train-test-validation splitting methods for each algorithm. In addition, we design a novel mixture of experts model tailored for incorporating polar field data for solar flare prediction context. Our experimental results indicate that polar field data can improve solar flare prediction performance. They also show that mixture of experts modeling is a simple yet effective alternative for combining polar field data and SHARP parameters, which can give similar or better performance to much more complicated

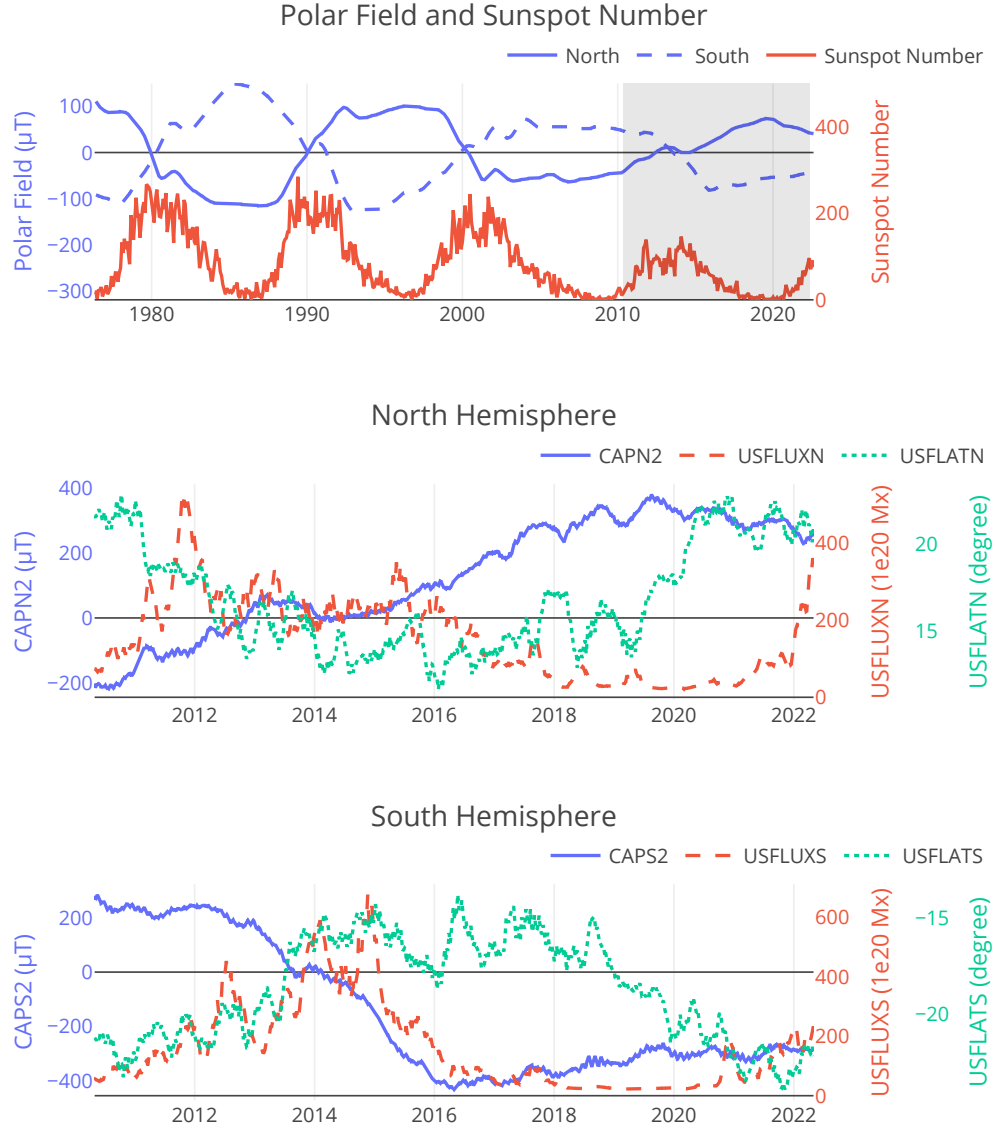


Figure 3.1: *Top*: The north and south polar strength and the monthly mean sunspot numbers plotted over 4 solar cycles. The shaded period is the time during which the HMI instrument has been generating data. Polar fields data are from the Wilcox Solar Observatory [273, 128]. Sunspot number data are from WDC-SILSO, Royal Observatory of Belgium, Brussels. *Middle*: Solar cycle variation of the north polar field strength (CAPN2), AR unsigned flux (USFLUXN), and flux-weighted AR latitude (USFLATN) obtained from HMI data. The unsigned flux in the northern hemisphere (USFLUXN) is positively correlated with the north polar fields (CAPN2), suggesting the inclusion of polar field data as additional features in a machine learning flare prediction model can potentially improve prediction accuracy. *Bottom*: Similar to the Middle panel but for the Southern Hemisphere of the sun.

neural network-based algorithms.

The rest of this chapter is organized as follows. In Section 2, we describe how we incorporate multiple data sources into flare prediction algorithms, and we introduce several data-splitting methods for assessing the performances of the models. In Section 3, we formally define the machine learning models that are compared in this study including the new mixture of experts model. In Section 4, we describe the experimental setup in detail and present our results.

3.2 Data Preparation

There are three distinct data sources we exploit in this study, namely, SHARP, GOES, and SPF datasets. The SHARP dataset provides parameters computed from the magnetograms produced by the HMI (See Table 3.3) instrument. These parameters are associated with the active regions that occurred from 2010 to 2020. We use the data product `hmi.sharp_cea_720s`, which contains definitive data in Cylindrical Equal-Area (CEA) coordinates and is available at the Joint Science Operations Center (JSOC). The GOES dataset includes solar flare event records measured by Geostationary Operational Environmental Satellites (GOES). This dataset can be downloaded using the SunPy package² and it covers the HMI time period from 2010 to 2020. The flare events are labeled by A, B, C, M, and X, based on the peak intensity of X-ray flux measures. The Solar Polar Fields (SPF) dataset consists of two solar polar field features, namely, CAPS2 and CAPN2. These features correspond to the mean radial field in latitudes N60-N90, and the mean radial field in latitudes S60-S90, respectively. Both HMI and MDI instruments record these features, which can also be downloaded from JSOC. The MDI instrument was operational from 1996 to 2010 and the HMI has been operational since 2010.

3.2.1 SPF Preprocessing

The noise levels on CAPS2 and CAPN2 features are significant, and the noise level is different in the HMI and MDI instruments. Hence, we consider smoothing by using a Kalman filter to filter out noise over these features. Note that the variance over the time series data is not solely due to noise. There is strong yearly modulation due to the Earth-solar B0 angle, where the solar rotation axis varies by ± 7 degrees. Furthermore, there is strong 27-day modulation due to solar rotation.

²<https://sunpy.org/>

Dataset	# of instances	B-class ratio	MX-class ratio	Parameters
SHARP	2663	0.49	0.51	USFLUXL, MEANGBL, RVALUE, AREA, TOTUSJH, TOTUSJZ, SAVNCP, ABSNJZH, TOTPOT, SIZEACR, NACR, MEANPOT, SIZE, MEANJZH, SHRGT45, MEANSHR, MEANJZD, MEANALP, MEANGBT, MEANGAM, MEANGBH, NPIX

Table 3.1: SHARP dataset statistics after preprocessing step. See Table 3 in the appendix for the definitions of the keywords.

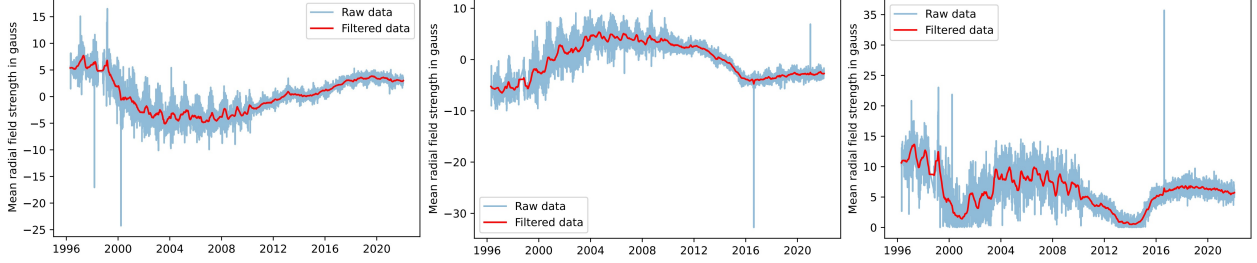


Figure 3.2: Raw (blue curve) and Kalman filtered (red curve) polar field features. Polar field strengths at the north pole CAPN2, south pole CAPS2, and the solar cycle approximation, defined as the unsigned polar field difference $|CAPS2(t) - CAPN2(t)|$, are shown in the left, middle, and right-hand panels respectively. The data source switched from MDI to HMI in 2010, which is accompanied by a visually obvious reduction in noise. The HMI data source has lower noise due to its high resolution.

We implement the Kalman filter using the local level model [212] where the state transition and observation models are defined as follows:

$$a_t = a_{t-1} + \epsilon_t^a, \quad (3.3)$$

$$y_t = a_t + \epsilon_t^y, \quad (3.4)$$

where y_t is the raw noisy observation at time t , a_t is the filtered data, and $\epsilon_t^y \sim \mathcal{N}(0, \lambda_y)$ and $\epsilon_t^a \sim \mathcal{N}(0, \lambda_a)$ are the noise processes over the observed and hidden spaces with variances λ_y and λ_a , respectively. By adjusting these levels instrument-wise, we perform smoothing for the data of both instruments. See Fig. 3.2 for the raw and filtered features.

3.2.2 Observation Preprocessing

In this section, we describe how we use the HMI SHARP dataset and the GOES flare intensity labels to construct the training and test data for our discriminative machine learning algorithms, developed in the next section. Our Preprocessing steps for these data follow the approach in [141, 49, 270].

We only include ARs having unique NOAA AR numbers [141] into our training set. We eliminate low-quality observations by excluding ARs that occur outside of range ± 70 relative to the central meridian [49]. Similarly to [31, 49] we focus on predicting whether a solar flare from a given active region will be strong (M/X class) or weak (B class). The A-class flares are excluded as they are almost indistinguishable from the background radiation. The C-class flares are frequently missing their NOAA AR numbers in the GOES database, making their locations difficult to track [3]. Exclusion of A and C flares and combining M and X class flares together creates a binary classification problem that enables the comprehensive comparative computational analysis we report in Sec. 4. This is a step along the way of establishing the value of polar field data for the more complex classification task of classifying between the five labels A, B, C, M, X, which is worthwhile of future study.

As in [141], the class label of an AR is designated as the strongest flare occurring in a 24-hour prediction period, according to the GOES database. Thus an AR is assigned a B label if there are no M/X flares within 24 hours, while it is assigned label M/X otherwise. This is accomplished as follows. For each of the flares occurring in an AR, variables are extracted from the time segment spanning a 24-hour time period from 48 hours to 24 hours before the occurrence of the flare. Since the cadence of the HMI observations is one sample every 12 minutes this yields a matrix of 22 SHARP predictor variables defined over 120-time points. If, for a given flare, the corresponding predictor variables have NaN values no class prediction is generated and we omit the flare from our analysis. Furthermore, in our analysis, we do not use any flare labels in the first 24 hours of the beginning of an ARs since such flares would have no associated observed predictor variables. Once these omissions are taken into account, the total number of observations (predictor variables-label pairs) is 2663. See Table 3.1 for a summary of our set of SHARP features.

3.2.3 Train-Validation-Test Splitting

We consider three different methods for splitting the set of AR samples into train-test sets. In all three methods, we split the AR's into 80% train and 20% test samples. i) Instance-based splitting. The indexes of the splits are generated randomly without replacement. This split can measure the goodness-of-fit but is prone to information leakage when the observations of

a particular AR are scattered among the splits [270]. ii) AR-based splitting. AR numbers are used instead of the observation indexes for random index generation. This split prevents the information leakage that happens in Random splitting but may introduce another form of information leakage due to sympathetic flaring between neighboring ARs [270]. iii) Month-based splitting. The timestamps are used to split the sets in terms of months with the same ratio. This split can also measure the goodness of fit and is closer to the operational forecasting setting. Each feature in the dataset is normalized over the time windows of the training data. Specifically, for each AR the mean and standard deviation of each feature is computed over the time points of the training sample. Then for each AR and each variable, the mean is subtracted and the result is divided by the standard deviation for both training and test samples.

3.3 Prediction algorithms

We consider four machine learning models that can exploit polar field features for the flare prediction task: i) Logistic Regression (LR); ii) Mixture of Experts (MOE); iii) Multi-Layer Perceptron; iv) Recurrent Neural Network.

3.3.1 Logistic Regression

Logistic Regression (LR) is a model for which the conditional distribution of the response variable as a Bernoulli distribution conditioned on the features of an AR sample in the case of binary classes, and categorical (multinomial) distribution for multiple classes. Specifically, for the binary case, LR models the conditional distribution of the response variable as

$$p(y_i|\mathbf{x}_i, \Theta) = \text{Ber}(y_i|\sigma(\Theta^T \mathbf{x}_i)), \quad (3.5)$$

where i indexes the data points, $y_i = \{0, 1\}$ denotes the response variable (i.e., flare class), and $\mathbf{x}_i$ corresponds to the SHARP parameters. The model parameters Θ are estimated from the training set by using Maximum Likelihood Estimation (MLE). In particular, the MLE objective over the N AR samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ in the training set:

$$\ell(\Theta) = \sum_{i=1}^N \log p(y_i|\mathbf{x}_i, \Theta), \quad (3.6)$$

which is convex in Θ , and has a unique solution. We use LBFG-S [186] to optimize the model parameters. In order to introduce polar field features to LR, a simple approach

is to concatenate them with the SHARP parameters and apply the algorithm with these augmented features.

3.3.2 Mixture of Logistic Regressions

A richer family of conditional distributions can be obtained by using a mixture of experts (MOE) framework [336]. For classification tasks, each expert is an LR model. The predictions of the experts are mixed probabilistically according to mixture coefficients, which are obtained by using another linear model called the gating function. The output of the gating function is constrained to be a simplex to support proper mixing coefficients. This is achieved by using a Softmax function $\mathcal{S}$, which is applied after the linear transformation of the input features of the gating function. We propose to use polar field features as the input to the gating function. The main intuition is that the experts will be forced to focus on specific portions of the solar cycle.

Let $\mathbf{z}_i$ denote a K dimensional categorical valued latent variable that indicates which expert is active for data point (AR) i , where K is the total number of experts initialized. The categorical conditional distribution of the latent variable $\mathbf{z}_i$ induced by the gating function is given as

$$p(\mathbf{z}_i|\mathbf{t}_i, \mathbf{V}) = \text{Cat}(\mathbf{z}_i|\mathcal{S}(\mathbf{V}^T\mathbf{t}_i)), \quad (3.7)$$

where $\mathbf{t}_i$ denotes the polar field features associated with data point i , and $\mathbf{V}$ represents the gating function parameters that map features to the logits of the categorical distribution. After the softmax function, we obtain proper mixture coefficients for data point i . Since the experts are logistic regression models, their conditional distributions are given as

$$p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i = k, \Theta_k) = \text{Ber}(\mathbf{y}_i|\sigma(\Theta_k^T\mathbf{x}_i)), \quad (3.8)$$

where Θ_k denotes the parameters of the expert k . The marginal likelihood of this model can be computed by integrating out the latent variables as follows

$$p(\mathbf{y}_i|\mathbf{x}_i, t_i) = \sum_{k=1}^K p(\mathbf{z}_i = k|\mathbf{t}_i, \mathbf{V})p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i = k, \Theta_k). \quad (3.9)$$

The objective is to maximize marginal log-likelihood with respect to the logistic regression and gating function parameters. The Expectation Maximization algorithm is a suitable optimization method for latent variable models. The algorithm optimizes the expected complete data log-likelihood (ECLL) instead of marginal, where the former forms a lower bound for the latter. There are two alternating steps, E and M steps. In the E-step of

the algorithm, the posterior distributions of the latent variables are computed and their expectations are plugged into the complete data log-likelihood to obtain ECLL. Subsequently, the ECLL is maximized with respect to the model parameters in the M-step. These steps are iterated until ECLL converges. For the mixture of experts model, the ECLL is given as

$$\sum_{i=1}^N \sum_{k=1}^K r_{ik} \log[\mathcal{S}_k(\mathbf{V}^T t_i) p(y_i | \mathbf{x}_i, \Theta_k)], \quad (3.10)$$

where we replace the explicit expression of the prior distribution of mixture coefficients. r_{ik} denotes the posterior distribution of z_i , which can be computed as

$$r_{ik} = p(z_i = k | \mathbf{x}_i, y_i, t_i, \Theta_k) \propto \mathcal{S}(\mathbf{V}^T t_i) p(y_i | \mathbf{x}_i, z_i = k, \Theta_k), \quad (3.11)$$

in the E-step of the EM algorithm. Replacing the posterior in the complete data log-likelihood, we get parameter-specific objectives to optimize in the M-step. Specifically, for the model parameters of the experts, the objective becomes

$$\ell(\Theta_k) = \sum_{i=1}^N r_{ik} \log p(y_i | \mathbf{x}_i, z_i = k, \Theta_k), \quad (3.12)$$

which can be computed by using any gradient-based optimizer. Note the similarity of this objective with Eq. 3.6. Hence, the former corresponds to weighted Logistic Regression. For the parameters of the gating function, the objective takes the following form

$$\ell(\mathbf{V}) = \sum_{i=1}^N \sum_{k=1}^K r_{ik} \log \mathcal{S}_k(\mathbf{V}^T t_i), \quad (3.13)$$

which is also similar to the Logistic Regression objective in Eq. 3.6 except that the response variable is simplex instead of discrete. This objective can also be maximized with any gradient-based optimizer. We use LBFG-S for all gradient-based optimization tasks. In summary, the E-step constitutes computing posterior distributions of the latent variables given the current values of the parameters by using Eq. 3.11, the M-step constitutes maximizing Eq. 3.12 and 3.13 with respect to Θ_k and $\mathbf{V}$. When Eq. 3.10 converges, the iterations are halted. If we don't use polar field features, mixing coefficients are unconditional as analogous to Gaussian Mixture Models.

3.3.3 Multi-layer Perceptron

The Multi-layer Perceptron (MLP) is a non-linear discriminative classification model. MLPs can be more powerful than their linear counterparts when the interactions between the features and the response variables can not be explained by linear decision boundaries. An MLP can be obtained by stacking multiple fully connected layers. Each layer performs an affine transformation, and subsequently applies a deterministic non-linear function, e.g., softmax, relu, tanh. Stacking multiple layers result in a complex deterministic function of the input features. Let f_{Θ} denote this function with the exception that the last layer does not apply any activation functions. Here, Θ represents the set of the affine transformation parameters of all the layers. Then, the conditional distribution for a binary classification task is given as

$$p(y_i|\mathbf{x}_i, \Theta) = \text{Ber}(y_i|\sigma(f_{\Theta}(\mathbf{x}_i))), \quad (3.14)$$

where the loss function is the same with LR as in Eq. 3.6.

The maximization of this distribution can be achieved by using a back-propagation algorithm [122]. We consider using polar field features as the additional features by concatenating them with the SHARP parameters.

Analogously to LR, a mixture of MLPs can be constructed by initializing multiple MLPs and mixing them using the probabilistic objective given in Eq. 3.9 [27]. However, in the MLP case, the EM algorithm is no longer applicable. Fortunately, back-propagation can still be used to maximize the objective. Similar to the Mixture of Logistic regression, a separate gating function can be initialized that maps polar field features to the mixture coefficients. This function can be linear or non-linear (MLP). In our experiments, we did not observe a significant performance change for Mixture of MLPs over NN-Augmented. Hence, we omit the Mixture of MLPs in this study.

3.3.4 Recurrent Neural Network

The aforementioned algorithms do not take the sequential nature of the observations into account. A more principled approach would model temporal data by assuming Markov chains between the consecutive inputs or outputs. Recurrent Neural Networks (RNN) do that by forming a first-order Markov chain between the hidden units. RNNs have shown significant success for sequential data modeling, and LSTM, a variant of RNNs, achieves state-of-the-art performance in flare forecasting from time series SHARP data [49, 188, 270]. For this study, we use a particular version to classify the sequences of the input observations, also known as time series classification. The conditional distribution of the model is given

as:

$$p(y_i|\mathbf{x}_{i,t=1:T}, \Theta) = \text{Ber}(y_i|\sigma(f_{\Theta}(\mathbf{x}_{i,t=1:T}))), \quad (3.15)$$

which implies that the observed label is dependent on a sequence of input data points. The length of the sequence is denoted as T and is a hyper-parameter of the model. The log-likelihood function associated with the RNN model is

$$\ell(\Theta) = \sum_{i=1}^N \log p(y_i|\mathbf{x}_{i,t=1:T}, \Theta), \quad (3.16)$$

where Θ represents the parameters of the model and can be optimized by using the back-propagation algorithm and $T = 120$ is the number of temporal samples in each AR training sample. The polar field features are incorporated as additional input features by concatenating them with the SHARP parameters.

3.4 Experiments

3.4.1 Configurations

The 4 algorithms described in Sec. 3.3 were compared in terms of relative performance improvements when adding the polar field to the SHARP features. Our SHARP dataset consists of 2663 samples with 22 parameters over 120 time points. As described above we consider 3 different data splits: instance-based, AR-based, and month-based. For polar field features, we create two modes: raw CAPS2 and CAPN2 used as two separate input features; and cycle, where the absolute difference of CAPS2 and CAPN2 is used as a lower dimensional solar cycle approximation (see Fig. 3.3). We also consider smoothed versions of CAPS2 and CAPN2 by using the Kalman filter, and unfiltered versions of these features. Hence, the total number of experimental configurations is 12 if we don't use the polar field feature, and 48 otherwise. We perform cross-validation by using training splits for each configuration to select the regularization strengths of the algorithms. Also, we perform 20 runs to assess the significance of the accuracies. Hence, the total number of runs is around 7k.

3.4.2 Tuning

The hyper-parameters of the machine learning algorithms must be tuned for fair performance comparison. Cross-validation on the training set is used to select the best hyper-parameters for each experimental configuration. First, the training dataset is divided into 5 folds. The first fold is held out for measuring the performance of a hyper-parameter setting while the

rest are used for training the algorithm. Then the hold-out set is alternated one by one so the total number of empirical realizations of performance for each hyper-parameter setting is 5. The average of these 5 values is then used as the score of that particular setting. We use the Heidke Skill Score HSS2 [55, 30] for quantifying performance, which is defined as follows:

$$HSS2 = \frac{2 \times [(TP \times TN) - (FN \times FP)]}{(TP + FN)(FN + TN) + (TP + FP)(FP + TN)}, \quad (3.17)$$

where TP, TN, FP, and FN correspond to true-positive, true-negative, false-positive, and false-negative values computed from binary classification predictions and ground truth, respectively.

All the algorithms studied were implemented with the standard L2-norm regularization approach to minimize overfitting. Regularization is accomplished by a penalty method, where the sum of the L2-norm squared of each parameter vector is weighted by a regularization parameter and added to the log-likelihood functions of each of the predictor models discussed in Sec. 3.3. The regularization parameters are tuned to optimize the performance of each algorithm. For LR, MLP, and RNN, the regularization strength is a single scalar that is common for all the elements of the parameters. The grid search range for this scalar is chosen as $[0.1, 1, 10]$. For MOE, we have separate parameters for the gating function and the experts. We use one scalar regularization parameter for the parameters of the gating function, and another for the parameters of the experts. The grid range for these two hyper-parameters is similarly determined as $[0.1, 1, 10]$.

In addition to regularization tuning, the mixture model has a hyper-parameter associated with the total number of experts being initialized. The grid search range for this hyper-parameter is set as $[2, 3, 4, 5, 6, 7, 8]$. We fix the network structure of RNN and MLP. In particular, 32 nodes with one hidden layer are initialized for both models.

3.4.3 Significance Assessment

To assess the significance of the results, we repeat each experimental configuration $n = 20$ times with and without using polar field data. We then compute p-values from one-sided paired t-tests. The t-statistic is computed as $t = \bar{d}\sqrt{n}/s$, where $\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i$ is the mean of the difference vector $\mathbf{d}$ whose elements are given by $d_i = y_i - x_i$ where y_i and x_i are HSS2 scores with and without polar field data for trial i , respectively. Then, under the Student-t distribution with degrees of freedom $n - 1$, we compute the right-tail p-value.

Algorithm	Filter	Mod	Split	p-value	t-score	hss2(s+p)	hss2(s)	hss2-diff	hss2(p)
LR	Yes	Raw	Month	0.0274	2.628	0.7745	0.7155	%8.2	0.3010
	No	Cyc	Inst	0.0439	2.1271	0.8143	0.802	%1.5	0.3628
MOE	No	Cyc	Inst	0.0002	4.3055	0.8223	0.7992	%2.9	0.3484
	Yes	Raw	Inst	0.0005	4.0369	0.8254	0.792	%4.2	0.3867
	Yes	Cyc	Inst	0.0155	2.6055	0.8205	0.8001	%2.5	0.3689
	Yes	Raw	Month	0.0386	2.4202	0.7502	0.6827	%9.9	0.2485
MLP	Yes	Raw	Inst	0.0000	6.8479	0.7992	0.7896	%4.9	0.4384
	No	Cyc	Month	0.0481	2.2856	0.7653	0.7149	%7.0	0.3658
RNN	Yes	Raw	Inst	0.0002	4.4235	0.826	0.7947	%3.9	0.4143
	Yes	Cyc	Month	0.0463	-2.3092	0.6117	0.6756	-%9.4	0.3627
	Yes	Raw	Month	0.0478	2.2888	0.7194	0.6544	%9.9	0.2791
	No	Raw	AR	0.0221	2.4469	0.6977	0.6337	%10.1	0.3952

Table 3.2: Table of statistically significant ($p\text{-value} \leq 5\%$) performance improvements due to adding polar field features to SHARP features, The hss2 skill score argument (p) means only polar fields data, (s) means only SHARP parameters, and (s+p) means both data sources are used as input to the algorithms. (diff) is the mean hss2 score percentage difference when the polar field is incorporated in addition to SHARP parameters. The highlight colors in the table specify the common configurations where MOE achieves statistically significant performance improvements due to adding the polar field. The full tables including the statistically non-significant configurations are provided in the supplementary.

3.4.4 Results

The results are compiled in Table 3.2. We report HSS2 computed on the test set as the performance metric. For each algorithm, we only keep the configurations with statistically significant performance improvement due to the incorporation of polar field data, where significance is defined as a p-value less than 0.05 (5% level) of a one-sided paired t-test of equal mean performance, computed from 20 runs. For the LR model, there are two configurations that exhibit significant performance differences when polar field features are incorporated. Both of them imply that performance improves with the incorporation polar field. Particularly when we use unfiltered cycle approximation for the polar field features and filtered raw features, we obtain performance boost for the month (8.2%) and instance-based splits (1.5%), respectively. For the MOE model, there are four significant cases with all of them yielding performance improvements. The performance boost is up to 9.9%. The model can exploit the polar field augmented data regardless of filtering and preprocessing options when the splits are instance and month based to improve flare classification performance. For filtered raw features, the MLP improves performance on instance-based splitting by 4.9% and non-filtered cycle approximation on month-based splitting improves by 7.0%. The RNN performance is boosted regardless of filtering, preprocessing, and splitting options for four

significant cases, achieving improvements as high as 10.1%. In summary, out of 12 significant cases for four algorithms, 11 of them provide performance increase, which suggests that polar field features to have high potential to improve solar flare prediction performance.

The column hss2(p) shows the prediction performance of the algorithms when only polar field data is used. The results indicate that sole use of polar field data gives poor skill scores, which is less than half of those of the SHARP data alone, for solar flare prediction. However, the hss2(p) scores are far from zero since solar cycle intensity and polar fields are correlated, as discussed in Section 3.1. Solar cycle intensity affects the expected intensities of the flares in particular time frames. Due to this connection, one can expect that coarse predictions about the classes of solar flares can be made with polar field data since it captures information about the solar cycle.

For the MOE model, the solar cycle dependency of the classification accuracy is shown in Fig. 3.3 by repeating five experiments using instance-based splitting. We plot sunspot numbers, polar fields data (cycle approximation), accuracy without polar fields data, and accuracy with polar fields data. The classification accuracy curves are smoothed out with a 90-day moving average. We observe that MOE can exploit polar field data to improve classification accuracy in some parts of the solar cycle. The average classification accuracies over the whole HMI period are 0.88 and 0.91 for without and with polar field data, respectively. One can also observe the high negative correlation between the sunspot number and the curve of the polar field, which is computed as -0.86.

Finally, we compare the performance improvements of the models in Table 3.2. Only those configurations for which the p-value was significant at a level of 5% are shown. Among all algorithms MOE stands out as the only one to attain highly significant performance improvement (p-value=0.0002) when no Kalman filtering is used, the magnitude difference north and south polar field data are used (the solar cycle approximation), and the splitting is instance based. MOE, MLP, and RNN, all have highly significant performance improvement when the polar field data is smoothed using the Kalman filter, the northern and southern polar fields are not combined into a cycle approximation (raw) and the split is instance based. Overall, among the five common configurations for which MOE achieves statistically significant improvement at the 5% level, MOE exhibits better performance in three of them, comparable performance in one of them, and slightly worse performance in the other one. We conclude that MOE is a promising machine learning model that can incorporate polar field data to improve flare classification accuracy.

We conclude this section by discussing some of the limitations of this study, which create opportunities for future research. Firstly, while we have established the value of polar field data when the train vs test data split is randomized over time, our conclusions may not

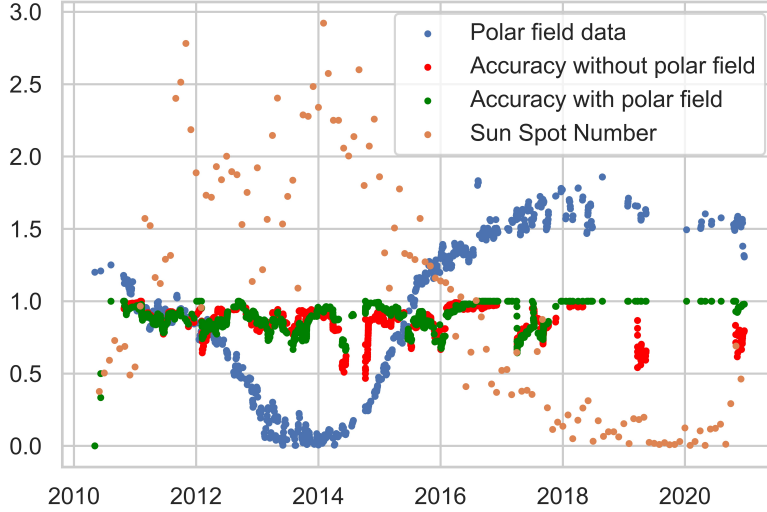


Figure 3.3: The classification accuracy curves of MOE with and without using polar fields data w.r.t. to the sunspot number when the splitting method is instance-based. Cycle approximation with filtered data is used as input. The curves are moving averaged over 90 days. The incorporation of the data substantially improves classification accuracy in some portions of the solar cycle.

translate to operational flare forecasting where train and test data are drawn from past and future observations, respectively. We could not implement such a temporally defined data splitting to train our algorithms due to the fact that HMI data is currently available only over a single solar cycle. Secondly, it is possible that we could strengthen our conclusions by accounting for the temporal and longitudinal correlations of the polar field and active region magnetograms. In particular, instead of summarizing the polar field as the difference between both north and south polar fields (CAPN2 and CAPS2) we might be able to improve our model by incorporating the spatial coordinates of the active regions and independently applying CAPN2 and CAPS2 to the northern and southern hemisphere active regions, respectively (see Fig. 3.1). Alternatively, it might be beneficial to incorporate a dynamic physics-based prediction model to account for dependencies between past active region activity and present polar field intensity. Furthermore, the added value of using polar field data may increase if we expanded the feature dimension to use the higher resolution HMI image data instead of just SHARP summarized data of active regions.

3.5 Conclusions

In this chapter, we investigated the utility of the sun’s north and south polar field strength for improving solar flare prediction from HMI data. We presented an extensive experimental study, that analyzes multiple ways of preprocessing schemes for polar field data, several machine learning models that can incorporate this additional data source, diverse experimental conditions, and multiple splitting methods. The mixture of experts model was introduced for this particular global data incorporation task. Our experimental results showed that polar field data is valuable for improving classification performance in multiple configurations. Additionally, we demonstrated that the proposed mixture of experts model can effectively supplement SHARP parameters with polar field data with for improved flare prediction.

The results suggest a causal connection between isolated ARs captured in the SHARP data and the global magnetic field captured in the polar field. The connection could come through the polar field itself due to the causal relationship between the event of decaying ARs and the subsequent drift of enhanced field towards poles. Hence, AR activity is anti-correlated to the polar field with a certain phase shift. In this study, we find that the probability of individual flare events is also correlated to the total sunspot number through the solar cycle, which closely follows the polar field strength. To put this work in context we point out that the polar field data has been used to inform solar cycle predictions with reasonable accuracy [227, 228]. We believe this work represents the first time the global field has been used to augment the prediction of solar flares.

In future work, there are several promising directions to enhance current results. The phase shift between the solar cycle/AR activity and polar field strengths can be estimated jointly by developing a more sophisticated model that can estimate and take into account phase shifts to improve prediction performance. The ARs in the northern and southern hemispheres can be modeled separately for more focused predictions. The correlation between the polar field strengths and other global indicators such as total active region area and flux-weighted latitude at a certain time can be investigated and possibly utilized for improved prediction performance. Finally, the data and the model may be further refined by using HMI high-dimensional image data instead of SHARP features.

Appendix A: Definitions of the SHARP parameters

Table 3.3 gives the SHARP parameters that are used in this work.

Parameters	Description
USFLUXL	Total unsigned flux
MEANGBL	Mean gradient of the line-of-sight field
RVALUE	Total line-of-sight unsigned flux
AREA	De-projected area of patch on sphere in microhemisphere
TOTUSJH	Total unsigned current helicity
TOTUSJZ	Total unsigned vertical current
SAVNCPP	Sum of the modulus of the net current per polarity
ABSNJZH	Absolute value of the net current helicity
TOTPOT	Proxy for total photospheric magnetic free energy density
SIZEACR	Deprojected area of active pixels (Bz magnitude larger than noise threshold) on image in microhemisphere (defined as one millionth of half the surface of the Sun)
NACR	The number of strong LOS magnetic field pixels in the patch
MEANPOT	Proxy for mean photospheric excess magnetic energy density
SIZE	Projected area of the image in microhemispheres
SHRGT45	Fraction of area with shear $> 45^\circ$
MEANSHR	Mean shear angle
MEANJZD	Vertical current density
MEANALP	Characteristic twist parameter, α
MEANGBT	Horizontal gradient of total field
MEANGAM	Mean angle of field from radial
MEANGBH	Horizontal gradient of horizontal field
NPIX	The number of pixels in a SHARP image

Table 3.3: List of SHARP Parameters and Brief Descriptions

CHAPTER 4

Multiway Covariance Models

4.1 Introduction

Probabilistic modeling of data is ubiquitous in statistics and machine learning, and arguably, estimating dependencies across covariates is one of the most fundamental modeling tasks. While first order “mean field” analysis can describe typical mean behavior of covariates, higher order analysis is required to capture interactions between covariates. Among the latter, covariance estimation is a second order analysis that has been extensively studied for empirically estimating the covariance or the inverse covariance (precision) matrix from multivariate data. Besides being of interest in their own right, covariance estimates are often a prerequisite in other applications, including: dimensionality reduction; multivariate clustering; supervised classification; linear prediction; and model selection [143, 202, 334, 20, 84, 279, 277, 280].

A striking feature of many modern datasets is their high-dimensionality and multiway nature, often involving a huge number of multi-indexed variables represented as tensor-valued data [153, 163]. Second order analysis of such tensor-valued data poses significant modeling and computational challenges due to the intrinsically high computational complexity of manipulating tensor representations [124]. To overcome these challenges, several methodological approaches have been developed over the past 50 years that reduce model complexity by imposing sparse or low-dimensional dependency structures in the covariance or inverse covariance. These include Kronecker generalizations of PCA (Kronecker PCA) [289, 103] and the matrix normal model [62], where the covariates are multivariate normal with multiway covariance structure modeled as a Kronecker product of matrices of much lower dimension. These Kronecker representations of covariance can have Kronecker factors that are either dense [76, 315] or sparse [10, 290].

The introduction of such parsimoniously structured tensor covariance models, and fast iterative computational algorithms for inferring these models, have made tensor covariance

models more practical for many applications. In particular, such models have been applied to image classification [86], spatio-temporal image processing [103, 101, 204, 66], bioinformatics [283, 100, 278, 74], atmospheric science [181, 102], neuroscience [306, 276, 26], medical imaging [190, 187], and digital advertising [117]. Several other applications of multiway covariance models are discussed in the unpublished survey [267].

In this article, we survey some of the recent advances in second order analysis of multiway data. The focus is on Kronecker structured representations and approximations to multiway covariance. After introducing some notation in Section 4.2, we describe how multiway covariance structure arises in different applications, including in linear prediction of dynamic processes whose evolution is governed by differential equations. We then overview recent approaches to Kronecker modeling and estimation of multiway covariance and inverse covariance in Section 4.3. In Section 4.4, we summarize and compare these Kronecker models and estimators based on their theoretical performance properties and a numerical comparison study. The numerical study focuses on model comparison for an important application in the domain of space weather: prediction of solar active region evolution over time. In Section 4.5, we conclude the chapter with a few open research directions in second order multiway data modeling. A software package [308] accompanies this survey with code implementing the methods described herein.

This survey is intended to complement previous surveys on aspects of high dimensional covariance modeling important to practitioners. These include surveys on robust structured covariance estimation [319, 149] and regularized covariance estimation [232, 165, 50]. Here, we focus on the class of Kronecker structured covariance models.

4.2 Background

4.2.1 Notation

Throughout the chapter, scalars are denoted by lowercase letters, vectors by boldface lowercase letters, and matrices by boldface capital letters. All matrix and vector quantities are assumed to be real valued, although all the multiway covariance models discussed in this chapter straightforwardly generalize to complex valued data. For a matrix $\mathbf{A} = (\mathbf{A}_{i,j}) \in \mathbb{R}^{d \times d}$, the spectral norm is denoted as $\|\mathbf{A}\|_2$, the Frobenius norm as $\|\mathbf{A}\|_F$, and the nuclear norm as $\|\mathbf{A}\|_*$. We define $\|\mathbf{A}\|_{1,\text{off}} := \sum_{i \neq j} |\mathbf{A}_{i,j}|$ as its off-diagonal ℓ_1 norm. For tensor algebra, we adopt the notations used by Kolda and Bader [153]. K -th order tensors are denoted by boldface Euler script letters, e.g., $\mathcal{X} \in \mathbb{R}^{d_1 \times \dots \times d_K}$, with the $(i_1, \dots, i_K)$ -th element denoted as $\mathcal{X}_{i_1, \dots, i_K}$. A fiber of a tensor is a higher order analogue of a row or a column of a ma-

trix. It is obtained by fixing all but one of the indices of the tensor. The vectorization of $\mathcal{X}$, denoted as $\text{vec}(\mathcal{X})$, stacks the entries of the tensor into a d -dimensional column vector $(\mathcal{X}_{1,1,\dots,1}, \mathcal{X}_{2,1,\dots,1}, \dots, \mathcal{X}_{d_1,d_2,\dots,d_K})^T$, where $d = \prod_{k=1}^K d_k$. Matricization, also known as unfolding, is the process of transforming a tensor into a matrix. The mode- k matricization of a tensor $\mathcal{X}$, denoted by $\mathcal{X}_{(k)}$, arranges the mode- k fibers to be the columns of the resulting matrix. The k -mode product of a tensor $\mathcal{X} \in \mathbb{R}^{d_1 \times \dots \times d_K}$ and a matrix $\mathbf{A} \in \mathbb{R}^{J \times d_k}$, denoted as $\mathcal{X} \times_k \mathbf{A}$, is a tensor of size $d_1 \times \dots \times d_{k-1} \times J \times d_{k+1} \times \dots \times d_K$, with entries defined as $(\mathcal{X} \times_k \mathbf{A})_{i_1, \dots, i_{k-1}, j, i_{k+1}, \dots, i_K} := \sum_{i_k=1}^{d_k} \mathcal{X}_{i_1, \dots, i_K} A_{j, i_k}$. We define the K -way Kronecker product (KP) as $\bigotimes_{k=1}^K \mathbf{A}_k = \mathbf{A}_1 \otimes \dots \otimes \mathbf{A}_K$, and the K -way Kronecker sum (KS) as $\bigoplus_{k=1}^K \mathbf{A}_k = \mathbf{A}_1 \oplus \dots \oplus \mathbf{A}_K = \sum_{k=1}^K \mathbf{I}_{[d_1:k-1]} \otimes \mathbf{A}_k \otimes \mathbf{I}_{[d_{k+1}:K]}$, where $\mathbf{I}_{[d_k:\ell]} = \mathbf{I}_{d_k} \otimes \dots \otimes \mathbf{I}_{d_\ell}$. For the case of $K = 2$, $\mathbf{A}_1 \oplus \mathbf{A}_2 = \mathbf{A}_1 \otimes \mathbf{I}_{d_2} + \mathbf{I}_{d_1} \otimes \mathbf{A}_2$.

Statistical convergence rate will be denoted by the $O_P(\cdot)$ notation, which is defined as follows. Consider a sequence of real random variables $\{X_n\}_{n \in \mathbb{N}}$ defined on a probability space $(\Omega, \mathcal{F}, P)$ and a deterministic (positive) sequence of reals $\{b_n\}_{n \in \mathbb{N}}$. By $X_n = O_P(1)$ is meant: $\sup_{n \in \mathbb{N}} P(|X_n| > K) \rightarrow 0$ as $K \rightarrow \infty$, where X_n is a sequence indexed by n . The notation $X_n = O_P(b_n)$ is equivalent to $\frac{X_n}{b_n} = O_P(1)$. By $X_n = o_p(1)$ is meant $P(|X_n| > \epsilon) \rightarrow 0$ as $n \rightarrow \infty$ for any $\epsilon > 0$. By $\lambda_n \asymp b_n$ is meant $c_1 \leq \frac{\lambda_n}{b_n} \leq c_2$ for all n , where $c_1, c_2 > 0$ are absolute constants. The asymptotic notation $a_n = O(b_n)$ means $\lim_{b \rightarrow \infty} |\frac{a_n}{b_n}| \leq C$ for some constant $C > 0$, while $c_n = \Omega(d_n)$ means $\liminf_{n \rightarrow \infty} |\frac{c_n}{d_n}| \geq C'$ for some constant $C' > 0$.

4.2.2 Multiway Data

Multiway data frame Multiway data are data that are indexed over multiple domains, called modes, and can be represented as a tensor-valued data frame. The number of modes is called the order of the tensor. Examples of multiway data include 3D images of the brain, where the modes are the 3 spatial dimensions, and spatio-temporal weather imaging data, a set of image sequences represented as 2 spatial modes and 1 temporal mode. One can also obtain a multiway data by stacking independent, identically distributed data samples.

Multiway patterned covariance matrix Given a multiway data tensor of order K , the covariance matrix of the vectorized data has a special patterned structure. For instance, suppose that $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$ is a multiway data frame of order 2, (i.e., a matrix). Then the covariance matrix of $\text{vec}(\mathbf{X})$ is a $d_1 d_2$ by $d_1 d_2$ matrix $\Sigma = (\Sigma_{ij})$ that corresponds to the covariance of the i -th and the j -th rows of $\mathbf{X}$. See Figure 4.1 for a visual illustration. As a special case, if the rows of $\mathbf{X}$ are independent and identically distributed with $d_2 \times d_2$ covariance Σ_{d_2} , then $\text{cov}(\mathbf{X}) = \mathbf{I}_{d_1} \otimes \Sigma_{d_2}$.

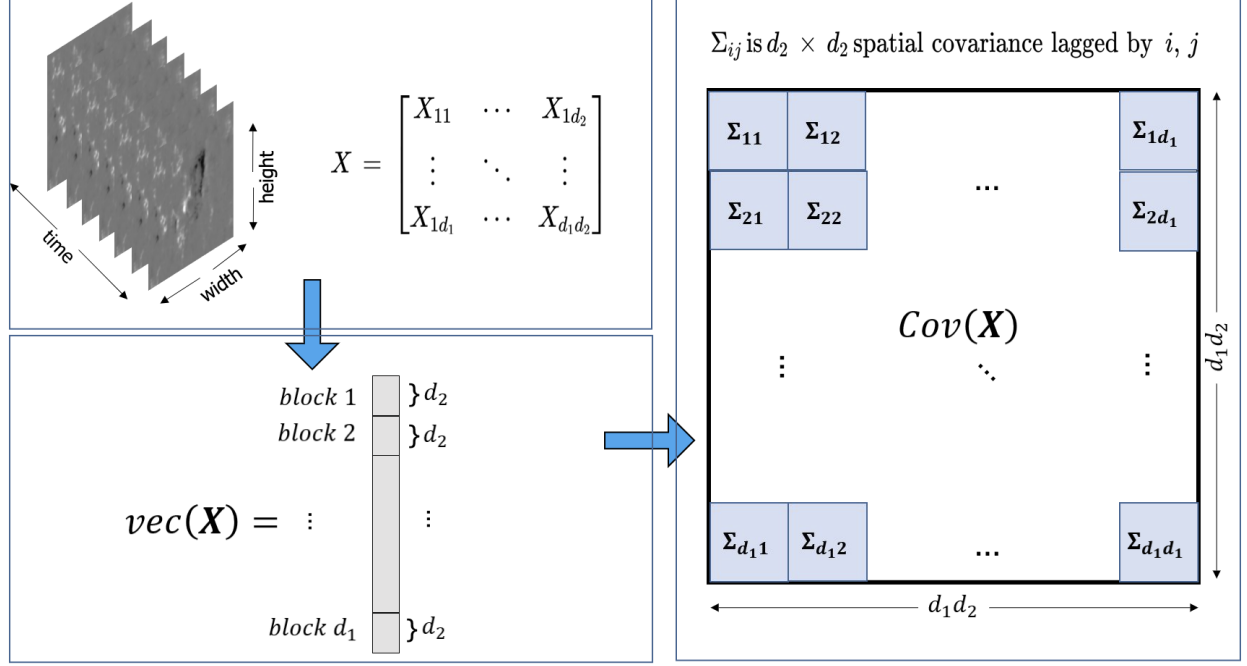


Figure 4.1: Multiway data and patterned covariance.

Utility of multiway data for prediction Arguably, regression analysis is one of the most straightforward approaches for prediction as it attempts to estimate the relation between the covariates and the response variable. Despite its long and fruitful history, traditional regression techniques often fall short of handling multiway data encountered in modern applications, due to the ultrahigh data dimensionality as well as complex data structure. Recently, several tensor regression models were proposed to exploit the special structure of tensor covariates [347, 117] or of tensor responses [234, 178, 348]. These models try to utilize the sparse and low-rank structures in the tensors – either in the regression coefficient tensor or the response tensor – to boost performance on the regression task by reducing the number of free parameters. Similarly, in Section 4.4.3, we demonstrate the utility of sparse precision/covariance matrices in linear prediction for tensor-valued data.

4.2.3 Multiway covariance representations for diffusion processes

While multiway representations have been successfully applied to many different types of data, they are particularly interpretable when modeling data generated from physical diffusion processes. In particular, consider a random process u whose sample paths obey a partial

differential equations of the form

$$\begin{aligned}\mathcal{D}u &= f \quad \text{in } \Omega, \\ u &= g \quad \text{on } \partial\Omega,\end{aligned}\tag{4.1}$$

where u is the unknown physical process, f is the driving process (e.g., white Gaussian noise), g is the function value of u on the boundary, $\mathcal{D}$ is some differential operator (e.g, a Laplacian or an Euler-Lagrange operator), and Ω is the domain. After finite difference discretization over the domain Ω , the model is equivalent to (ignoring discretization error) the matrix equation

$$\mathbf{D}\mathbf{u} = \mathbf{f}.$$

Here, $\mathbf{D}$ is a sparse matrix since $\mathcal{D}$ is a differential operator. Additionally, as shown below, $\mathbf{D}$ admits the Kronecker structure as a mixture of Kronecker sums and Kronecker products.

The matrix $\mathbf{D}$ reduces to a Kronecker sum when $\mathcal{D}$ involves no mixed derivatives. As an example, we consider the Poisson equation, an elliptical PDE that governs many physical processes including electromagnetic induction, heat transfer, and convection. On a rectangular region $\Omega = (0, d_1) \times (0, d_2)$ in the 2D Cartesian plane, the Poisson equation with homogeneous Dirichlet boundary condition is expressed as

$$\begin{aligned}\mathcal{D}u &= (\partial_x^2 + \partial_y^2)u = f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega\end{aligned}\tag{4.2}$$

where $f : \Omega \rightarrow \mathbb{R}$ is the given source function and $u : \Omega \rightarrow \mathbb{R}$ is the unknown process of interest. Using the finite difference method with a square mesh grid with unit spacing, the unknown and the source can be expressed as d_1 -by- d_2 matrices, $\mathbf{U}$ and $\mathbf{F}$, respectively, that are related to each other via

$$U_{i+1,j} + U_{i-1,j} + U_{i,j+1} + U_{i,j-1} - 4U_{i,j} = F_{i,j}\tag{4.3}$$

for any interior grid point (i, j) . Defining n -by- n square matrix

$$\mathbf{A}_n = \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix},$$

the above relation can be expressed as the (vectorized) Sylvester equation with $K = 2$:

$$(\mathbf{A}_{d_1} \oplus \mathbf{A}_{d_2})\mathbf{u} = \mathbf{f}, \quad (4.4)$$

where $\mathbf{u} = \text{vec}(\mathbf{U})$, $\mathbf{f} = \text{vec}(\mathbf{F})$. Note that $\mathbf{A}_n$ is tridiagonal. In the case where $\mathbf{f}$ is white noise with variance σ^2 , the inverse covariance matrix of $\mathbf{u}$ has the form $\text{cov}^{-1}(\mathbf{u}) = \sigma^{-2}(\mathbf{A}_{d_1} \oplus \mathbf{A}_{d_2})^T(\mathbf{A}_{d_1} \oplus \mathbf{A}_{d_2})$ and hence sparse.

More generally, any physical process generated from Equation (4.1) also has sparse inverse covariance matrices due to the sparsity of general discretized differential operators. Note that similar connections between continuous state physical processes and sparse “discretized” statistical models have been established by Lindgren et al. [183], who elucidated a link between Gaussian fields and Gauss Markov Random Fields via stochastic partial differential equations.

In some applications, e.g. weather forecasting and multivariate time series, the data frame $\mathbf{X}$ can be modeled as having been generated by an underlying dynamical process obeying the aforementioned physical diffusion model. This model can be used for various tasks such as forecasting, classification, and two-sample comparison. For example, when the dynamical model is of known linear form, the Kalman filter can be used for these tasks [34]. In the case of non-linear dynamics the ensemble Kalman filter (EnKF) [81] is applicable. Recently, a sparsity-penalized EnKF was introduced for modeling non-linear state dynamics [133]. This EnKF was implemented with a sparse inverse covariance estimator in order to stabilize the error covariance update of the Kalman Filter. When the state is high dimensional and tensor-valued, [304] used Kronecker-structured models for the (inverse) covariance and demonstrated improvements in tracking certain classes of dynamical models. A similar application of Kronecker covariance modeling will be further illustrated in Section 4.4.2.

4.3 Multiway Covariance Models

In this section, we survey several covariance models for multiway data and review associated estimation algorithms that have been introduced in the literature. The common assumption of these models is that the multimodal covariance (or inverse covariance) of multiway data has a factorized representation involving a Kronecker product or a Kronecker sum of smaller matrices, or as a linear combination thereof. Each of the smaller matrices corresponds to a single mode of data and these are often interpreted as mode-wise (inverse) covariances. The models and the algorithms reviewed in this section are summarized in Table 4.1.

Table 4.1: Overview of the models and the algorithms reviewed in Section 4.3. Here, Σ and Ω denote the covariance and the inverse covariance (precision), respectively. For the sake of brevity, the model in the second column is described for $K = 2$.

	Model	Algorithm	Note
KP-Covariance (Section 4.3.2)	$\Sigma = \sum_{l=1}^r \mathbf{A}_l \otimes \mathbf{B}_l$	KPCA [289]	Algorithm requires $K = 2$
		Robust KPCA [100]	Algorithm requires $K = 2$
		AdaKron [177]	Analysis restricted to $K = 2$
		TRCM [10]	Analysis restricted to $K = 2$
		Gemini [349]	Analysis restricted to $K = 2$
KP-Precision (Section 4.3.3)	$\Omega = \mathbf{A} \otimes \mathbf{B}$	KGlasso [10, 290]	Analysis restricted to $K = 2$
		Tlasso [192]	Applicable to $K \geq 2$
KS-Precision (Section 4.3.4)	$\Omega = \mathbf{A} \oplus \mathbf{B}$	Teralasso [102]	Applicable to $K \geq 2$
		BiGlasso [147]	Analysis restricted to $K = 2$
Sylvester GM (Section 4.3.5)	$\Omega = (\mathbf{A} \oplus \mathbf{B})^2$	SyGlasso [306]	Applicable to $K \geq 2$
		SG-PALM [305]	Applicable to $K \geq 2$

4.3.1 The Kronecker Product Covariance Model

For a K -mode tensor-valued data frame, the Kronecker product (KP) model of its covariance matrix Σ is

$$\Sigma = \Sigma_1 \otimes \cdots \otimes \Sigma_K \quad (4.5)$$

where $\Sigma_k \in \mathbb{R}^{d_k \times d_k}$, $k = 1, \dots, K$ are symmetric positive-semidefinite matrices, which are called the Kronecker factors of Σ .

The KP model with $K = 2$ imposes a simple two-way covariance representation recast as $\Sigma = \mathbf{A} \otimes \mathbf{B}$, arguably the simplest non-trivial one. Then, under this model, each $d_2 \times d_2$ block $[\Sigma]_{ij}$ of the $d_1 d_2 \times d_1 d_2$ patterned covariance matrix Σ shown in Figure 4.1 is the same matrix, namely $[\Sigma]_{ij} = \mathbf{B}$, up to a multiplicative factor. Despite its apparent simplicity, the KP model with $K = 2$ has been used effectively in many different applications, including spatio-temporal data frames arising in MIMO wireless communications [333, 315], geostatistics [58], genomics [330, 131], face recognition [344], and recommender systems [10].

The KP model was introduced by Dawid [62] in the context of the matrix normal distribution for data with $K = 2$ tensor modes. An alternating optimization algorithm for maximum likelihood estimation (MLE) of the $K = 2$ Kronecker factors was introduced by Mardia and Goodall [197] and analyzed by Dutilleul [76]. A different approach based on the covariance matching principle was taken by Strobach [265], who cast the problem of estimating the $K = 2$ Kronecker factors as a minimum Frobenius norm covariance matching

problem:

$$\min_{\mathbf{A}, \mathbf{B}} \|\mathbf{S} - \mathbf{A} \otimes \mathbf{B}\|_F^2. \quad (4.6)$$

Using the fact that the matrix Frobenius norm is invariant to permutation of the entries, the above optimization (4.6) boils down to finding the rank-one approximation to the matrix $\mathcal{R}(\mathbf{S})$, where $\mathcal{R}$ is the rearrangement-permutation operator that maps the square matrix $\mathbf{S} \in \mathbb{R}^{d_1 d_2 \times d_1 d_2}$ to a rectangular matrix $\mathcal{R}(\mathbf{S}) \in \mathbb{R}^{d_1^2 \times d_2^2}$ [295]. The minimizers of (4.6) can thus be expressed in closed form using the left and right principal components $\mathbf{u}$ and $\mathbf{v}$ of $\mathcal{R}(\mathbf{S})$ as $\hat{\mathbf{A}} = \mathcal{R}^{-1}(\mathbf{u})$ and $\hat{\mathbf{B}} = \mathcal{R}^{-1}(\mathbf{v})$, and the resultant estimator can be represented as $\hat{\mathbf{\Sigma}} = \hat{\mathbf{A}} \otimes \hat{\mathbf{B}}$. Compared to the MLE, this least square estimate does not require an iterative algorithm, but is generally not asymptotically efficient. A weighted Frobenius norm formulation based on (4.6) was considered by Werner et al. [315], leading to an asymptotically efficient estimator under the matrix normal assumption.

KP models with $K > 2$ also arise in applications and are explored in the literature. Mardia and Goodall [197] considered a spatio-temporal multivariate environmental monitoring data set and adopted a KP covariance with $K = 3$. They were the first to propose a maximum likelihood estimation via triple iteration. As there were insufficient replicates, they simplified the model by assuming the temporal components were statistically independent. Galecki [87] discussed five 3-way KP models with different combinations of covariance factor structures of identity, compound symmetry, AR(1), unstructured. Akdemir and Gupta [6] provided a heuristic method for $K \geq 2$, generalizing the flip-flop method originally proposed for matrices. The convergence of the algorithm was empirically verified using simulations with K up to 4, and using real world data with K up to 3. Hoff [129] discussed maximum likelihood estimation and Bayesian estimation for tensor normal models for general K , the latter approach verified on a 4-way international trade dataset. We note here that, in both Hoff [129] and Ohlson et al. [221], the KP model was referred to as the “separable covariance” model. An empirical study of the three-stage iterative algorithm for MLE of KP structured covariance with $K = 3$ was given by Manceur and Dutilleul [195], who provided numerical and simulation results showing convergence of the algorithm. Pouryazdian et al. [233] proposed to select the order of Canonical Decomposition (CANDECOMP) in modeling 3-way data under the assumption that the noise covariance is a KP model with $K = 3$ modes, justified by the structure of electroencephalogram (EEG) data spanning the time, frequency, and space dimensions.

4.3.2 Kronecker Covariance Decomposition and Kronecker PCA

There is a less restrictive model than (4.5) that approximates the covariance as a sum of several Kronecker products. For example, when $K = 2$, the covariance under this model is expressed as

$$\mathbf{\Sigma} = \sum_{l=1}^r \mathbf{A}_l \otimes \mathbf{B}_l, \quad (4.7)$$

where $r \geq 1$ is a positive integer, $\mathbf{A}_l \in \mathbb{R}^{d_1 \times d_1}$ and $\mathbf{B}_l \in \mathbb{R}^{d_2 \times d_2}$ for all $1 \leq l \leq r$. The sum of Kronecker product representation (4.7) is universal in the sense that any matrix $\mathbf{\Sigma} \in \mathbb{R}^{d_1 d_2 \times d_1 d_2}$ can be expressed as a sum of r Kronecker products with $r \leq \min\{d_1^2, d_2^2\}$ in view of SVD, cf. Van Loan and Pitzianis [295, Section 2]. Indeed, even when $K \geq 3$, every tensor $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$ can be represented by a sum of Kronecker products $\mathbf{\Sigma} = \sum_{l=1}^r \mathbf{A}_l^{(1)} \otimes \cdots \otimes \mathbf{A}_l^{(K)}$ where $d = \prod_{k=1}^K d_k$ and $\mathbf{A}_l^{(k)} \in \mathbb{R}^{d_k \times d_k}$ for all $k \in [K]$ and $l \in [r]$. The minimum positive integer r such that $\mathbf{\Sigma}$ admits a representation of the form (4.7) is called the (tensor) rank of $\mathbf{\Sigma}$ [153]; note that the KP model from Section 4.3.1 corresponds to the special case with tensor rank 1 with additional constraints enforcing the factor matrices to be symmetric positive semidefinite.

A model of the form (4.7) (with small r) allows for a compact, yet quite flexible, representation of the matrix $\mathbf{\Sigma}$. Recently, Dantas et al. [61] investigated the utility of this model for hyperspectral image denoising through the lens of dictionary learning. We remark that they used the model (4.7) to approximate the first-order “mean-field” behavior of data rather than its covariance (cf. Section 4.1), and can be considered a higher-order analogue of low-rank matrix models.

In the context of second-order (covariance) analysis, Tsiligkaridis and Hero [289] proposed a penalized optimization approach for estimating a (tensor) rank r Kronecker product decomposition (4.7) of covariance, proposing an estimate that solves

$$\min_{\{\mathbf{A}_l, \mathbf{B}_l\}} \left\| \mathbf{S} - \sum_{l=1}^r \mathbf{A}_l \otimes \mathbf{B}_l \right\|_F^2 + \lambda \left\| \sum_{l=1}^r \mathbf{A}_l \otimes \mathbf{B}_l \right\|_*$$

for a user-supplied regularization parameter $\lambda > 0$. The solution to this penalized optimization is specified by the first r principal components of the singular value decomposition (SVD) of $\mathcal{R}(\mathbf{S})$ where r is determined by λ through a soft-thresholding of the SVD spectrum. In analogy to the ordinary PCA algorithm, the soft-thresholding SVD solution to this optimization problem was called Kronecker PCA (KPCA) in [103]. In [289], it was shown that, when the sample size $n > d_1 d_2$, KPCA provides a consistent estimator of $\mathbf{\Sigma}$. Bounds

on statistical convergence rates were also established.

Subsequently, regularized variants of KPCA were proposed. Greenewald and Hero [103] proposed to add one pair of diagonal matrices in the representation (4.7), namely, $(\mathbf{A}_0, \mathbf{B}_0)$, to provide a better fit to homogeneous noise in spatiotemporal modeling. This was further extended to a robust Kronecker PCA [100], which is equipped with additional capability to handle sparse unstructured “outlier correlations” that do not fit the vanilla KPCA model à la robust PCA [42, 45].

While the KPCA model is flexible in the sense that every covariance tensor admits a representation (4.7) for some r , there is no a priori guarantee that the factors $\mathbf{A}_l, \mathbf{B}_l$ are symmetric, positive semidefinite (PSD) matrices. Indeed, one can see from a dimension counting argument¹ that if $d_1, d_2 \geq 2$, then there exists a PSD matrix $\Sigma \in \mathbb{R}^{d_1 d_2 \times d_1 d_2}$ that cannot be expressed as a sum of Kronecker products of symmetric matrices $\mathbf{A}_l \in \mathbb{R}^{d_1 \times d_1}$ and $\mathbf{B}_l \in \mathbb{R}^{d_2 \times d_2}$. As a PSD matrix is symmetric, the aforementioned matrix Σ cannot be expressed as a sum of Kronecker products of PSD matrices. Unless the factors $\mathbf{A}_l, \mathbf{B}_l$ are PSD, we may not be able to interpret them as mode-wise ‘principal covariance components.’

It remains an open question how closely an arbitrary covariance Σ can be approximated by a sum of Kronecker product of PSD matrices, or whether a typical covariance Σ can be represented as a sum of Kronecker product of PSD matrices under reasonable data-generative models.

4.3.3 Kronecker Graphical Lasso and Tensor Lasso

As discussed in Section 4.2.3, physical processes often give rise to sparse covariance or sparse inverse covariance. Furthermore, when the number of independent samples, n , is smaller than the data dimension, $d = \prod_{i=1}^K d_i$, imposition of a sparse structure can stabilize the multi-way covariance estimates. Leng and Pan [177] considered the problem of estimating a sparse covariance in the KP model (4.5) using a sparsity-penalized moment matching approach. Under the matrix normal model, this corresponds to a sparse model on the pairwise marginal dependencies in the covariates. It is sometimes more natural to impose sparsity on the conditional dependencies, equivalent to imposing sparsity on the inverse covariance instead of on the covariance itself, as demonstrated by Yuan and Lin [334] and Banerjee *et al.* [20]. A fast algorithm for sparse estimation of the inverse covariance in the matrix normal framework was proposed by Friedman *et al.* [84], who called the algorithm

¹The set of $d_1 d_2 \times d_1 d_2$ PSD matrices has nonempty interior in the vector space of $d_1 d_2 \times d_1 d_2$ symmetric matrices that has dimension $\binom{d_1 d_2 + 1}{2}$. Note that the tensor product of the vector space of $d_1 \times d_1$ symmetric matrices and that of $d_2 \times d_2$ symmetric matrices is the span of Kronecker products and has dimension $\binom{d_1 + 1}{2} \binom{d_2 + 1}{2}$, which is strictly smaller than $\binom{d_1 d_2 + 1}{2}$ if $d_1, d_2 \geq 2$.

the graphical lasso (Glasso).

Allen and Tibshirani [11] extended the Glasso to the matrix normal model ($K = 2$ modes) and this algorithm, which we call the Kronecker graphical lasso (KGlasso), was shown to be consistent with statistical convergence bounds in [290]. The KGlasso uses an alternating flip-flop optimization on the sparsity penalized KP log-likelihood function, similarly to the algorithm proposed by Mardia and Goodall [197] for the original unpenalized KP maximum likelihood problem. Besides the diffusion processes mentioned in Section 4.2.3, the sparse KP inverse covariance structures also arise in classification problems where the predictors are matrix-valued [205].

The KGlasso solves the following non-convex optimization, equivalent to maximizing the penalized log likelihood,

$$\min_{\mathbf{A}, \mathbf{B}} \text{tr}\{\mathbf{S}(\mathbf{A} \otimes \mathbf{B})\} - \log \det(\mathbf{A} \otimes \mathbf{B}) + \lambda_1 \|\mathbf{A}\|_1 + \lambda_2 \|\mathbf{B}\|_1 \quad (4.8)$$

where λ_1, λ_2 are sparsity regularization parameters applied to the ℓ_1 norms of the Kronecker product factors $\mathbf{A}$ and $\mathbf{B}$. Here, the ℓ_1 norms are often restricted to range only over the off-diagonals of $\mathbf{A}$ and $\mathbf{B}$. Like the Glasso, iterative optimization methods are required to solve the optimization problem (4.8), however, algorithmic implementation of such methods can be significantly more difficult because it is a non-convex problem.

The tensor lasso (Tlasso) introduced by Lyu *et al.* [192] is a generalization of the KGlasso to $K \geq 2$ modes. As in the KGlasso, an alternating flip-flop maximization method is used to estimate the precision matrix factors. Consistency and a minimax property are established for the Tlasso. As pointed out in [192], normalization of the K inverse covariance factors is necessary to ensure model identifiability as the tensor normal model is invariant to reciprocal scaling and other transformations on the Kronecker factors.

Following KGlasso and Tlasso, recent works have focused on more efficient estimation algorithms for solving the non-convex problem posed by (4.8) with provable guarantees. For example, instead of alternatingly applying Glasso, Xu et al. [325] proposed an alternating block gradient descent algorithm that has been shown to converge linearly to the true unknown precision matrices with optimal statistical error rates. More recently, Min et al. [203] proposed a parallel estimation scheme as an alternative to previous alternating/cyclic approaches for estimating the K inverse covariance factors. They demonstrated that the parallelized version achieves similar convergence rates as alternating block descent.

4.3.4 Tensor Graphical Lasso

The tensor graphical lasso (TeraLasso) of Greenewald *et al.* [102] is a $K \geq 2$ generalization of the bigraphical lasso (BiGlasso) of Kalaitzis *et al.* [147]. For simplicity, we confine the discussion to the simpler $K = 2$ mode case. In contrast to the KGlasso/Tlasso, the TeraLasso replaces the Kronecker product representation of the inverse covariance, $\mathbf{\Omega} = \mathbf{A} \otimes \mathbf{B}$, with a Kronecker sum representation: $\mathbf{\Omega} = \mathbf{A} \oplus \mathbf{B} = \mathbf{A} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{B}$. The rationale of [147, 102] that motivates the Kronecker sum model is that, for equal number of non-zero entries in factors $\mathbf{A}$ and $\mathbf{B}$, it gives a sparser inverse covariance than the Kronecker product. This can be appreciated from the fact that the Kronecker sum representation of the inverse covariance corresponds to a Cartesian product graphical model instead of a direct product graphical model given by the Kronecker product representation. This KS structure has recently been studied in the context of matrix-valued time-series analysis – Jiang *et al.* [140] introduced a novel multivariate autoregressive model to infer the graph topology encoded in the coefficient matrix, which captures the sparse Granger causality dependency structures describing both the sparse spatial relationship between sensors and the multiple measurement relationship. They decompose the graph by imposing a KS structure on the coefficient matrix.

Furthermore, the Kronecker sum representation converts non-convex optimization problem in the KGlasso (4.8) into the convex optimization problem:

$$\min_{\mathbf{A}, \mathbf{B}} \text{tr}\{\mathbf{S}(\mathbf{A} \oplus \mathbf{B})\} - \log \det(\mathbf{A} \oplus \mathbf{B}) + \lambda_1 \|\mathbf{A}\|_1 + \lambda_2 \|\mathbf{B}\|_1 \quad (4.9)$$

where λ_1, λ_2 are sparsity regularization parameters applied to the ℓ_1 norms of the Kronecker sum factors $\mathbf{A}$ and $\mathbf{B}$. As in the KGlasso, these ℓ_1 norms range over the off-diagonals of $\mathbf{A}$ and $\mathbf{B}$. This objective can be solved via either an iterative application of Glasso [147] or by a projected proximal gradient descent algorithm [102]. More recent work has also extended these estimation schemes to be more scalable and to handle non-identifiability [331] and count data [179].

4.3.5 Sylvester Graphical Lasso and SG-PALM

The Sylvester graphical lasso (SyGlasso) [306] models the precision matrix as the square of a sparse matrix that has Kronecker sum structure. Different from KGlasso, Tlasso and TeraLasso, the Sylvester graphical model uses the Sylvester tensor equation to define a generative process for the underlying multivariate tensor data. The Sylvester tensor equation has been studied in the context of finite-difference discretization of high-dimensional elliptical partial differential equations [99, 156]. Consider the simplest case of $K = 2$, any solution $\mathbf{X}$

to such a PDE must have the (discretized) form:

$$\mathbf{X}\mathbf{A} + \mathbf{B}\mathbf{X} = \mathbf{Z} \iff (\mathbf{A} \oplus \mathbf{B}) \text{vec}(\mathbf{X}) = \text{vec}(\mathbf{Z}), \quad (4.10)$$

where $\mathbf{Z}$ is the driving source on the domain, and $\mathbf{A} \oplus \mathbf{B}$ represents the discretized differential operators for the PDE, e.g., Laplacian, Euler-Lagrange operators, and associated coefficients. These operators are often sparse and structured.

The Sylvester generative (SG) model (4.10) leads to a matrix-valued random variable $\mathbf{X}$ with a precision matrix $\mathbf{\Omega} = (\mathbf{A} \oplus \mathbf{B})^2$, given that $\mathbf{Z}$ is isotropic (white) Gaussian distributed. The Sylvester generating factors $\mathbf{A}$ and $\mathbf{B}$ can be obtained via minimization of the penalized negative log-pseudolikelihood

$$\min_{\mathbf{A}, \mathbf{B}} \text{tr}\{\mathbf{S}(\mathbf{A} \oplus \mathbf{B})^2\} - \log \det \left(\text{diag}(\mathbf{A}) \oplus \text{diag}(\mathbf{B}) \right) + \lambda_1 \|\mathbf{A}\|_1 + \lambda_2 \|\mathbf{B}\|_1. \quad (4.11)$$

This differs from the true penalized Gaussian negative log-likelihood in the exclusion of off-diagonals of $\mathbf{A}$ and $\mathbf{B}$ in the log-determinant term. (4.11) is motivated and derived directly using the Sylvester equation defined in (4.10), from the perspective of solving a sparse linear system. This is a generalization and extension of the convex pseudolikelihood framework for high dimensional partial correlation estimation (CONCORD) [150] to tensor-variate variables. The pseudolikelihood function is convex and is maximized using a entrywise regression approach similar to CONCORD. It is known that inference using pseudo-likelihood is consistent and enjoys the same $\sqrt{N}$ convergence rate as the MLE in general [22, 296].

Estimation of the generating parameters of the SG model is challenging since the sparsity penalty applies to the square root factors of the precision matrix, which leads to a highly coupled likelihood function. The nodewise regression approach in SyGlasso recovers only the off-diagonal elements of each Sylvester factor. This is a deficiency in many applications where the factor-wise variances are desired. Moreover, the convergence rate of the cyclic coordinate-wise algorithm used in SyGlasso is unknown and the computational complexity of the algorithm is higher than other sparse Glasso-type procedures. To overcome these deficiencies, Wang *et al.* [305] proposed a proximal alternating linearized minimization method that is more flexible and versatile, called SG-PALM, for finding the minimizer of (4.11). SG-PALM is designed to exploit structures of the coupled objective function and yields simultaneous estimates for both off-diagonal and diagonal entries. It also achieves state-of-the-art iterative convergence result (i.e., linear convergence of the optimization error) and an order of magnitude improvement in practical runtime per iteration. Additionally, SG-PALM easily adapts to non-convex regularization functions (e.g., SCAD, MCP) and enjoys similar rate of iterative convergence.

4.4 Illustrations and Performance Comparisons

In this section, we discuss theoretical guarantees on the multiway covariance estimation procedures described in Section 4.3 and provide numerical illustrations of them. In Section 4.4.1, we collect and compare theoretical guarantees on the statistical consistency and computational complexity for these covariance estimation algorithms. Then, in Section 4.4.2, we provide empirical performance comparisons when the covariance / precision matrix estimation methods are used for characterizing multiway physical processes. Finally, in Section 4.4.3, we demonstrate the use of tensor graphical models on the real-world application of solar active region evolution prediction. The numerical studies in Section 4.4.2 and 4.4.3 were performed using the publicly available software package [308] developed by the authors, and the readers can explore other properties beyond those presented here.

4.4.1 Theoretical Guarantees

All of the methods described in Section 4.3 have non-asymptotic theoretical guarantees on the statistical consistency of the estimators and the rate of computational convergence, which can be converted to upper bounds on the sample- and computational complexity of the methods, respectively. The best known upper bounds on the estimation error (in Frobenius norm) and those on computational complexity (in number of flops per iteration) are summarized in Table 4.2.

Statistical Error For each row of Table 4.2, the expression for statistical error applies only when the associated model is correct, i.e., there is no model mismatch (bias) between the model generating the observations and the model assumed to estimate covariance/precision. For structured covariance estimation, the more recent AdaKron algorithm achieves better statistical error (given that the true covariance matrix is sparse) than KPCA/Robust KPCA as the dependence on d_i 's is linear. For structured precision matrix estimation, all the structured methods enjoy significantly reduced statistical error compared to the traditional Glasso. TeraLasso outperforms Tlasso and SyGlasso/SG-PALM approximately by a factor of d_k and s_k , respectively. Moreover, it is noteworthy that Tlasso and TeraLasso achieve single-sample convergence of the ℓ_2 norm (operator norm) errors that are not included in the table, meaning that the error vanishes with a single sample ($N = 1$) as the dimension grows to infinity [102, 192].

Computational Complexity KPCA-based methods require expensive full SVD/eigenvalue factorization in general, which dominates their runtime complexity. Although speedup

can be achieved via randomized algorithms or iterative algorithms for sparse matrices along with small r , the computational complexities of these improvements remain unknown for general problem instances.

All the structured precision estimation algorithms are variants of Glasso, implemented with techniques tailored to the model assumptions for speedup. Generally speaking, the resulting complexity consists of the mode-wise complexity (d_k^3) and the cost of updating the objective: dK for TeraLasso, $N \sum_k d_k m_k^2$ for Tlasso, and $N \sum_k \sum_{j \neq k} d_j m_j^2$ for SG-PALM. Note that AdaKron for structured covariance models uses a pre-processing step that avoids this operation involving d and/or N flops per-iteration, but this works only for $K = 2$. The mode-wise complexity of TeraLasso is dominated by matrix inversion, which is hard to scale for general problem instances. For Tlasso/KGlasso, the mode-wise complexity is the same as that of running a Glasso-type algorithm for each mode, which could be improved by applying state-of-the-art optimization techniques developed for vector-variate Gaussian graphical models. For SG-PALM, the mode-wise operations involve only sparse-dense matrix multiplications, which could be improved to $O(d_k \cdot \text{nnz})$, where nnz counts the number of non-zero elements of the sparse matrix (i.e., the estimated $\mathbf{\Omega}_k$ at each iteration). This could greatly reduce the computational cost for extremely sparse $\mathbf{\Omega}_k$, e.g., with only $O(d_k)$ non-zero elements. Further, Tlasso and SG-PALM both incur a cost of $O(N d_k m_k^2)$ for each mode-wise update. This can also be reduced to be $\approx d$ for sparse estimated $\mathbf{\Omega}_k$'s at each iteration. Overall, for sample-starved setting where we only have access to a handful of data samples, structured KP and KS models run similarly fast, while the Sylvester GM runs slower theoretically due to the extra and richer structures that it takes into account.

4.4.2 Numerical Experiments with Synthetic Data

Here we compare the performance of the Kronecker covariance / precision estimation methods when used to estimate multiway physical processes, based on numerical experiments using synthetic data. To this end, we first describe two generative models that extend the spatial Poisson equation (cf. Section 4.2.3) to incorporate temporal dynamics, and the resulting multiway (inverse) covariance structure in Section 4.4.2.1. Thereafter, in Section 4.4.2.2, we illustrate the performance of the various (inverse) covariance estimation methods using synthetic data generated from the two models of spatio-temporal dynamics.

4.4.2.1 Two Generative Models for Spatio-temporal Dynamics

We describe two dynamic models that extend the spatial Poisson equation (cf. Section 4.2.3) to incorporate temporal dynamics, and the resulting multiway (inverse) covariance structure.

Table 4.2: Comparison of known theoretical guarantees on sample complexity (statistical error) and computational complexity. Here, $M = \max\{d_1, d_2, N\}$, $m_k = \prod_{i \neq k} d_i$ is the co-dimension of the k -th mode, $d = \prod_{k=1}^K d_i$, and s_k characterizes the sparsity of each of the inverse covariance Kronecker factors $s_k = |\{(i, j) : i \neq j, [\mathbf{\Omega}_k]_{i,j} \neq 0\}|$, s is the sparsity of the full inverse covariance $s = |\{(i, j) : i \neq j, \mathbf{\Omega}_{i,j} \neq 0\}|$ and $s = \sum_{k=1}^K m_k s_k$ if $\mathbf{\Omega}$ satisfies the Kronecker sum model.

Model	Algorithm	Statistical Error	Computational Complexity
KP-Covariance (Section 4.3.2)	KPCA [289]	$O_P\left(\sqrt{\frac{r(d_1^2 + d_2^2 + \log M)}{N}}\right)$	$O(d_1^3 d_2^3)$
	Robust KPCA [100]	$O_P\left(\sqrt{\frac{r(d_1^2 + d_2^2 + \log M)}{N}}\right)$	$O(d_1^3 d_2^3)$
	AdaKron [177]	$O_P\left(\sum_{i=1}^2 \sqrt{\frac{(s_i + d_i) \log d_i}{N}}\right)$	$O(d_1^3 + d_2^3)$
KP-Precision (Section 4.3.3)	KGlasso [10, 290]	$O_P\left(\sqrt{\frac{(d_1 + d_2) \log M}{N}}\right)$	$O(d_1^3 + d_2^3 + N d_1 d_2)$
	Tlasso [192]	$O_P\left(\sqrt{\frac{d_k(d_k + s_k) \log d_k}{N d}}\right)$ mode- k	$O\left(\sum_{k=1}^K (d_k^3 + N d_k m_k^2)\right)$
KS-Precision (Section 4.3.4)	TeraLasso [102]	$O_P\left(\sqrt{K+1} \cdot \sqrt{\frac{(d+s) \log d}{N \min_k m_k}}\right)$	$O(dK + \sum_{k=1}^K d_k^3)$
Sylvester GM (Section 4.3.5)	SyGlasso [306]	$O_P\left(\sqrt{K} \cdot \max_k \sqrt{\frac{s_k d_k \log d}{N}}\right)$	$O\left(d \sum_{k=1}^K (N d_k + \sum_{j \neq k} (d_k + d_j))\right)$
	SG-PALM [305]	$O_P\left(\sqrt{K} \cdot \max_k \sqrt{\frac{s_k d_k \log d}{N}}\right)$	$O\left(\sum_{k=1}^K (d_k^3 + N \sum_{j \neq k} d_j m_j^2)\right)$

These models will be used in our numerical experiments to generate data to compare the performance of estimation algorithms in Section 4.4.2.2.

Poisson-AR(1) Process. The first extension, which we call the Poisson-AR(1) process, imposes an autoregressive temporal model of order 1 on the source function f in the Poisson equation (4.2). Specifically, we say a sequence of discretized spatial observations $\{\mathbf{U}^k \in \mathbb{R}^{d_1 \times d_2}\}_k$ indexed by time step $k = 1, \dots, T$ is from a Poisson-AR(1) process if

$$(\mathbf{A}_{d_1} \oplus \mathbf{A}_{d_2}) \text{vec}(\mathbf{U}^k) = \text{vec}(\mathbf{Z}^k), \quad (4.12)$$

$$\text{vec}(\mathbf{Z}^k) = a \text{vec}(\mathbf{Z}^{k-1}) + \text{vec}(\mathbf{W}^k), \quad |a| < 1, \quad (4.13)$$

where $\{\mathbf{W}^k \in \mathbb{R}^{d_1 \times d_2}\}_k$ is spatiotemporal white noise, i.e., $W_{i,j}^k \sim \mathcal{N}(0, \sigma_w^2)$, i.i.d. Assuming $\mathbf{Z}^0 = \mathbf{0}$ and defining the T -by- T matrix

$$\mathbf{B} = \begin{bmatrix} 1 & -a & & \\ & 1 & \ddots & \\ & & \ddots & -a \\ & & & 1 \end{bmatrix},$$

the above linear system of equations can be written as $(\mathbf{A}_{d_1} \oplus \mathbf{A}_{d_2})\mathbf{U}\mathbf{B} = \mathbf{W}$, or equivalently,

$$(\mathbf{B}^T \otimes (\mathbf{A}_{d_1} \oplus \mathbf{A}_{d_2})) \text{vec}(\mathbf{U}) = \text{vec}(\mathbf{W}), \quad (4.14)$$

where $\mathbf{U} = [\text{vec}(\mathbf{U}^1) \text{vec}(\mathbf{U}^2) \dots \text{vec}(\mathbf{U}^T)] \in \mathbb{R}^{d_1 d_2 \times T}$ and $\mathbf{W}$ is defined likewise. The inverse covariance of $\mathbf{U}$, despite having a large size of $d_1 d_2 T \times d_1 d_2 T$, is sparse and has a mixed Kronecker sum and product structure.

Convection-diffusion Process. The second time-varying extension of the Poisson PDE model (4.2) is based on the convection-diffusion process [46]

$$\frac{\partial u}{\partial t} = \theta \sum_{i=1}^2 \frac{\partial^2 u}{\partial x_i^2} - \epsilon \sum_{i=1}^2 \frac{\partial u}{\partial x_i}. \quad (4.15)$$

Here, $\theta > 0$ is the diffusivity; and $\epsilon \in \mathbb{R}$ is the convection velocity of the quantity along each coordinate. Note that for simplicity of discussion here, we assume these coefficients do not change with space and time (see, Stocker [263], for example, for a detailed discussion). These equations are closely related to the Navier-Stokes equation commonly used in stochastic modeling for weather and climate prediction [46, 263]. Coupled with Maxwell's equations, these equations can be used to model magneto-hydrodynamics [243], which characterize solar activities including flares.

A solution of Equation (4.15) can be approximated similarly as in the Poisson equation case, through a finite difference approach. Denote the discrete spatial samples of $u(\mathbf{x}, t)$ at time t_k as a matrix $\mathbf{U}^k \in \mathbb{R}^{d_1 \times d_2}$. We obtain a discretized update propagating $u(\mathbf{x}, t)$ in space and time, which locally satisfies

$$\begin{aligned} \frac{U_{i,j}^k - U_{i,j}^{k-1}}{\Delta t} = & \theta \left(\frac{U_{i+1,j}^k + U_{i-1,j}^k + U_{i,j+1}^k + U_{i,j-1}^k - 4U_{i,j}^k}{h^2} \right) \\ & - \epsilon \left(\frac{U_{i+1,j}^k - U_{i-1,j}^k + U_{i,j+1}^k - U_{i,j-1}^k}{2h} \right), \end{aligned} \quad (4.16)$$

where $\Delta t = t_{k+1} - t_k$ is the time step and h is the mesh step (spatial grid spacing). Similarly to the Poisson-AR(1) process, in the following, we consider a “blocked” version of the convection-diffusion process.

We define the first-order and second-order discretized differential operators, denote by $\mathbf{D}$

and $\mathbf{A}$, respectively:

$$\mathbf{D} = \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix}.$$

Then, Equation (4.16) can be written as

$$\begin{aligned} \frac{1}{\Delta t}(\mathbf{D} \otimes \mathbf{I} \otimes \mathbf{I}) \text{vec } \mathbf{U} &= \frac{\theta}{h^2}(\mathbf{I} \otimes \mathbf{A} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{I} \otimes \mathbf{A}) \text{vec } \mathbf{U} \\ &\quad - \frac{\epsilon}{2h}(\mathbf{I} \otimes \mathbf{D} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{I} \otimes \mathbf{D}) \text{vec } \mathbf{U}, \end{aligned} \quad (4.17)$$

where $\mathbf{U} = [\text{vec}(\mathbf{U}^1) \text{vec}(\mathbf{U}^2) \dots \text{vec}(\mathbf{U}^T)] \in \mathbb{R}^{d_1 d_2 \times T}$. Assuming the process is driven by some white noise $\mathbf{W}$, similarly defined as in the Poisson-AR equation, the inverse covariance of $\mathbf{U}$ is again sparse and has a mixed Kronecker sum and product structure.

4.4.2.2 Experiments with Synthetic Data: Performance Comparisons

We consider spatio-temporal processes (2D space + time) that evolve $T = 50$ time steps on a 8×8 spatial grid according to the Poisson-AR(1) and the Convection-Diffusion dynamics (cf. Section 4.4.2.1). We generated $N = 50$ independent realizations of each type of process. Figure 4.2 shows one realization for each type of process, with spatial resolution increased for better illustration. Assuming $K = 2$ with a spatial dimension 64 and a temporal dimension 50, we estimated the covariance / inverse covariance using several sparse multiway inverse covariance estimation methods including Glasso [84], KPCA (Section 4.3.2), Tlasso (Section 4.3.3), TeraLasso (Section 4.3.4), SG-PALM (Section 4.3.5). Note that none of the above-mentioned models operate under the true generative processes (i.e., there is model mismatch with the data). Here, the sparsity-regularized methods are all implemented with an ℓ_1 penalty function, and the penalty parameters were selected similarly and guided by the theoretical results in Table 4.2. For example, for SG-PALM, we use a penalty parameter of $\lambda_k = C \sqrt{\frac{d_k \log d}{N}}$ where C is chosen by optimizing a normalized Frobenius norm error between the estimate and the truth, over a range of λ values parameterized by C . For the KPCA algorithm, both the nuclear norm penalty parameter and the separation rank are selected by optimizing a normalized Frobenius norm error via grid search.

A summary of the accuracy of estimation (normalized Frobenius norm error), as well as support recovery (Mathews Correlation Coefficient [200]), are reported in Table 4.3. In Figure 4.3 and Figure 4.4 we show the true and the estimated inverse covariance matrices

obtained for all the methods except KPCA, under both the Poisson-AR (Figure 4.3) and the Convection-Diffusion processes (Figure 4.4). The inverse covariances under both generating processes admit structures with a mix of Kronecker sums and Kronecker products of sparse matrices. In both of the cases, the SG-PALM produces estimates having the lowest error. We believe this is due to the goodness of fit because the Sylvester graphical model imposes a squared KS structure on the precision matrix. Tlasso has comparable performances and achieves the best matrix approximation error under the convection-diffusion generating process. This might be due to the fact that the KP model corresponds to an underlying spatio-temporal autoregressive process. TeraLasso seems to produce the biggest model mismatch as indicated by the MCC scores. In Figure 4.5, we compare the true covariance matrices and the estimates obtained by KPCA. Here, although the KPCA model does not match the underlying generating processes, the estimates were able to capture certain blocking patterns that similarly exist in the true covariance.

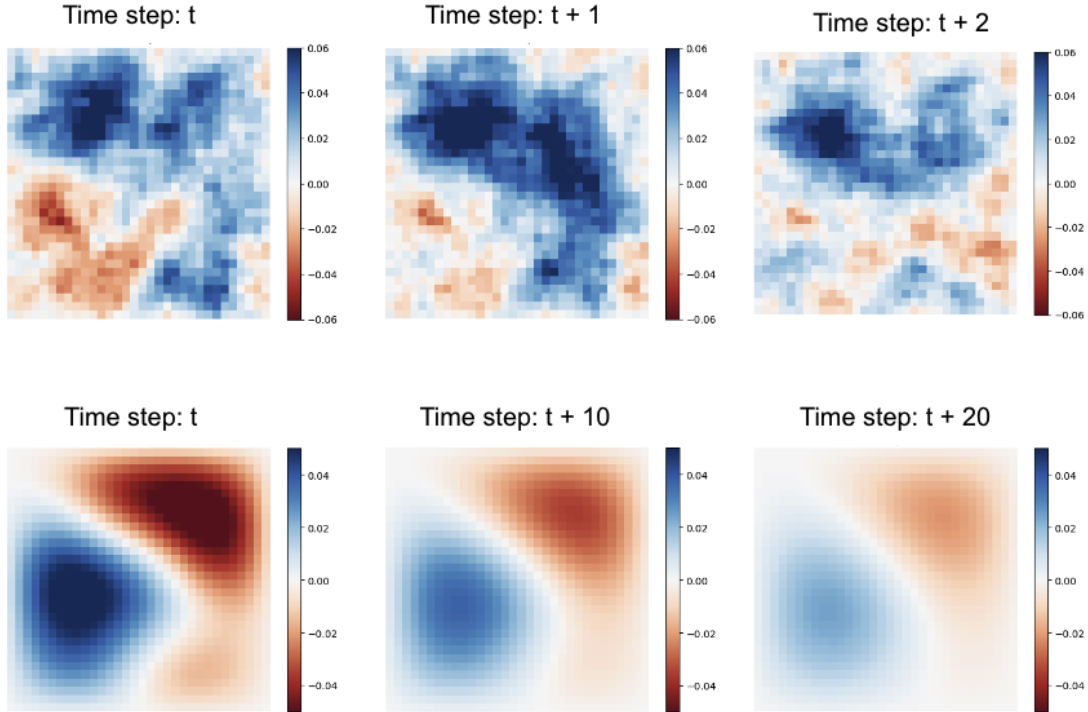
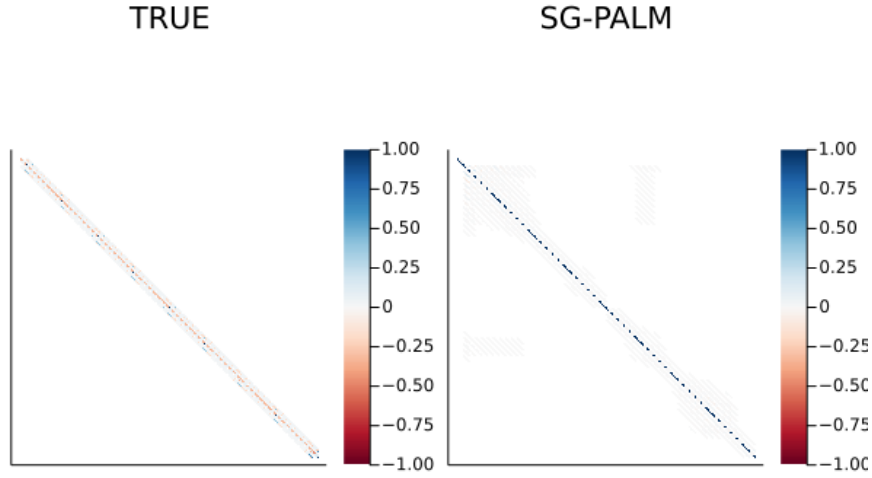
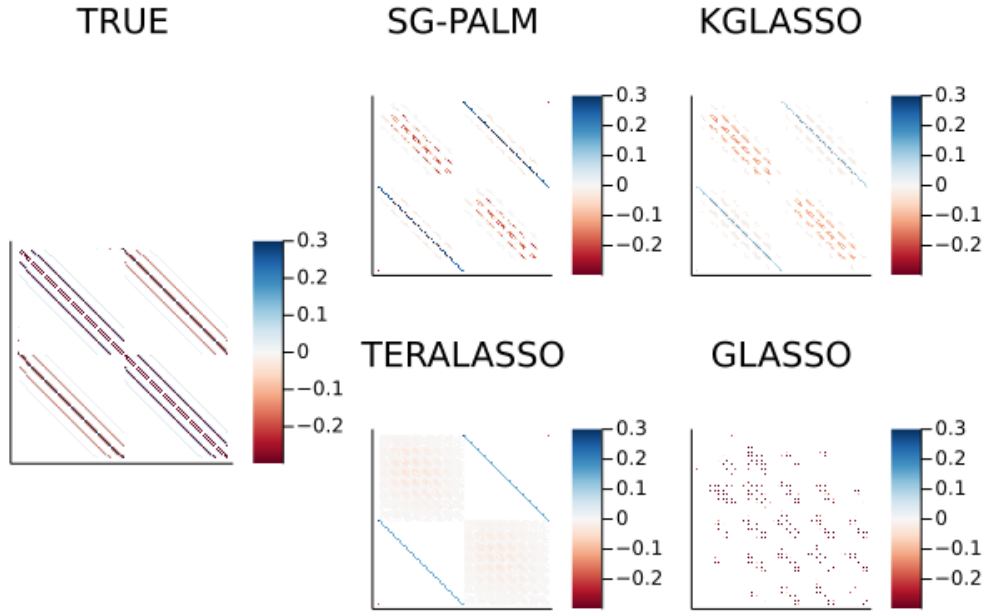


Figure 4.2: Poisson-AR(1) (top) and 2D Convection-diffusion (bottom) state variables of size 64×64 at three different time steps, which are chosen such that t is sufficiently large for the processes to be in steady state and that the sampling times are adjusted to the temporal dynamics of each process. Here, models parameters used for simulations are: $a = -0.5$ for Poisson-AR, $\theta = 0.05$ and $\epsilon = 0.01$ for convection-diffusion process; the driving noise is generated from white Gaussian with $\sigma_w = 0.1$ in both cases.

Computational efficiencies of the various covariance/precision estimation algorithms are

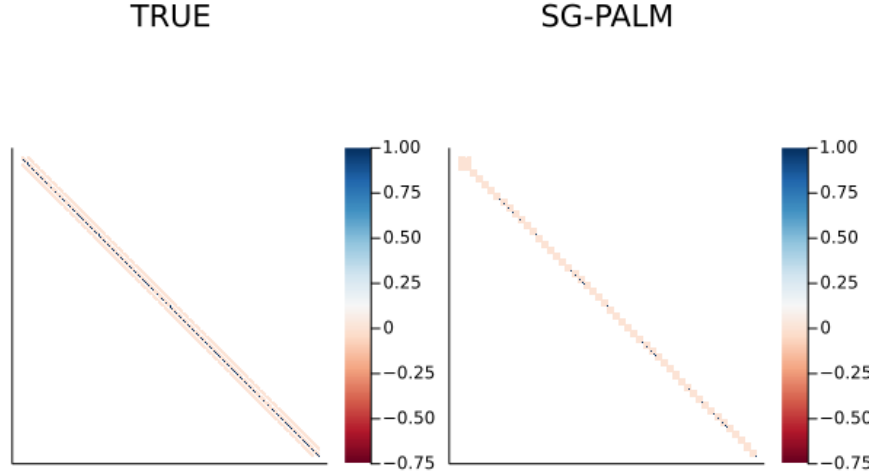


(a) Poisson-AR inverse covariance structure (left) and the estimate obtained by SG-PALM (right).

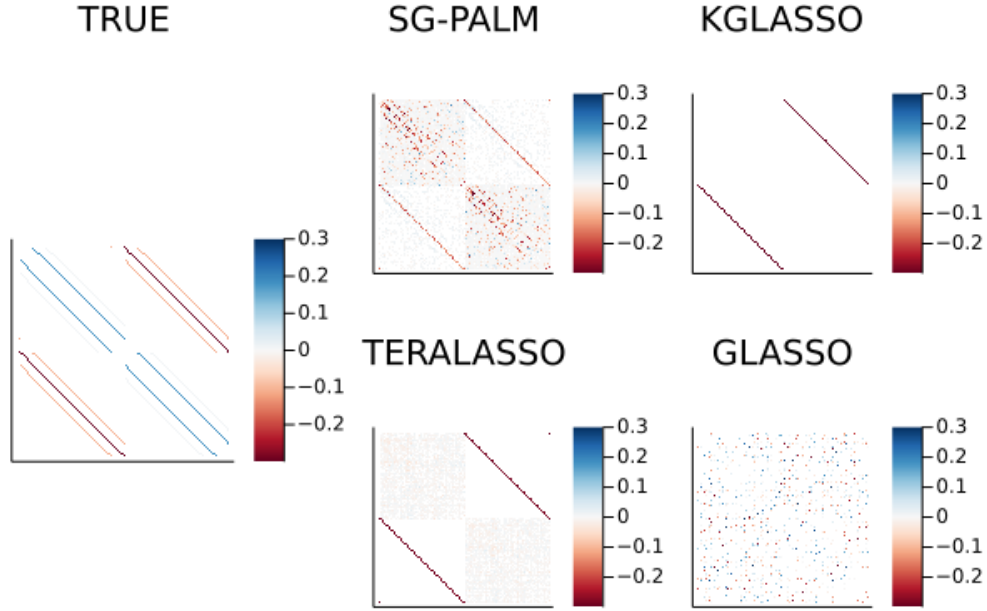


(b) Zoomed-in (middle 128 rows and columns) Poisson-AR inverse covariance structure (left) and the estimate obtained by SG-PALM, KGLasso, Glasso, TeraLasso (right, clockwise).

Figure 4.3: Visualizations of the inverse covariance structures for Poisson-AR(1) dynamics and its estimates. Here, white entries indicate zeros in the inverse covariance matrices. The zoomed-in plots show two temporal blocks (each of size 64×64) of spatial inverse correlation structures with the diagonal elements removed for clearer visualization. SG-PALM and the associated Sylvester graphical model produce the richest structures.

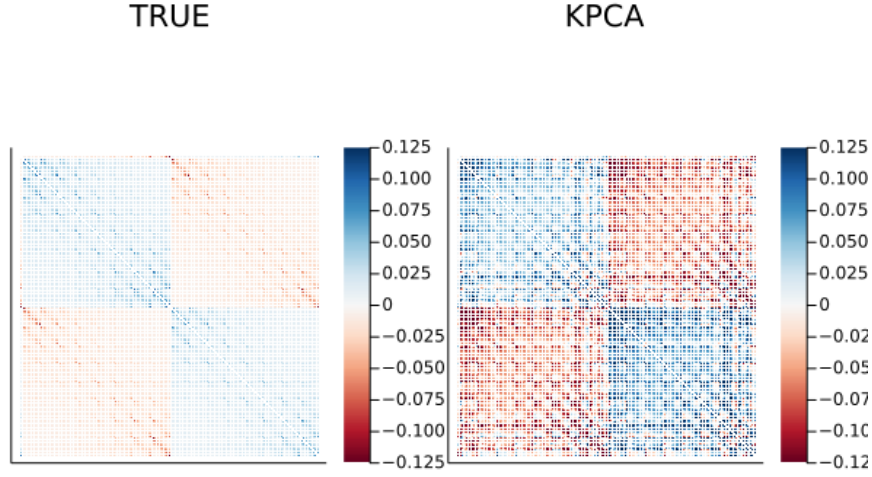


(a) Convection-diffusion inverse covariance structure (left) and the estimate obtained by SG-PALM (right).

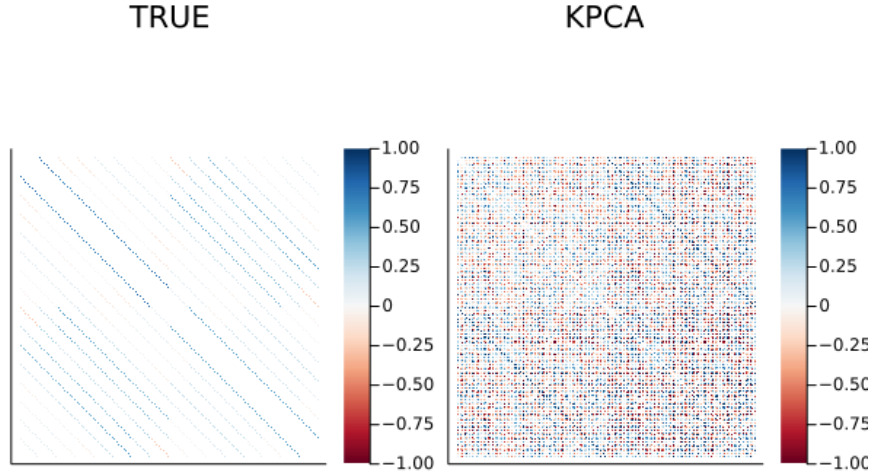


(b) Zoomed-in (middle 128 rows and columns) convection-diffusion inverse covariance structure (left) and the estimate obtained by SG-PALM, KGLasso, Glasso, TeraLasso (right, clockwise).

Figure 4.4: Visualizations of the inverse covariance structures for the Convection-Diffusion dynamics and its estimates. Here, white entries indicate zeros in the inverse covariance matrices. The zoomed-in plots show two temporal blocks (each of size 64×64) of spatial inverse correlation structures with the diagonal elements removed for clearer visualization. SG-PALM and the associated Sylvester graphical model produce the richest structures.



(a) Poisson-AR covariance structure (left) and the estimate obtained by KPCA (right).



(b) Convection-Diffusion covariance structure (left) and the estimate obtained by KPCA (right).

Figure 4.5: Visualizations of the middle 128 rows and columns of the covariance structures for Poisson-AR(1) and Convection-Diffusion dynamics and their estimates, which show two temporal blocks of spatial correlation structures, each of size 64×64 , with the diagonal elements removed for clearer visualization of the pattern. Here, white entries indicate zeros in the covariance matrices. Since the covariances are not sparse in general, all matrices are thresholded for clearer inspections of patterns.

Table 4.3: Comparisons of performances measured by $\log(\|\hat{\Sigma} - \Sigma\|_F / \|\Sigma\|_F)$ for KPCA as well as $\log(\|\hat{\Omega} - \Omega\|_F / \|\Omega\|_F)$ and the Mathews Correlation Coefficient (MCC) for SG-PALM, Tlasso, TeraLasso, Glasso. The MCC is a measure of the quality of sparsity recovery considered as a binary classification problem, where ± 1 indicates perfect agreement or disagreement between the truth and the estimation. Here the Frobenius norm errors are included in the first row under each generating type while the MCCs are in the second row. Note that the best performers under each type/criteria are highlighted.

Type	Metric	SG-PALM	KGlasso	TeraLasso	Glasso	KPCA
P-AR	Fnorm	-0.2622	1.1777	0.6312	0.9775	0.3289
	MCC	0.4300	0.3395	0.2061	0.0560	N/A
C-D	Fnorm	-0.0420	1.4919	-0.0208	2.2041	0.0642
	MCC	0.2122	0.1884	0.2018	0.0349	N/A

also vitally important in practice to facilitate real-time tracking of physical systems. Table 4.4 shows the runtime of different covariance and inverse covariance estimation algorithms for the synthetic experiments. It shows that by recognizing and exploiting multiway structures in the data, sparse multiway inverse covariance estimation methods, TeraLasso, Tlasso, and SG-PALM significantly reduce the runtime complexity of Glasso that ignores such special multiway structures. Remark that KPCA runs considerably slower than other methods as it involves expensive singular value decomposition of a large-dimensional re-arranged sample covariance matrix of the data.

Table 4.4: Average runtime and its standard error (in seconds) of estimating spatio-temporal (inverse) covariance matrices of size $d \times 50$, where d is varying, using various algorithms. Comparisons under various problem sizes (i.e., different d and N) are shown. Note the sparse multiway precision models (SG-PALM, KGlasso, TeraLasso) are comparably fast and are all faster than Glasso (for large problems) and KPCA.

d	N	Glasso	SG-PALM	TeraLasso	KGlasso	KronPCA
		sec	sec	sec	sec	sec
8^2	25	0.40(0.20)	0.46(0.15)	0.15(0.35)	0.65(0.11)	37.22(0.20)
	50	0.48(0.21)	0.47(0.08)	0.22(0.50)	0.70(0.10)	38.22(0.55)
	100	0.76(0.05)	0.44(0.13)	0.26(0.28)	0.69(0.30)	39.09(1.05)
16^2	25	6.43(1.45)	3.37(1.09)	5.38(0.58)	5.14(1.99)	495.47(2.69)
	50	9.12(0.98)	3.27(1.81)	4.62(1.98)	3.39(2.00)	516.64(2.19)
	100	11.84(2.01)	4.85(1.10)	6.71(0.72)	5.67(0.57)	498.04(4.01)

4.4.3 Application to Predicting Evolution of Solar Active Regions

Solar active regions are temporary centers of strong and complex magnetic field on the sun, the principal source of violent eruptions such as solar flares [294]. While weak flares of, for example, B-class, have only limited terrestrial effect, strong flares of M- and X-class can produce tremendous amount of electromagnetic radiation, causing disturbance or damage to satellites, power grids, and communication systems. Therefore, it would be great value to be able to predict how active regions evolve before the onset of solar flares.

Although there are numerous studies that use active region images or physical parameters to predict flare activities [169, 49, 141, 301, 269], fewer studies have attempted to predict the complicated preflare evolution of active regions without physical modeling [19]. In this section, to demonstrate the application of the tensor graphical models, we use multiwavelength active region observations to predict the evolution of two types of active regions that lead to either a weak (B-class) flare or a strong (M- or X-class) flare ².

We construct a multiwavelength active region video dataset from the curated dataset generated by Galvez et al. [88]. The video data are taken in four wavelengths (94Å, 131Å, 171Å, and 193Å) by the Atmospheric Imaging Assembly [AIA, 176] aboard the Solar Dynamics Observatory (SDO) satellite. Each video is a 24-hour image sequence of an active region at 1-hour cadence before a strong (M- or X-class) or a weak (B-class) flare occurs in the region. We spatially interpolate the videos so that each video is represented as a $d_1 \times d_2 \times d_3 \times d_4$ tensor, where $d_1 = 13$ denotes the number of frames in the video, $d_2 = 50$ denotes the height of the frames after interpolation, $d_3 = 100$ denotes the width of the frames after interpolation, and $d_4 = 4$ represents the number of different channels/wavelength at which the images are recorded. To prevent information leakage, we chronologically split the active region videos into a training set (year 2011 to 2014) and a test set (year 2015 to 2014). In the training set, there are 186 active region videos that lead to a B-class flare and 48 active region videos that lead to a M/X-class flare. In the test set, the sample sizes are 93 and 24 for the B-class and the M/X-class, respectively.

To perform active region prediction, we first fit the tensor graphical models on the training set to estimate the covariance or prediction matrices for each of the two types of active region videos, and then we use the best linear predictor to predict the last frame from all previous frames for videos in the test set. The forward linear predictor is constructed in a multi-output least squares regression setting as

$$\hat{\mathbf{y}}_t = -\mathbf{\Omega}_{2,2}^{-1}\mathbf{\Omega}_{2,1}\mathbf{y}_{t-1:t-(p-1)} \quad (4.18)$$

²A similar and more comprehensive version of this numerical study was included in Wang and Hero [305]

when the precision estimate is available. Here, $t = d_1$ for predicting the last frame of a video. For notational convenience, let $p = d_1$ and $q = d_2 d_3 d_4$, then $\mathbf{y}_{t-1:t-(p-1)} = \mathbf{y}_{p-1:1} \in \mathbb{R}^{(p-1)q}$ is the stacked set of pixel values from the previous $p - 1$ time instances and $\mathbf{\Omega}_{2,1} \in \mathbb{R}^{q \times (p-1)q}$ and $\mathbf{\Omega}_{2,2} \in \mathbb{R}^{q \times q}$ are submatrices of the $pq \times pq$ estimated precision matrix:

$$\hat{\mathbf{\Omega}} = \begin{pmatrix} \mathbf{\Omega}_{1,1} & \mathbf{\Omega}_{1,2} \\ \mathbf{\Omega}_{2,1} & \mathbf{\Omega}_{2,2} \end{pmatrix}.$$

The tensor graphical models to be evaluated are SG-PALM (Section 4.3.5), Tlasso (Section 4.3.4), and TeraLasso (Section 4.3.3), with regularization parameters chosen by optimizing the prediction NRMSE (i.e., the root mean squared error normalized by the difference between maximum and minimum pixels) on the training set over a range of λ values guided by the convergence rates presented in Table 4.2. In particular, the SG-PALM estimator was implemented using a regularization parameter $\lambda_N = C_1 \sqrt{\frac{\min(d_k) \log(d)}{N}}$ for all k with the constant C_1 chosen by optimizing the prediction NRMSE on the training set over a range of λ values parameterized by C_1 . The TeraLasso estimator and the Tlasso estimator were implemented using $\lambda_{N,k} = C_2 \sqrt{\frac{\log(d)}{N \prod_{i \neq k} d_i}}$ and $\lambda_{N,k} = C_3 \sqrt{\frac{\log(d_k)}{Nd}}$ for $k = 1, 2, 3$, respectively, with C_2, C_3 optimized in a similar manner. Also considered is a baseline estimator, referred to as IndLasso, obtained by applying independent and separate ℓ_1 -penalized regressions to each pixel in $\mathbf{y}_t$. Each sparse regression in the IndLasso estimator was implemented and tuned independently with regularization parameters chosen in a similar manner. Note that we did not include Glasso and KPCA (Section 4.3.2) in the comparison due to their prohibitively expensive runtime complexity for large multiway data.

Table 4.5 shows the NRMSE over the testing samples for the forecasts based on SG-PALM, Tlasso, TeraLasso, and IndLasso. We observe that SG-PALM outperforms all three other methods, indicated by the lowest NRMSEs on the test set. Figure 4.6 depicts examples of predicted images, comparing with the ground truth. The SG-PALM estimates produced most realistic image predictions that capture the spatially varying structures and closely approximate the pixel values (i.e., maintaining contrast ratios). The latter is important as the flares are being classified into weak (B-class) and strong (M/X-class) categories based on the brightness of the images, and stronger flares are more likely to lead to catastrophic events, such as those that can lead to damaging terrestrial perturbations and health of astronauts.

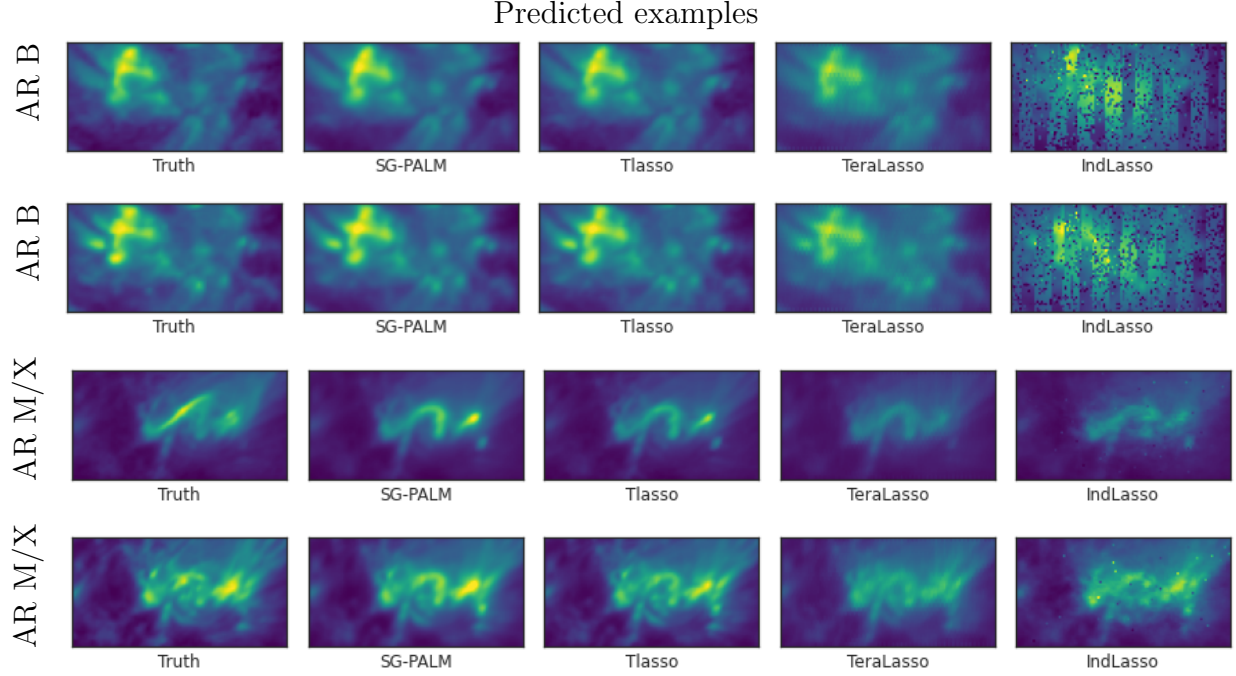


Figure 4.6: Examples of one-hour ahead prediction of the first two AIA channels of last frames of 13-frame videos, leading to B- (first two rows) and M/X-class (last two rows) flares, produced by the SG-PALM, Tlasso, TeraLasso, IndLasso algorithms, comparing to the ground truth image (far left column). The column ordering from left to right is in the order of increasing normalized mean squared error (NMSE) - see Table 4.5. Note that in general linear forward predictors tend to underestimate the contrast ratio of the images. The proposed SG-PALM produced the best-quality images in terms of both the spatial structures and contrast ratios. This figure is excerpted from Wang and Hero [305] (cf. Figure 3 in the reference).

Table 4.5: Comparison of the SG-PALM, Tlasso, TeraLasso, IndLasso performances measured by NRMSE in predicting the last frame of 13-frame video sequences leading to B- and M/X-class solar flares. The NRMSEs are computed by averaging across testing samples and AIA channels for each pixel. SG-PALM achieves the best overall NRMSEs across pixels. B flares are generally easier to predict due to both a larger number of samples in the training set and typically smoother transitions from frame to frame within a video.

Flare Type	SG-PALM	Tlasso	TeraLasso	IndLasso
	NRMSE	NRMSE	NRMSE	NRMSE
B	0.0379	0.0386	0.0579	0.1628
M/X	0.0620	0.0790	0.0913	0.1172

4.5 Conclusion and discussion

Physical interpretability. While the Kronecker products expansion used in KPCA captures dense structures in the covariance matrix of data generated from more complex spatio-temporal physical processes, it lacks physical interpretability. In contrast to the case of Sylvester graphical model and Poisson-AR(1) processes, it is not obvious whether the sum of Kronecker products structure corresponds to any true physical models. Recent work in quantum informatics [52] has demonstrated a link between estimation of the density matrix for entangled quantum states and the structured tensor approximation via $\sum_{i=1}^r \mathbf{A}_i \otimes \mathbf{B}_i$. Further characterizing and extending these connections to other classes of discretized PDEs is an interesting future direction. Furthermore, in both the Poisson-AR(1) and convection-diffusion examples, a mixed Kronecker sum and Kronecker product structure emerges that can be related to the state covariance of a dynamical system.

Heavy-tailed multiway covariance/precision models. Most existing work on multiway covariance and inverse covariance models focus on modeling Gaussian variables. It would be interesting to explore whether the pseudo-likelihood framework we adopted for SyGlasso and SG-PALM can be extended to non-Gaussian heavy-tailed models, e.g., using copula's or elliptically contoured distributions. This could have important practical applications, in particular to solar flare and active region prediction problems presented in Section 4.4.3. The images that characterize the active regions generally include a small number of pixels of extreme high-intensity. These pixels might not be captured by a Gaussian-like distribution. Recently, there have been advances [311, 149] in covariance estimation for heavy-tailed, non-Gaussian vector-variate data. Multiway (inverse) covariance estimation is an open problem.

Kronecker-structured autoencoders. Low-rank covariance models have close connections with variational autoencoders (VAEs). Dai et al. [60] studied the relationship between (robust) PCA and VAEs. Since the Kronecker product for matrices is a generalization of the outer product for vectors, KPCA can be considered as a generalization of a the low-rank approximation method of PCA. It is thus natural to exploit similar relationships between KPCA and VAEs. In this case VAE may be considered as a nonlinear/non-Gaussian extension to KPCA for low separation rank covariance models. Additionally, recent advances in efficient training of the VAE-type neural network architecture (e.g., using stochastic gradient descent) could improve the computational complexity of KPCA that is currently limited by an expensive singular value decomposition [289, 100].

Computational complexity. Estimating the Kronecker product model is generally hard due to: (i) tensor factorization is NP-hard [124] when $K \geq 3$; and (ii) the optimization objective is nonconvex. The NP-hardness implies computational intractability for large scale problems in the worst case. However, this does not necessarily preclude the existence of tractable solutions for most cases. Specifically, it is believed that the model can be estimated within polynomial time when the signal-to-noise ratio (SNR) is sufficiently high³, but only when it is significantly higher than the information-theoretical threshold for model identifiability [240, 130]. While such studies have been completed for some important problems e.g, the statistical-computational gap [44, 340] and the optimization landscape [91], an analogous framework for Kronecker covariance remains to be developed.

While Kronecker sum models do not suffer from the nonconvexity of the Kronecker product model, scalable estimation methods that apply to very large covariance and precision matrices remains computationally challenging. Therefore, it would be worthwhile to develop improvements for the Kronecker sum that mirror the successes in related problems, e.g., introducing more amenable surrogate objectives[22, 150].

Software. A software package called `TensorGraphicalModels` [308] has been developed to accompany this review article. Detailed code and instructions for reproducing the numerical results presented in this article (and beyond) are provided in the reference. The software provides a suite of tools in `Julia`, which was designed specifically for numerical and scientific computing and has recently become popular for large-scale scientific simulation and modeling, such as those conducted by climate scientists [225].

³For example, the convergence analysis of Tlasso [192] requires a ‘good’ initialization; a high SNR is required for a random initialization to satisfy the condition.

CHAPTER 5

Julia Software for Kronecker Models

5.1 Introduction

Many datasets are high-dimensional and multiway in nature, often involving a huge number of multi-indexed variables represented as tensors [153, 163]. Second order statistical analysis of such data poses significant modeling and computational challenges [124]. To overcome these challenges, several methodological approaches have been developed over the past 50 years that reduce model complexity by imposing sparse or low-dimensional dependency structures in the covariance or precision matrices. These include Kronecker generalizations of PCA (Kronecker PCA) [289, 103] and the matrix normal model [62], where the data are modeled as multivariate Gaussian variables with the covariance represented as a Kronecker product and/or sum of matrices of much lower dimensions. These Kronecker representations can have Kronecker factors that are either dense [76, 315] or sparse [10, 290, 102, 192, 306, 305]. These models give the analyst flexibility for exploring various degrees of sparsity and separability of the covariance or precision matrices. Sparse separable precision matrix structure characterizes a range of physical processes that satisfy an ODE or PDE, e.g., heat equations or magnetic hydrodynamic equations, when discretized in time or space [304]. Thus, when viewed through the lens of Gaussian process models, these covariance models are instances of physics-based machine learning models. While, much of the machine learning community focuses on black-box models like deep neural networks [275, 191, 343, 298, 238, 300], shallower models that capture mechanism in a physically interpretable manner are of increasing interest [312, 237, 39].

This chapter introduces `TensorGraphicalModels`, a suite of tools in Julia that provides a unified framework for estimating multiway covariance and precision matrices from tensor-variate Gaussian data and integrating them into a dynamical systems nonlinear predictor model, the ensemble Kalman filter. `TensorGraphicalModels` has three main features. First, it has been developed in a modern high-performance scientific programming language

– **Julia**. **Julia** was designed to combine the convenience of high level rapid prototyping languages, such as **Python**, **Matlab** and **R**, with the optimized runtime performance of a low level language like **C** and **C++** [23]. It has become increasingly popular in the scientific computing community and, in particular, for large-scale scientific simulation and modeling in the climate sciences [225, 77, 307, 281]. The development of this new language was intended to embrace interdisciplinary researchers and practitioners from all areas of data-driven science. There are many high-quality, mature libraries for numerical computing already written in **C** and **Fortran** that can be directly called from **Julia**. Furthermore, **Julia** has a “no boilerplate” philosophy [23]: functions can be called directly from **Julia** without any “glue” code, code generation, or compilation – even from the interactive prompt. A second feature is that **Julia** utilizes native multi-threaded linear algebraic libraries together with Intel’s Math Kernel Library (MKL) backend, which offers sparse functionality to speed up operations on sparse arrays. For example, sparse/dense and sparse/sparse matrix-matrix and matrix-vector multiplications are greatly accelerated. This is important for efficient implementations of the underlying optimization and numerical routines for Gaussian graphical models and Kalman filters, as they rely on gradient-based iterative solvers that are rich in sparse matrix and tensor operations. A third feature is that **Julia** has an scientific machine learning (SciML) ecosystem that has been rapidly growing. This facilitates the integration of physical models that require formatting and solving (partial) differential equation systems. Our **TensorGraphicalModels** package exploits this ecosystem using, for example the **Julia** packages **DifferentialEquations.jl** [235] and **ModelingToolkit.jl** [193] to efficiently implement the ensemble Kalman filter that accomplishes data assimilation with PDE-driven dynamical models.

5.2 Impact on covariance model development

The software we develop has impact on two different but closely related lines of research: inference for high-dimensional covariance and precision matrices and forecasting of dynamical processes.

Multiway covariance/precision models. High-dimensional multiway/tensor-variate data arise naturally in physical sciences. For example, weather satellites measure spatio-temporal climate variables such as temperature, wind velocity, sea level, pressure, etc. Due to the non-homogeneous nature of these data, second-order information that encodes (conditional) dependency structure within the data is of interest. Assuming the data are drawn from a tensor normal distribution, a straightforward way to estimate this structure is to vectorize the tensor and estimate the underlying Gaussian graphical model associated with

the vector. However, such an approach ignores the tensor structure and requires estimating a rather high dimensional precision matrix, often with insufficient sample size. In many scientific applications the sample size can be as small as one when only a single tensor-valued measurement is available.

To address sample complexity challenges in learning second-order representations for multiway (tensor) data, sparsity is often imposed on the covariance matrix Σ or the precision matrix $\Omega = \Sigma^{-1}$. Such approaches include the sparse Kronecker product (KP) or Kronecker sum (KS) decomposition of Σ or Ω . Statistical models and corresponding learning algorithms can be derived using generative models or matrix approximations. The former include: KGlasso/Tlasso [290, 192] for estimating $\Omega = \mathbf{A} \otimes \mathbf{B}$, using a representation $\mathbf{A}\mathbf{X}\mathbf{B} = \mathbf{Z}$ for data $\mathbf{X}$ when $\mathbf{Z}$ is white noise. Another generative model is SyGlasso/SG-PALM [306, 305] that models the precision matrix as $\Omega = (\mathbf{A} \oplus \mathbf{B})^2$, which corresponds to assuming the data $\mathbf{X}$ obeys a Sylvester equation $\mathbf{X}\mathbf{A} + \mathbf{B}\mathbf{X} = \mathbf{Z}$. Matrix approximation methods include: KPCA [289, 100] that approximates the covariance matrix as $\Sigma = \sum_{i=1}^r \mathbf{A}_i \otimes \mathbf{B}_i$. Another matrix approximation method is the TeraLasso [102] that models the precision matrix as $\Omega = \mathbf{A} \oplus \mathbf{B}$. TeraLasso is equivalent to approximation of the conditional dependency graph (encoded by the precision matrix) with a Cartesian product of smaller graphs ¹.

Our software implements the state-of-the-art estimation algorithms for all contemporary tensor-variate Gaussian graphical models. It also features synthetic data simulation routines, which facilitates quick comparison and illustration of various models. As a first of its kind Gaussian graphical modeling packages in the Julia ecosystem, the proposed software opens doors for new researchers and practitioners interested in the field. As an example, the following code snippet generates data from a multiway KP model and estimates its precision matrix using KGlasso included in the software:

```

1  using TensorGraphicalModels
2  model_type = "kp"
3  # this defines the structure of the Kronecker factors, sb = star-block
4  sub_model_type = "sb"
5  K = 2
6  N = 1000
7  d_list = [25, 30]
8  # multi-dimensional array (tensor) of dimension d_1 x ... x d_K x N
9  X = TensorGraphicalModels.gen_kronecker_data(model_type, sub_model_type
10         , K, N, d_list)
10 Psi_hat_list = kglasso(X)

```

¹Note that Tlasso, TeraLasso, Syglasso/SG-PALM are generalizable to precision matrices of the form $\otimes_{k=1}^K \mathbf{A}_k$, $\oplus_{k=1}^K \mathbf{A}_k$, and $(\oplus_{k=1}^K \mathbf{A}_k)^2$, respectively, for $K \geq 2$.

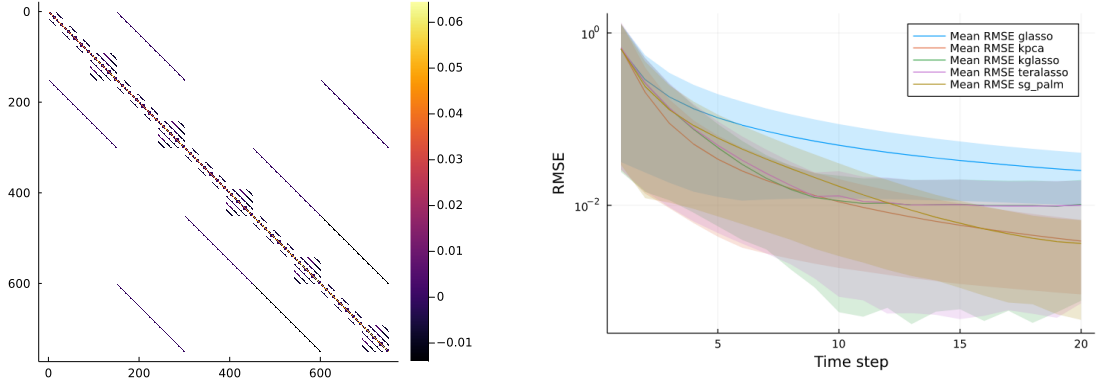


Figure 5.1: Visualization of an estimated multiway sparse precision matrix (left) and accuracy of multiway PEnKF algorithm (right), using the proposed software.

The estimated sparse KP structured matrix can be visualized on the left plot of Figure 5.1.

Multiway second-order characterization of dynamic processes. Physical systems often exhibit sparsity and low-rank structures in their covariance or precision matrix. This is due to the fact that many physical systems are governed by differential equations, which are characterized by sparse differential operators. For instance, [305] showed that for multivariate data generated by the Poisson equation, the discretized data has a precision matrix equal to a squared Kronecker sum of smaller sparse matrices. In related work, [183] elucidated a link between certain classes of Gaussian Fields (GF) and Gaussian Markov Random Fields (GMRF) via stochastic partial differential equations, and showed that efficient learning algorithms can be developed using the fact that GMRFs have sparse precision matrices.

For multiway/tensor-variate Gaussian data, the aforementioned multiway (inverse) covariance estimators have been shown to be statistically consistent in high-dimensional regimes when the sample size (N) is much smaller than the dimensionality (d) of the covariates. An important question is whether these Kronecker structures can be integrated into the Kalman filter for tracking the states of a physical system that generates multiway data.

The package `TensorGraphicalModels` implements the ensemble Kalman filter (EnKF) method for data assimilation that is particularly useful for tracking the time evolution of a usually non-linear physical system. It does so by using an empirical approximation to the well-known Kalman filter. Specifically, we implement a penalized version (PEnKF) proposed by [133] and [304] that use sparse estimators for the state (inverse) covariance, which constitutes an important step of the filtering algorithm. The sparse and Kronecker-structure regularization allow for both an efficient computation and hence fast temporal evolution of the system and elimination of spurious correlations [104] that are often present

in high-dimensional regime.

An example code snippet that generates a dynamical process governed by a convection-diffusion equation is given by

```
1 using TensorGraphicalModels
2 #specify the dynamical model: convection_diffusion, poisson, or
   poisson_ar
3 dynamic_type = "convection_diffusion"
4 #specify the observation model: identity, linear_perm, or
   linear_perm_miss
5 obs_type = "linear_perm_miss"
6 T = 20 #time stamps
7 N = 50 #number of ensembles
8 px = py = (32, 32) #state dimension
9 obs_noise = 0.01
10 process_noise = 0.01
11 add_process_noise = false
12 X, Y, H = TensorGraphicalModels.gen_kalmanfilter_data(dynamic_type,
   obs_type,
13     T, px, py, obs_noise, process_noise, add_process_noise)
```

Then, the proposed multiway PEnKF approach can be applied using various multiway covariance and precision matrix estimators, using the following code, for example:

```
1 method_list = ["glasso", "kpca", "kglasso", "teralasso", "sg_palm"]
2 NRMSEs_list = []
3 for method in method_list
4     ## run enfk
5     Xhat, Xhat_bar, Omegahat = enfk(
6         Y,
7         TensorGraphicalModels.method_str_to_type(method),
8         dynamic_type,
9         H,
10        px,
11        py,
12        N,
13        obs_noise,
14        process_noise,
15        add_process_noise
16    )
17    ## compute NRMSEs
18    NRMSEs = TensorGraphicalModels.compute_nrmse(X, Xhat)
19    push!(NRMSEs_list, NRMSEs)
20 end
```

The root mean square errors (RMSE) of the estimated state propagation in the example given can be visualized on the right plot of Figure 5.1.

5.3 Ongoing research and further development

There is a large range of practical problems for which Kronecker covariance decomposition models and our software have application. Some applications where such models have been applied include: longitudinal gene expression [288], EEG signal analysis [276, 306, 288], spatial imaging [66], radar imaging [101, 266], radar target tracking [324], and analysis of healthcare data [119]. It is hoped that the computationally efficient and accessible `Julia` notebooks provided with this chapter will help further expand the impact on applications of these powerful techniques.

An ongoing line of research that could benefit from the Kalman filtering functionality of our software is space weather prediction. Specifically, global maps of the solar photospheric magnetic field are fundamental drivers for simulations of the corona, solar wind, and therefore geoeffective events [123]. However, since there is no direct measurement from the far-side of the sun, data assimilation methods have to be used to infer the synchronic global maps of solar photospheric magnetic field. Photospheric flux transport models [302] evolves the magnetic flux on the sun by modeling differential rotation, meridional flow, supergranular diffusion, flux cancellation, and small scale random flux emergence. Different forms of Photospheric flux transport models exist, but they are mostly different forms of convection-diffusion equations. The Air Force Data Assimilative Photospheric Flux Transport (ADAPT) model [14] develops a framework of data assimilation that uses the photospheric flux transport dynamics. ADAPT includes several data assimilation techniques including ensemble least squares (EnLS), ensemble transform Kalman filters (ETKF), and local ensemble transform Kalman filters (LETKF). In particular, ADAPT using LETKF developed by [123] uses a tapered covariance matrix to suppress spurious long-distance correlations due to the small ensemble sizes. However, it is not obvious how to determine the local area of interest that are used in assimilation. On the other hand, tensor graphical models offer an alternative to the artificial assumption of a tapered covariance matrix, replacing it by a more natural assumption that the precision matrix is sparse and separable. Such precision matrix structure captures the Markovian and Cartesian or direct product structure of processes generated discretized PDEs. Thus, integration of our software into the ADAPT framework may be able to improve inference of the synchronic global maps of the solar photospheric magnetic field.

We conclude with several areas of future development of the software:

- **Heavy-tailed multiway covariance/precision models.** Most existing research on multiway covariance and precision matrix models focus on modeling Gaussian variables. We are exploring possible extensions of `TensorGraphiclModels` to non-Gaussian heavy-tailed models using copula and elliptically contoured distributions.
- **User-defined dynamical models** The EnKF framework in the package currently allows for testing of the data assimilation methods for physical systems generated by a few specific PDEs. We are working towards an interface that allows for tracking and data assimilation under user-defined dynamical systems.
- **Unified interface with existing machine learning frameworks** Popular machine learning toolkit, such as `sklearn` in Python and `Flux` in Julia use unified workflow for fitting, testing, and prediction for data modeling. One of our ongoing development is to build a similar workflow for analysis using the `TensorGraphiclModels` to allow for easy transitions of researchers and practitioners from using traditional ML libraries.

CHAPTER 6

Recalibration in Binary Classification

6.1 Introduction

Generating reliable probability estimates alongside accurate class labels is crucial in classification tasks. A probabilistic classifier is considered “well calibrated” when its predicted probabilities closely align with the empirical frequencies of the corresponding labels [63]. Calibration is highly desirable, particularly in high-stakes applications such as meteorological forecasting [207, 210, 65, 95], econometrics [96], personalized medicine [139, 136], and natural language processing [215, 43, 68, 345]. Unfortunately, many machine learning algorithms lack inherent calibration [109].

To tackle this challenge, various methods have been proposed for designing methods for recalibration, i.e., post-hoc correction of the output of a probabilistic classifier. These functions are used to assess calibration error [342, 231, 105, 28], detect miscalibration [166], and provide post-hoc recalibration [337, 338, 109, 293, 162, 111]. Despite the rapid development of recalibration algorithms, there is still a lack of a comprehensive theory that encompasses both calibration and sharpness (retaining predictive power) from a principled standpoint. Furthermore, existing methods often rely on diverse calibration metrics [105, 28], and the selection of hyperparameters is often based on heuristic approaches [245, 110] without rigorous justifications. This highlights the need for identifying an optimal metric to evaluate calibration, which can facilitate the development of a unified theory and design principles for recalibration functions.

In addition, the deployment of machine learning models to data distributions that differ from the training phase is increasingly common. These distribution shifts can occur naturally due to factors such as seasonality or other variations, or they can be induced artificially through data manipulation methods such as subsampling or data augmentation. Distribution shifts pose challenges to the generalization of machine learning models. Therefore, it becomes necessary to adapt the trained model to these new settings. One significant category of

distribution shifts is label shift, where the marginal probabilities of the classes differ between the training and test sets while the class conditional feature distributions remain the same. With calibrated probabilistic predictions, the impact of label shift can be corrected assuming access to class marginal probabilities [79]. However, miscalibration combined with label shift is common and remains a challenging problem [8, 90, 285, 230].

In this chapter, we aim to address these issues in a twofold manner. Firstly, we develop a unified framework for recalibration that incorporates both calibration and sharpness in a principled manner. Secondly, we propose a composite estimator for recalibration in the presence of label shift that converges to the optimal recalibration. Our framework enables the adaptation of a classifier to the label-shifted domain in a sample-efficient manner.

6.1.1 Related work

Recalibration algorithms. Recalibration methods can be broadly categorized into parametric and nonparametric approaches. Parametric methods model the recalibration function in a parametric form and estimate the parameters using calibration data. Examples of parametric methods include Platt scaling [229], temperature scaling [109], Beta calibration [160], and Dirichlet calibration [161]. However, it has been reported that scaling methods are often less calibrated than supposed, and quantifying the degree of miscalibration can be challenging [162]. In contrast, nonparametric recalibration methods do not assume a specific parametric form for the recalibration function. These methods include histogram binning [337], isotonic regression [338], kernel density estimation [342, 231], splines [114], Gaussian processes [314], among others. Hybrid approaches, integrating both parametric and nonparametric techniques, have also been proposed. For instance, Kumar et al. [162] combine nonparametric histogram binning with parametric scaling to reduce variance and improve recalibration performance. Nevertheless, this hybrid approach is biased when its parametric assumptions fail. In this work, we consider a nonparametric histogram binning method called uniform-mass binning (UMB), which is asymptotically unbiased.

Histogram binning method. Histogram binning methods are widely used for recalibration due to their simplicity and adaptability. The binning schemes can be pre-specified (e.g., uniform-width binning [109]), data-dependent (e.g., uniform-mass binning [337]), or algorithm-induced [338]. When selecting a binning scheme, it is crucial to consider the trade-off between approximation and estimation. Coarser binning reduces estimation error (variance), leading to improved calibration, but at the expense of increased approximation error (bias), which diminishes sharpness. Thus, determining the optimal binning scheme and hyperparameters, such as the number of bins (B), remains an active area of research.

[214] proposes a Bayesian binning method, but verifying the priors is often challenging. [245] suggests choosing the largest B that preserves monotonicity, which is heuristic and computationally inefficient. [110] offers a heuristic for choosing the largest B subject to a calibration constraint, lacking a quantitative characterization of sharpness. Our work builds upon the existing upper bounds for calibration risks of binning methods [162, 110] and derived upper bounds for a complementary risk component known as the sharpness risk. We quantitatively characterize the calibration-sharpness tradeoff, which yields an optimal choice for the number of bins that achieves the minimum risk.

Adaptation to label shift. Label shift presents a challenge in generalizing models trained on one distribution (source) to a different distribution (target). As such, adapting to label shift has received considerable attention in the literature [79, 248, 184, 17, 8, 90]. In practical scenarios, it is common to encounter model miscalibration and label shift simultaneously [285, 230]. Empirical observations have highlighted the crucial role of probability calibration in label shift adaptation [8, 80], which is justified by subsequent theories [90]. However, to the best of our knowledge, there has been no prior work that specifically addresses the recalibration with a limited amount of labeled data from the target distribution. Our theoretical analysis points out that using only target labels achieves risk bounds of the same order as the methods using only target features [184, 17, 90].

6.1.2 Contributions

This chapter contributes to the theory of recalibration across three key dimensions.

Firstly, we develop a comprehensive theory for recalibration in binary classification by adopting the mean-squared-error (MSE) decomposition framework commonly used in meteorology and space weather forecasting [37, 209, 59, 270]. Our approach formulates the probability recalibration problem as the minimization of a specific risk function, which can be orthogonally decomposed into calibration and sharpness components.

Secondly, utilizing the aforementioned framework, we derive a rigorous upper bound on the finite-sample risk for uniform-mass binning (UMB) (Theorem 6.1). Furthermore, we minimize this risk bound and demonstrate that the optimal number of bins for UMB, balancing calibration and sharpness, scales on the order of $n^{1/3}$, yielding the risk bound of order $n^{-2/3}$, where n denotes the sample size.

Lastly, we address the challenge of recalibrating classifiers for label shift when only a limited labeled sample from the target distribution is available, a challenging situation for a direct recalibration approach. We propose a two-stage approach: first recalibrating the classifier on the abundant source-domain data, and then transferring it to the label-shifted

target domain. We provide a finite-sample guarantee for the risk of this composite procedure (Theorem 6.2). Notably, to control the risk under ε , our approach requires a much smaller sample size from the target distribution than a direct recalibration on the target sample ($\Omega(\varepsilon^{-1})$ vs. $\Omega(\varepsilon^{-3/2})$, cf. Remark 6.4).

6.1.3 Organization

This chapter is organized as follows. In Section 6.2, we introduce notation and provide an overview of calibration and sharpness. Section 6.3 introduces the notion of minimum-risk recalibration by defining the recalibration risk that takes into account both calibration and sharpness. In Section 6.4, we describe the uniform-mass histogram binning method for recalibration and provide a risk upper bound with rate analysis. We extend our approach to handle label shift in Section 6.5. To validate our theory and framework, we present numerical experiments in Section 6.6. Finally, in Section 6.7, we conclude the chapter with a discussion and propose future research directions.

6.2 Preliminaries

6.2.1 Notation

Let $\mathbb{N}$ and $\mathbb{R}$ denote the set of positive integers and the set of real numbers, respectively. For $n \in \mathbb{N}$, let $[n] := \{1, \dots, n\}$. For $x \in \mathbb{R}$, let $\lfloor x \rfloor = \max\{m \in \mathbb{Z} : m \leq x\}$. For any finite set $\mathcal{S} = \{s_i : i \in [n]\} \subset \mathbb{R}$ and any $k \in [n]$, we let $s_{(k)}$ denote the k -th order statistic, which is the k -th smallest element in $\mathcal{S}$.

Let $(\Omega, \mathcal{E}, P)$ denote a generic probability space. For an event $A \in \mathcal{E}$, the indicator function $\mathbb{1}_A : \Omega \rightarrow \{0, 1\}$ is defined such that $\mathbb{1}_A(x) = 1$ if and only if $x \in A$. We write A happens P -almost surely if $P(A) = 1$. For a probability measure P , define $\mathbb{E}_P$ as the expectation, with subscript P omitted when the underlying probability measure is clear. For a probability measure P and a random variable $X : \Omega \rightarrow \mathbb{R}^d$, let $P_X := P \circ X^{-1}$ be the probability measure induced by X .

Letting $f, g : \mathbb{R} \rightarrow \mathbb{R}$, we write $f(x) = O(g(x))$ as $x \rightarrow \infty$ if there exist $M > 0$ and $x_0 > 0$ such that $|f(x)| \leq Mg(x)$ for all $x \geq x_0$. Likewise, we write $f(x) = \Omega(g(x))$ if $g(x) = O(f(x))$. We write $f(x) \asymp g(x)$ if $f(x) = O(g(x))$ and $g(x) = O(f(x))$. We write $f(x) = \tilde{O}(g(x))$ if there is $k \geq 1$ such that $f(x) = O(g(x) \log^k(g(x)))$.

6.2.2 Calibration and sharpness

Consider the binary classification problem; let $X \in \mathcal{X}$ and $Y \in \mathcal{Y} := \{0, 1\}$ denote the feature and label random variables. Letting P denote a probability measure, a probabilistic classifier is a function $f : \mathcal{X} \rightarrow \mathcal{Z} = [0, 1]$ that estimates the conditional probability, *i.e.*,

$$f(X) \approx P[Y = 1 \mid X]. \quad (6.1)$$

Since estimating the probability conditioned on the high dimensional X is difficult, the notion of calibration captures the intuition of (6.1) in a weaker sense [209, 63, 113].

Definition 6.1. *A function $f : \mathcal{X} \rightarrow \mathcal{Z}$ is (perfectly) calibrated with respect to probability measure P , if*

$$f(X) = P[Y = 1 \mid f(X)] \quad P\text{-almost surely.}$$

Calibration itself does not guarantee a useful predictor. For instance, a constant predictor $f(X) = \mathbb{E}Y$ is perfectly calibrated, but it does not change with the features. Such a predictor lacks sharpness [159], also known as resolution [211], another desired property which measures the variance in the target Y explained by the probabilistic prediction $f(X)$.

Definition 6.2. *The sharpness of a function $f : \mathcal{X} \rightarrow \mathcal{Z}$ with respect to probability measure P refers to the quantity*

$$\text{Var}(\mathbb{E}[Y \mid f(X)]) = \mathbb{E}[(\mathbb{E}[Y \mid f(X)] - \mathbb{E}[Y])^2].$$

The following decomposition [159] of the mean squared error (MSE) suggests why it is desirable for a classifier f to be calibrated and have high sharpness; note that $\text{Var}[Y]$ is a quantity intrinsic to the problem, unrelated to f .

$$\text{MSE}(f) := \underbrace{\mathbb{E}[(Y - f(X))^2]}_{\text{mean-squared error}} = \text{Var}[Y] - \underbrace{\text{Var}(\mathbb{E}[Y \mid f(X)])}_{\text{sharpness}} + \underbrace{\mathbb{E}[(f(X) - \mathbb{E}[Y \mid f(X)])^2]}_{\text{lack of calibration}} \quad (6.2)$$

6.3 Optimal recalibration

For an arbitrary predictor $f : \mathcal{X} \rightarrow \mathcal{Z}$, the aim of recalibration is to identify a post-processing function $h : \mathcal{Z} \rightarrow \mathcal{Z}$ such that $h \circ f$ is perfectly calibrated while maintaining the sharpness of f as much as possible. The calibration and sharpness can be evaluated using the following two notions of risks. We suppress the dependency of risks on f and P when it leads to no confusion.

Definition 6.3. Let $f : \mathcal{X} \rightarrow \mathcal{Z}$ and $h : \mathcal{Z} \rightarrow \mathcal{Z}$. The calibration risk of h over f under P is defined as

$$R^{\text{cal}}(h) = R_P^{\text{cal}}(h; f) := \mathbb{E}_P \left[\left(h \circ f(X) - \mathbb{E}_P[Y \mid h \circ f(X)] \right)^2 \right]. \quad (6.3)$$

Definition 6.4. Let $f : \mathcal{X} \rightarrow \mathcal{Z}$ and $h : \mathcal{Z} \rightarrow \mathcal{Z}$. The sharpness risk of h over f under P is defined as

$$R^{\text{sha}}(h) = R_P^{\text{sha}}(h; f) := \mathbb{E}_P \left[\left(\mathbb{E}_P[Y \mid h \circ f(X)] - \mathbb{E}_P[Y \mid f(X)] \right)^2 \right]. \quad (6.4)$$

Note that the calibration risk $R^{\text{cal}}(h; f) = 0$ if and only if $h \circ f$ is perfectly calibrated, cf. Definition 6.1. The sharpness risk $R^{\text{sha}}(h; f)$ quantifies the decrement in sharpness of f incurred by applying the recalibration map h , and $R^{\text{sha}}(h; f) = 0$ when h is injective [162].

Next we define a comprehensive notion of risk that we will use to evaluate recalibration functions.

Definition 6.5. Let $f : \mathcal{X} \rightarrow \mathcal{Z}$ and $h : \mathcal{Z} \rightarrow \mathcal{Z}$. The recalibration risk of h over f under P is defined as

$$R(h) = R_P(h; f) := \mathbb{E}_P \left[\left(h \circ f(X) - \mathbb{E}_P[Y \mid f(X)] \right)^2 \right]. \quad (6.5)$$

The following proposition shows that the recalibration risk can be decomposed into calibration risk and sharpness risk. The proof is deferred to Appendix 6.8.

Proposition 6.1 (Decomposition of recalibration risk). *For any $f : \mathcal{X} \rightarrow \mathcal{Z}$ and any $h : \mathcal{Z} \rightarrow \mathcal{Z}$,*

$$R_P(h; f) = R_P^{\text{cal}}(h; f) + R_P^{\text{sha}}(h; f). \quad (6.6)$$

Note that $R_P(h) = 0$ if and only if $R_P^{\text{cal}}(h) = 0$ and $R_P^{\text{sha}}(h) = 0$. This happens if and only if $h \circ f$ is calibrated, and the recalibration h preserves the sharpness of f in predicting Y .

If $R(h) = 0$, then we call h an *optimal recalibration function* (or *minimum-risk recalibration function*) of f under P . Let $h_{f,P}^* : \mathcal{Z} \rightarrow \mathcal{Z}$ be the function

$$h_{f,P}^*(z) = \begin{cases} \mathbb{E}_P[Y \mid f(X) = z], & \text{if } z \in \text{supp } P_{f(X)}, \\ 0, & \text{if } z \notin \text{supp } P_{f(X)}. \end{cases} \quad (6.7)$$

Then $R(h_{f,P}^*) = 0$. Indeed, h is a minimum-risk recalibration function of f if and only if $h = h_{f,P}^*$ $P_{f(X)}$ -almost surely.

Problem 6.1 (Recalibration). *Suppose that we have a measurable function $f : \mathcal{X} \rightarrow \mathcal{Z}$ and a dataset $\mathcal{D} = \{(x_i, y_i) : i \in [n]\}$ that is an independent and identically distributed (IID) sample drawn from P . The goal of recalibration is to estimate a recalibration function $\hat{h} \approx h_{f,P}^*$ using f and $\mathcal{D}$.*

Problem 6.1 can be viewed as a regression problem, where we estimate the functional form of $\mathbb{E}[Y | Z]$ from data $\{(z_i, y_i) : i \in [n]\}$, where $z_i = f(x_i)$, $\forall i \in [n]$.

6.4 Recalibration via uniform-mass binning

6.4.1 Uniform-mass binning algorithm for recalibration

Given a dataset of prediction-label pairs $\{(z_i, y_i) \in \mathcal{Z} \times \mathcal{Y} : i \in [n]\}$, the histogram binning calibration method partitions $\mathcal{Z} = [0, 1]$ into a set of smaller bins, and estimate $\mathbb{E}[Y | Z]$ by taking the average in each bin. We consider the uniform mass binning, constructed using quantiles of predicted probabilities.

Definition 6.6 (Uniform mass binning). *Let $S = \{z_i \in [0, 1] : i \in [n]\}$. A binning scheme $\mathcal{B} = \{I_1, I_2, \dots, I_B\}$ is the uniform-mass binning (UMB) scheme induced by S if*

$$I_1 = [u_0, u_1], \quad \text{and} \quad I_b = (u_{b-1}, u_b] \quad \forall b \in [B] \setminus \{1\}, \quad (6.8)$$

where $u_0 = 0$, $u_B = 1$, and $u_b = z_{(\lfloor nb/B \rfloor)}$ for $b \in [B - 1]$.

For our subsequent discussions, we make the following assumptions on the distribution of Y and Z :

- (A1) The cumulative distribution function of Z , denoted by F_Z , is absolutely continuous.
- (A2) $h_{f,P}^*$, defined in (6.7), is monotonically non-decreasing on $\text{supp } P_Z$.
- (A3) There exists $K > 0$ such that if $z_1 \leq z_2$, then $h_{f,P}^*(z_2) - h_{f,P}^*(z_1) \leq K \cdot (F_Z(z_2) - F_Z(z_1))$.

Assumption (A1) is made for the sake of analytical convenience without loss of generality; see discussions in [110, Appendix C]. An important implication of Assumption (A1) is that all intervals in $\mathcal{B}$ are non-empty P_Z -almost surely if $S = \{Z_i : i \in [n]\}$ are IID under P_Z . Assumption (A2) makes sure Z is informative to preserve the rankings of $P[Y | Z]$ [338]. Lastly, Assumption (A3) posits that Z is sufficiently informative that $P[Y = 1 | Z = z]$ does not change too rapidly in any interval I where $P_Z(I)$ is small. This assumption is mild but not trivial; see Appendix 6.9.2.

Now we describe how to construct a recalibration function $\hat{h}$ using UMB.

Algorithm.

1. Given $\{(z_i, y_i) : i \in [n]\}$, and $B \in \mathbb{N}$, let $\mathcal{B} = \{I_1, I_2, \dots, I_B\}$ be the UMB scheme of size B induced by $\{z_i : i \in [n]\}$.
2. Let $\hat{h} = \hat{h}_{\mathcal{B}} : \mathcal{Z} \rightarrow \mathcal{Z}$ be a function such that

$$\hat{h}(z) = \sum_{I \in \mathcal{B}} \hat{\mu}_I \cdot \mathbb{1}_I(z) \quad \text{where} \quad \hat{\mu}_I := \frac{\sum_{i=1}^n y_i \cdot \mathbb{1}_I(z_i)}{\sum_{i=1}^n \mathbb{1}_I(z_i)}, \quad \forall I \in \mathcal{B}. \quad (6.9)$$

6.4.2 Theoretical guarantees

Here we establish a high-probability upper bound on the recalibration risk $R(\hat{h})$ for the UMB estimator, $\hat{h}$ defined in (6.9), which converges to 0 as the sample size n increases to ∞ .

Theorem 6.1. *Let P be a probability measure and $f : \mathcal{X} \rightarrow \mathcal{Z}$ be a measurable function. Suppose that (A1) & (A2) hold. Let $\mathcal{B}$ be the UMB scheme induced by an IID sample of $f(X)$, and let $\hat{h} = \hat{h}_{\mathcal{B}}$ be the recalibration function based on $\mathcal{B}$, cf. (6.9). Then there exists a universal constant $c > 0$ such that for any $\delta \in (0, 1)$, if $n \geq c \cdot |\mathcal{B}| \log\left(\frac{2|\mathcal{B}|}{\delta}\right)$, then with probability at least $1 - \delta$,*

$$R^{\text{cal}}(\hat{h}) \leq \left(\sqrt{\frac{\log(4|\mathcal{B}|/\delta)}{2(\lfloor n/|\mathcal{B} \rfloor - 1)}} + \frac{1}{\lfloor n/|\mathcal{B} \rfloor} \right)^2, \quad \text{and} \quad R^{\text{sha}}(\hat{h}) \leq \frac{2}{|\mathcal{B}|}.$$

In addition, if (A3) also holds, then

$$R^{\text{sha}}(\hat{h}) \leq \frac{8K^2}{|\mathcal{B}|^2}.$$

Remark 6.1. *We note that the upper bound on $R^{\text{cal}}(\hat{h})$ in Theorem 6.1 coincides with the result presented in [110] up to a constant factor in the failure probability. However, Theorem 6.1 provides an additional upper bound on $R^{\text{sha}}(\hat{h})$, thereby effectively managing the overall recalibration risk $R(\hat{h})$.*

Optimal choice of the number of bins $|\mathcal{B}|$. Based on Theorem 6.1, for any fixed sample size n , as the number of bins $B = |\mathcal{B}|$ increases, the calibration risk bound increases, while the sharpness risk bound decreases. This trade-off suggests we may get an optimal number of bins B by minimizing the upper bound for the overall risk, $R(\hat{h}) = R^{\text{cal}}(\hat{h}) + R^{\text{sha}}(\hat{h})$.

For the simplicity of our analysis, we assume n is divisible by B , and Assumption (A3) holds. Firstly, observe that $n/B \geq 2$, and thus, $n/B - 1 \geq n/(2B)$. Also, since $B \geq 1$

and $\delta < 1$, $\log(\frac{4B}{\delta}) \geq 1$, it follows that $\frac{1}{n/B} \leq \sqrt{\frac{1}{2(n/B-1)} \log(\frac{4B}{\delta})} \leq \sqrt{\frac{1}{n/B} \log(\frac{4B}{\delta})}$. Thus, we obtain a simplified risk bound $R(\hat{h}) \leq \zeta(B; n, \delta)$ where $\zeta(B; n, \delta) := \frac{4B}{n} \log(\frac{4B}{\delta}) + \frac{8K^2}{B^2}$. With $\log B$ terms ignored, minimizing $\zeta(B; n, \delta)$ over B yields optimal B^* and resulting risk bounds, respectively:

$$B^* \asymp n^{1/3} \cdot \log^{-1/3}(1/\delta), \quad (6.10)$$

$$R(\hat{h}) \asymp R^{\text{cal}}(\hat{h}) \asymp R^{\text{sha}}(\hat{h}) = O(B^{*-2}) = O\left(n^{-2/3} \cdot \log^{2/3}(1/\delta)\right). \quad (6.11)$$

Furthermore, the asymptotic risk bound in (6.11) implies that $R(\hat{h})$, $R^{\text{cal}}(\hat{h})$, and $R^{\text{sha}}(\hat{h})$ are all bounded by ε with high probability if the sample size

$$n = \Omega(\varepsilon^{-3/2}). \quad (6.12)$$

Comparison with the hybrid method [162]. Kumar et al. [162] proposed a hybrid recalibration approach that involves fitting a recalibration function in a parametric family $\mathcal{H}$, which is then discretized using UMB. Note that the hypothesis class $\mathcal{H}$ may or may not include the optimal recalibration function $h_{f,P}^*$. If $h_{f,P}^* \in \mathcal{H}$, using a similar analysis as above, we derive their high probability risk bound as $O\left(\frac{1}{n} \log \frac{B}{\delta} + \frac{1}{B^2}\right)$ under Assumption (A3), which achieves $\tilde{O}(n^{-1})$ when $B^* \asymp \sqrt{n}$, exhibiting a faster decay than our $R(\hat{h}) = O(n^{-2/3})$. While the faster rate is anticipated from employing parametric methods, we note that when $h_{f,P}^* \notin \mathcal{H}$, their method exhibits inherent bias (approximation error) induced by parametric function fitting, whereas our method is asymptotically unbiased. This distinction is corroborated by numerical simulations in Appendix 6.12.2.

Proof sketch of Theorem 6.1. First, we demonstrate that the uniform-mass binning scheme $\mathcal{B} = \{I_b\}_{b=1}^B$ satisfies two regularity conditions with high probability, when the sample size n is not too small. Specifically, we show that (i) $\mathcal{B}$ is 2-well-balanced [162] with respect to $f(X)$, resulting in B bins having comparable probabilities (Lemma 6.3); and that (ii) the empirical mean in each bin of $\mathcal{B}$ uniformly concentrates to the population conditional mean of Y conditioned on $f(X)$ being contained within the bin (Lemma 6.4). Thereafter, we prove that if $\mathcal{B}$ satisfies these two properties, then the calibration risk R^{cal} and the sharpness risk R^{sha} can be upper bounded as stated in Theorem 6.1; see Lemmas 6.5 and 6.6 (or Lemma 6.7 when (A3) holds). The detailed proof is in Appendix 6.9.

6.5 Recalibration under label shift

This section extends the results from Section 6.4 to address label shift. In Section 6.5.1, we introduce the label shift assumption (Definition 6.7) and reframe the recalibration problem accordingly. We show that the optimal recalibration function in this context can be expressed as a composition of the optimal recalibration function (cf. Section 6.3) and a shift correction function. Building on this observation, we propose a two-stage estimator in Section 6.5.2, where each stage estimates one of the component functions. The composite estimator's overall performance is supported by theoretical guarantees.

6.5.1 Revisiting the problem formulation

Let P and Q denote the probability measures of the source and the target domains, respectively. We assume P and Q satisfy the label shift assumption defined below.

Definition 6.7 (Label shift). *Probability measures P and Q are said to satisfy the label shift assumption if the following two conditions are satisfied:*

(B1) $P[X \in B \mid Y = k] = Q[X \in B \mid Y = k]$ for all $k \in \{0, 1\}$ and all $B \in \mathcal{B}(\mathcal{X})$.

(B2) $P[Y = 1] \in (0, 1)$ and $Q[Y = 1] \in (0, 1)$.

According to Condition (B1), the class conditional distributions remain the same, while the marginal distribution of the classes may change. Condition (B2) requires all classes to be present in the source population, which is a standard regularity assumption in the discussion of label shift [184, 90]; it also posits the presence of every class in the target population.

Optimal recalibration under label shift. Under the label shift assumption between P and Q , we define the label shift correction function $g^* : \mathcal{Z} \rightarrow \mathcal{Z}$ such that

$$g^*(z) = \frac{w_1^* z}{w_1^* z + w_0^* (1 - z)} \quad \text{where} \quad w_k^* = \frac{Q[Y = k]}{P[Y = k]}, \quad \forall k \in \{0, 1\}. \quad (6.13)$$

The conditional probabilities under P and Q can be related [248] as follows:

$$Q[Y = 1 \mid X \in B] = g^*\left(P[Y = 1 \mid X \in B]\right), \quad \forall B \in \mathcal{B}(\mathcal{X}). \quad (6.14)$$

Recall that the optimal recalibration function for a predictor $f : \mathcal{X} \rightarrow \mathcal{Z}$ under probability measure P is defined as $h_{f,P}^*(z) = P[Y = 1 \mid f(X) = z]$; see (6.7). In the presence of a label

shift between P and Q , we may write the optimal recalibration function for f under Q as

$$h_{f,Q}^* = g^* \circ h_{f,P}^* \quad (6.15)$$

because $h_{f,Q}^*(z) \stackrel{(a)}{=} Q[Y = 1 \mid f(X) = z] \stackrel{(b)}{=} g^*(P[Y = 1 \mid f(X) = z]) \stackrel{(c)}{=} (g^* \circ h_{f,P}^*)(z)$, where (a) and (c) follows from the definition of $h_{f,P}^*$ and (b) is due to (6.14).

Recalling the definition of the risk $R_P(h; f)$ from (6.5), we observe that $R_Q(h_{f,Q}^*; f) = 0$, which is consistent with the risk characterization of the optimal recalibration. Our goal is to estimate the optimal recalibration function $h_{f,Q}^* = g^* \circ h_{f,P}^*$ from data.

Problem 6.2 (Recalibration under label shift). *Suppose that we have a measurable function $f : \mathcal{X} \rightarrow \mathcal{Z}$ and two IID datasets $\mathcal{D}_P = (x_i, y_i)_{i=1}^{n_P} \sim P$ and $\mathcal{D}_Q = (x'_i, y'_i)_{i=1}^{n_Q} \sim Q$. The goal of recalibration under label shift is to estimate $\hat{h} \approx h_{f,Q}^*$ using f , $\mathcal{D}_P$ and $\mathcal{D}_Q$.*

Remark 6.2. *The source (training) dataset $\mathcal{D}_P$ may not be accessible due to privacy protections, proprietary data, or practical constraints, as is often the case when recalibrating a pre-trained black box classifier to new data. In these cases, it suffices to have estimates of the recalibration function $h_{f,P}^*$ and the marginal probabilities $P[Y = k]$, $k \in \{0, 1\}$ under P , for our method and analysis.*

6.5.2 Two-stage recalibration under label shift

Method. We propose a composite estimator of $h_{f,Q}^* = g^* \circ h_{f,P}^*$, which comprises two estimators $\hat{g} \approx g^*$ and $\hat{h}_P \approx h_{f,P}^*$. Here we describe a procedure to produce this composite estimator.

1. Use $\mathcal{D}_P$ to construct $\hat{h}_P : \mathcal{Z} \rightarrow \mathcal{Z}$, the estimated recalibration function (6.9) (for f under P).
2. Use $\mathcal{D}_P$ and $\mathcal{D}_Q$ to construct $\hat{g} : \mathcal{Z} \rightarrow \mathcal{Z}$ such that

$$\hat{g}(z) = \frac{\hat{w}_1 z}{\hat{w}_1 z + \hat{w}_0(1 - z)} \quad \text{where} \quad \hat{w}_k = \frac{\hat{Q}[Y = k]}{\hat{P}[Y = k]}, \quad \forall k \in \{0, 1\}, \quad (6.16)$$

where $\hat{P}[Y = k] := \frac{1}{|\mathcal{D}_P|} \sum_{i=1}^{|\mathcal{D}_P|} \mathbf{1}[y_i = k]$ and $\hat{Q}[Y = k] := \frac{1}{|\mathcal{D}_Q|} \sum_{i=1}^{|\mathcal{D}_Q|} \mathbf{1}[y'_i = k]$ are the empirical estimates of the class marginal probabilities.

3. Let

$$\hat{h}_Q = \hat{g} \circ \hat{h}_P. \quad (6.17)$$

Note that the recalibration estimator $\hat{h}_P$ (Step 1) remains the same with that in Section 6.4.1. Furthermore, the shift correction estimator $\hat{g}$ (Step 2) is a plug-in estimator of the label shift correction function g^* in (6.14) based on the estimated weights, $\hat{w}_1$ and $\hat{w}_0$.

Theory. We present a recalibration risk upper bound for the proposed two-stage estimator. We let $p_k := P[Y = k]$, $q_k := Q[Y = k]$, and $w_k^* = \frac{q_k}{p_k}$ for $k \in \{0, 1\}$. Moreover, we let $p_{\min} := \min_k p_k$, $q_{\min} := \min_k q_k$, $w_{\min}^* := \min_k w_k^*$ and $w_{\max}^* := \max_k w_k^*$.

Theorem 6.2 (Convergence of $\hat{h}_Q$). *Let P, Q be probability measures and let $f : \mathcal{X} \rightarrow \mathcal{Z}$ be a measurable function. Let $\mathcal{D}_P \sim P$ be an IID sample of size n_P and $\mathcal{D}_Q \sim Q$ be an IID sample of size n_Q . Suppose that Assumptions (B1) & (B2) hold. Let $\mathcal{B}$ be the UMB scheme induced by $\mathcal{D}_P$. Let $\hat{h}_P = \hat{h}_{P, \mathcal{B}}$ be the recalibration function (6.9) based on $\mathcal{B}$, and let $\hat{g}$ denote the shift correction function as defined in (6.16). Then*

$$R_Q(\hat{g} \circ \hat{h}_P) \leq 2 \left\{ \left(\frac{\rho_0 - \rho_1}{\rho_0 + \rho_1} \right)^2 + \frac{w_{\max}^*}{w_{\min}^*} \cdot R_P(\hat{h}_P; f) \right\} \quad \text{where} \quad \rho_k := \frac{\hat{w}_k}{w_k^*}, \quad k \in \{0, 1\}. \quad (6.18)$$

Furthermore, suppose that (A1), (A2) & (A3) hold. Then there exists a universal constant $c > 0$ such that for any $\delta \in (0, 1)$, if

$$n_P \geq \max \left\{ c, \frac{27}{p_{\min}} \right\} \cdot |\mathcal{B}| \log \left(\frac{4|\mathcal{B}|}{\delta} \right) \quad \text{and} \quad n_Q \geq \frac{27}{q_{\min}} \log \left(\frac{16}{\delta} \right),$$

then with probability at least $1 - \delta$,

$$R_Q(\hat{g} \circ \hat{h}_P) \leq 2 \frac{w_{\max}^*}{w_{\min}^*} \cdot \left\{ \left(\sqrt{\frac{1}{2(\lfloor n_P/|\mathcal{B} \rfloor - 1)}} \log \left(\frac{8|\mathcal{B}|}{\delta} \right) + \frac{1}{\lfloor n_P/|\mathcal{B} \rfloor} \right)^2 + \frac{8K^2}{|\mathcal{B}|^2} \right\} \\ + 54 \max \left\{ \frac{1}{p_{\min} \cdot n_P}, \frac{1}{q_{\min} \cdot n_Q} \right\} \cdot \log \left(\frac{16}{\delta} \right). \quad (6.19)$$

Remark 6.3. Note that $\rho_k \rightarrow 1$ as $n_P, n_Q \rightarrow \infty$, and thus, the upper bound (6.18) reduces to $2 \frac{w_{\max}^*}{w_{\min}^*} \cdot R_P(\hat{h}_P; f)$. Moreover, when $P = Q$, we have $w_{\min}^* = w_{\max}^* = 1$, and this further simplifies to the recalibration risk without label shift, up to multiplicative constant 2.

Remark 6.4 (Target sample complexity). Assume that $n_P \geq n_Q$ and the number of bins satisfies $|\mathcal{B}| \asymp n_P^{1/3}$. Then (6.19) implies $R_Q(\hat{g} \circ \hat{h}_P; f) = O(n_P^{-2/3} + n_Q^{-1})$. This result indicates that the proposed recalibration method using (6.17) requires a significantly smaller target sample size $n_Q = \Omega(\varepsilon^{-1})$ to control the risk, as compared to $n_Q = \Omega(\varepsilon^{-3/2})$ in (6.12).

Remark 6.5 (Comparison with some other estimators). *An alternative way to recalibrate under label shift is to perform UMB on $\mathcal{D}_Q$ directly and get the estimate $\hat{h}_Q^{\text{target}}$. With the number of bins on the order of $n_Q^{1/3}$, $\hat{h}_Q^{\text{target}}$ is a consistent estimator for h_Q^* with $R_Q(\hat{h}_Q^{\text{target}}) = O(n_Q^{-2/3})$. However, because $\hat{h}_Q^{\text{target}}$ does not take advantage of the source domain data, not only does it require more target sample ($n_Q = \Omega(\varepsilon^{-3/2})$) compared to $n_Q = \Omega(\varepsilon^{-1})$ for $\hat{h}_Q$, but it also requires access to the features in the target sample $\mathcal{D}_Q$. We also note that both $\hat{g}$ and $\hat{h}_P$ are important for the propose composite estimator $\hat{h}_Q$: neither $\hat{g}$ or $\hat{h}_P$ is a consistent estimator for h_Q^* , except in special cases (h_Q^* coincides with g^* or h_P^* , respectively).*

Remark 6.6 (Comparison with label shift using unlabeled target data). *When the source sample size n_P is sufficiently large, we achieve a risk of $R_Q(\hat{g} \circ \hat{h}_P) = O(n_Q^{-1})$ with high probability. It is important to note that in this scenario, we only utilize the labels from the target sample to address label shift. Remarkably, the same rate applies when employing the algorithms proposed in [90], which only uses features from the target sample. Please refer to Appendix 6.11 for the proof.*

6.6 Numerical experiments

In this section, we present the results of our numerical simulations conducted to validate and reinforce the theoretical findings discussed earlier. The simulations are based on a family of recalibration functions called beta calibration [160]: $\mathcal{H}_{\text{beta}} = \{h_{\text{beta}}(z; a, b, c) : a \geq 0, b \geq 0, c \in \mathbb{R}\}$, where $h_{\text{beta}}(\cdot; a, b, c) : [0, 1] \rightarrow [0, 1]$ is defined as

$$h_{\text{beta}}(z; a, b, c) = \frac{1}{1 + 1/\left(e^c \frac{z^a}{(1-z)^b}\right)}. \quad (6.20)$$

In addition, consider the joint distributions $\mathcal{D}(\pi)$ of X and Y , where $Y \sim \text{Bernoulli}(\pi)$, $X | Y = 0 \sim N(-2, 1)$, and $X | Y = 1 \sim N(2, 1)$, and a pre-trained probabilistic classifier $f(x) = \sigma(x) := 1/(1 + e^{-x})$. To accommodate the limitations of space, we summarize the results in Figure 6.1, 6.2, 6.3, and Table 6.1, 6.2, providing a concise overview. Detailed information about the simulation settings, implementation details, and further experimental results and discussions can be found in Appendix 6.12. Our simulation code is available at https://github.com/ZeyuSun/calibration_label_shift.

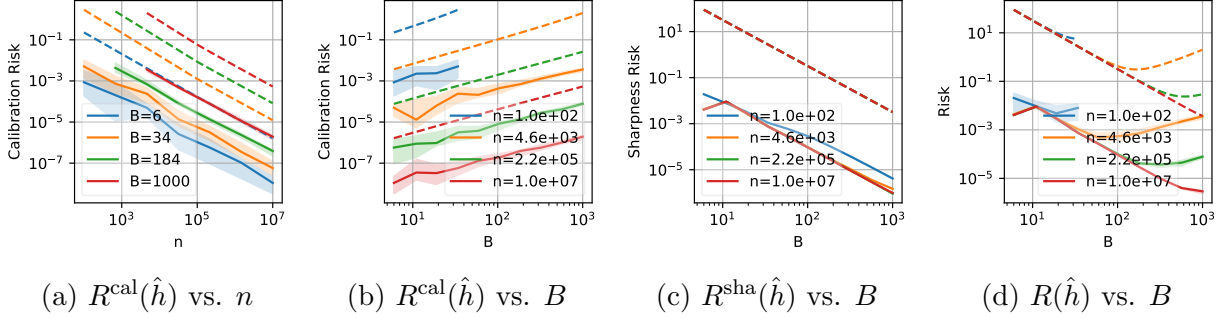


Figure 6.1: Medians (solid lines) and 10-90 percentile ranges (shaded areas) of quadrature estimates of population risks over 10 realizations and theoretical risk upper bounds ($\delta = 0.1$) (dashed lines) for various n and B . **(a)-(c)** The empirical rates, $R^{\text{cal}} = O(n^{-0.99}B^{0.98})$ and $R^{\text{sha}} = O(B^{-1.83})$, align with theoretically predicted rates, $\tilde{O}(B/n)$ and $O(B^{-2})$, in Thm. 6.1. **(d)** The empirically observed R and our upper bound exhibit similar trends as a function of B .

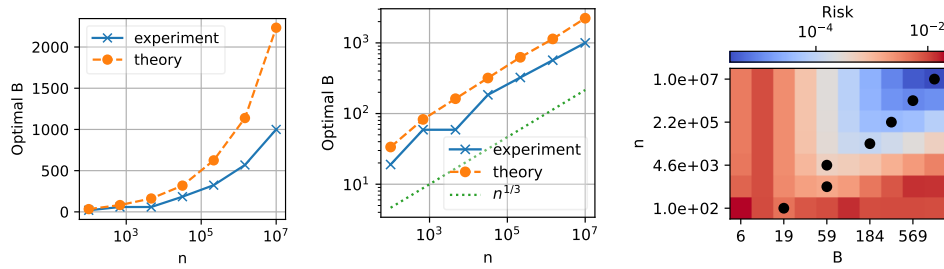
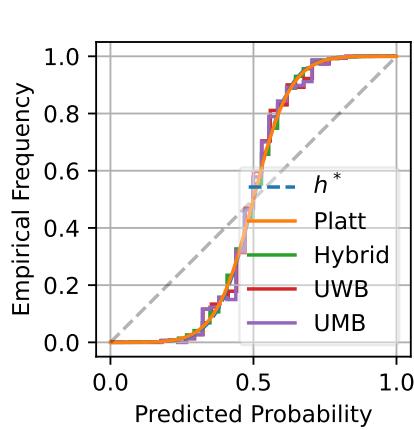


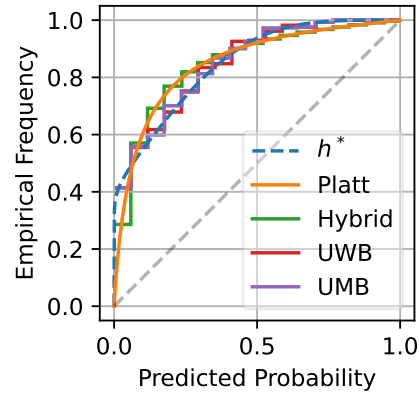
Figure 6.2: The optimal number of bins B for different sample sizes n plotted in linear scale (*left*) and log scale (*middle*), and the population risk for various combinations of n and B , with the optimal B marked by black dots (*right*). Note that the risk surface is relatively smooth around its minimum, suggesting the robustness of the optimal B .

Table 6.1: 90%-quantiles of the risks of Platt scaling [229], the hybrid method [162], uniform-width binning (UWB) [109], and uniform-mass binning (UMB) over 100 random calibration datasets drawn from $Z \sim \text{Uniform}[0, 1]$, and **(a)** $Y | Z \sim \text{Bernoulli}(h_{\text{beta}}(z; 4, 4, 0))$, or **(b)** $Y | Z \sim \text{Bernoulli}(h_{\text{beta}}(z; 0.1, 4, 0))$. While Platt scaling and the hybrid method achieve lower R under the correct parametric assumption, UWB and UMB may outperform when the parametric assumption fails.

(a) Correct parametric assumption					(b) Misspecified parametric assumption				
Metric ($\times 10^{-3}$)	R^{cal}	R^{sha}	R	MSE	Metric ($\times 10^{-3}$)	R^{cal}	R^{sha}	R	MSE
Platt	0.122	0.000	0.122	10.338	Platt	3.682	0.000	3.682	21.994
Hybrid	0.119	0.212	0.315	10.532	Hybrid	3.117	0.251	3.360	21.672
UWB	0.661	0.194	0.855	11.071	UWB	0.572	0.238	0.810	19.122
UMB	0.647	0.212	0.839	11.055	UMB	0.560	0.251	0.797	19.109



(a) Correct parametric assumption



(b) Misspecified parametric assumption

Figure 6.3: Optimal recalibration function h^* and recalibration function estimates by Platt Scaling [229], the hybrid method [162], UWB, and UMB when the parametric assumption is **(a)** correct and **(b)** misspecified. UMB traces the h^* in both cases, whereas the hybrid method traces Platt scaling, exhibiting an intrinsic bias from h^* in **(b)**.

Table 6.2: Risks under label shift from $\mathcal{D}(0.5)$ to $\mathcal{D}(0.1)$, with $n_P = 10^3$ and $n_Q = 10^2$. Standard deviations are computed from 10 random realizations. LABEL-SHIFT, only applying an injective $\hat{g}$, achieves $R^{\text{sha}} = 0$ but incurs high R^{cal} . SOURCE, recalibrated with $B = n_P^{1/3}$ on $\mathcal{D}_P$, incurs high R^{cal} . TARGET, recalibrated with $B = n_Q^{1/3}$ on $\mathcal{D}_Q$, incurs low R^{cal} . Our proposed COMPOSITE achieves the lowest nonzero R^{sha} and lowest R^{cal} .

Method	Estimator	R^{cal}	R^{sha}	R	MSE
SOURCE	$\hat{h}_P$	0.016±0.005	0.0032±0.0014	0.019±0.006	0.029±0.006
TARGET	$\hat{h}_Q^{\text{target}}$	0.0020±0.0025	0.049±0.006	0.051±0.006	0.060±0.006
LABEL-SHIFT	$\hat{g}$	0.026±0.006	0	0.026±0.006	0.035±0.006
COMPOSITE	$\hat{h}_Q$	0.00019±0.00017	0.0032±0.0014	0.0034±0.0013	0.0127±0.0013

6.7 Discussion

This chapter presents a comprehensive theory for recalibration, considering both calibration and sharpness within the mean-squared-error (MSE) decomposition framework. We use this framework to quantify the optimal calibration-sharpness balance and establish a rigorous upper bound on the finite-sample risk for uniform-mass binning (UMB). Additionally, we address the challenge of recalibration under label shift with limited access to labeled target data. Our proposed two-stage approach effectively estimates the recalibration function using ample data from the source domain and adjusts for the label shift using target domain data. Importantly, our findings suggest that transferring a calibrated classifier requires a significantly smaller target sample than recalibrating from scratch on the new domain. Numerical simulations confirm the tightness of the finite sample bounds, validate the optimal number of bins, and demonstrate the effectiveness of the label shift adaptation.

In concluding this chapter, we identify several promising directions for future research.

Relaxation of the assumptions It would be worthwhile to explore whether or not the assumptions made in our analysis could be relaxed. For instance, the widely adopted monotonicity assumption (A2) and its variants [338, 342, 236, 245] may not hold in real-world settings. Thus, exploring potential relaxations of this assumption is valuable. In addition, a mild but non-trivial smoothness assumption (A3) (c.f. Remark 6.8 and 6.9) is introduced to obtain a tight sharpness risk upper bound (c.f. Remark 6.10); investigating its practical implications and exploring potential relaxations could be interesting future work.

Calibration-sharpness framework analysis Applying our framework to analyze recalibration methods beyond UMB, such as isotonic regression [338] and kernel density estimation [342], can offer further insights into their performance and properties, providing guidance to practitioners in selecting suitable algorithms based on specific conditions and requirements.

Multiclass probability recalibration The concept of calibration considered in this work extends to multiclass classification settings, known as canonical calibration [293]. Weaker, but more tractable notions of multiclass calibration have also been explored in the literature [293, 111]. While partition-based methods, as multiclass extensions of binning methods, are known to have consistency [293] and vanishing calibration error [230], establishing upper bounds for their sharpness risk remains difficult. Additionally, designing a partition scheme in a multidimensional space is a challenging task [111]; the interplay between calibration and sharpness could potentially guide the development of partition strategies that balance both aspects in multiclass classification.

Applications to real-world data Investigating the calibration-sharpness tradeoff in real-world applications, which often involve multiple classes, presents an interesting challenge. It is crucial to develop effective estimators for both calibration risk and sharpness risk in such scenarios. While estimators for calibration risk exist (e.g., binning-based and KDE-based) [38, 162, 342, 231] and a lower bound for sharpness risk has been established [224], a direct estimator of sharpness risk is still lacking.

6.8 Appendix: Proof of Proposition 6.1

Proof of Proposition 6.1. First of all, we recall the definition of the two risks from (6.3) and (6.4):

$$\begin{aligned} R^{\text{cal}}(h; f) &= \mathbb{E} \left[\left(f(X) - \mathbb{E}[Y \mid f(X)] \right)^2 \right] \\ R^{\text{sha}}(h; f) &= \mathbb{E} \left[\left(\mathbb{E}[Y \mid h \circ f(X)] - \mathbb{E}[Y \mid f(X)] \right)^2 \right]. \end{aligned}$$

To keep our notation concise, we use $Z = f(X)$ as a shorthand notation, and also let $Y_Z := \mathbb{E}[Y \mid Z]$ and $Y_{h(Z)} = \mathbb{E}[Y \mid h(Z)]$ throughout this proof. We can decompose the recalibration risk from Definition 6.5:

$$\begin{aligned} R(h) &= \mathbb{E}[(h(Z) - Y_Z)^2] \\ &= \mathbb{E}[(h(Z) - Y_{h(Z)} + Y_{h(Z)} - Y_Z)^2] \\ &= \mathbb{E}[(h(Z) - Y_{h(Z)})^2] + \mathbb{E}[(Y_{h(Z)} - Y_Z)^2] + 2\mathbb{E}[(h(Z) - Y_{h(Z)})(Y_{h(Z)} - Y_Z)] \\ &= \mathbb{E}[(h(Z) - Y_{h(Z)})^2] + \mathbb{E}[(Y_{h(Z)} - Y_Z)^2] + 2\mathbb{E}[\mathbb{E}[(h(Z) - Y_{h(Z)})(Y_{h(Z)} - Y_Z) \mid h(Z)]] \\ &= \mathbb{E}[(h(Z) - Y_{h(Z)})^2] + \mathbb{E}[(Y_{h(Z)} - Y_Z)^2] + 2\mathbb{E}[(h(Z) - Y_{h(Z)})(Y_{h(Z)} - \mathbb{E}[Y_Z \mid h(Z)])] \\ &= \mathbb{E}[(h(Z) - Y_{h(Z)})^2] + \mathbb{E}[(Y_{h(Z)} - Y_Z)^2] + 2\mathbb{E}[(h(Z) - Y_{h(Z)})(Y_{h(Z)} - Y_{h(Z)})] \\ &= \underbrace{\mathbb{E}[(h(Z) - Y_{h(Z)})^2]}_{R^{\text{cal}}(h)} + \underbrace{\mathbb{E}[(Y_{h(Z)} - Y_Z)^2]}_{R^{\text{sha}}(h)}. \end{aligned}$$

□

6.9 Appendix: Proof of Theorem 6.1

In this section, we present a proof of Theorem 6.1. Let $(Y, Z) \in \mathcal{Y} \times \mathcal{Z}$ be random variables that admits a joint distribution $P_{Y,Z}$, which we assume to be fixed throughout this section. Let $S = \{(y_i, z_i) \in \mathcal{Y} \times \mathcal{Z} : i \in [n]\}$ and let $\mathcal{B} = \{I_1, I_2, \dots, I_B\}$ be the uniform-mass binning scheme (cf. Definition 6.6) of size B induced by $(z_i$'s in S). Note that if S is a random sample from $P_{Y,Z}$, then the binning scheme $\mathcal{B}$ induced by S is also a random variable following a derived distribution. To facilitate our analysis, we introduce the notion of well-balanced binning.

Definition 6.8 (Well-balanced binning; [162]). *Let $B \in \mathbb{N}$, let Z be a random variable that takes value in $[0, 1]$, and let $\alpha \in \mathbb{R}$ such that $\alpha \geq 1$. A binning scheme $\mathcal{B}$ of size B is*

α -well-balanced with respect to Z if

$$\frac{1}{\alpha B} \leq P[Z \in I_b] \leq \frac{\alpha}{B}, \quad \forall b \in [B].$$

In addition, we define two (parameterized families of) Boolean-valued functions Φ_{balance} and Φ_{approx} as follows: for any binning scheme $\mathcal{B}$,

$$\forall \alpha \in \mathbb{R}, \quad \Phi_{\text{balance}}(\mathcal{B}; \alpha) := \mathbb{1} \left\{ \frac{1}{\alpha |\mathcal{B}|} \leq P[Z \in I] \leq \frac{\alpha}{|\mathcal{B}|}, \quad \forall I \in \mathcal{B} \right\}, \quad (6.21)$$

$$\forall \varepsilon \in \mathbb{R}, \quad \Phi_{\text{approx}}(\mathcal{B}; \varepsilon) := \mathbb{1} \left\{ \max_{I \in \mathcal{B}} |\hat{\mu}_I - \mu_I| \leq \varepsilon \right\}, \quad (6.22)$$

where $\mathbb{1}(A) = 1$ if and only if the predicate A is true, and for each interval $I \in \mathcal{B}$,

$$\hat{\mu}_I = \frac{\sum_{i=1}^n y_i \cdot \mathbb{1}_I(z_i)}{\sum_{i=1}^n \mathbb{1}_I(z_i)} \quad \text{and} \quad \mu_I = \mathbb{E}_{(Y,Z) \sim P_{Y,Z}} [Y \cdot \mathbb{1}_I(Z)]. \quad (6.23)$$

Note that if $\Phi_{\text{balance}}(\mathcal{B}; \alpha) = 1$ for $\alpha \geq 1$, then $\mathcal{B}$ is α -well-balanced with respect to Z (cf. Definition 6.8). Also, if $\Phi_{\text{approx}}(\mathcal{B}; \varepsilon) = 1$ for $\varepsilon \geq 0$, then the conditional empirical mean of Y in each bin $I \in \mathcal{B}$ approximates the conditional expectation with error at most ε , uniformly for all bins.

The rest of this section is organized as follows. In Section 6.9.1, we ensure that for an appropriate choice of $\alpha, \varepsilon \in \mathbb{R}$, it holds with high probability (with respect to the randomness in $\mathcal{B}$) that $\Phi_{\text{balance}}(\mathcal{B}; \alpha) = \Phi_{\text{approx}}(\mathcal{B}; \varepsilon) = 1$. In Section 6.9.2, we establish upper bounds on the reliability risk R^{cal} and the sharpness risk R^{sha} under the premise that $\Phi_{\text{balance}}(\mathcal{B}; \alpha) = \Phi_{\text{approx}}(\mathcal{B}; \varepsilon) = 1$. Finally, in Section 6.9.3, we conclude the proof of Theorem 6.1 by combining these results together.

6.9.1 High-probability certification of the conditions

Well-balanced binning scheme. First of all, we observe that the uniform-bass binning scheme $\mathcal{B}$ induced by an IID random sample from $P_{Y,Z}$ is 2-well-balanced with high probability, if the sample size is sufficiently large. Here we paraphrase a result from [162] in our language.

Lemma 6.3 ([162, Lemma 4.3]). *Let $S = \{Z_i : i \in [n]\}$ be an IID sample drawn from P_Z and let $\mathcal{B}$ be the uniform-mass binning scheme of size B induced by S . There exists a universal constant $c' > 0$ such that for any $\delta \in (0, 1)$, if $n \geq c' \cdot B \log(B/\delta)$, then $\Phi_{\text{balance}}(\mathcal{B}, 2) = 1$ with probability at least $1 - \delta$.*

Lemma 6.3 states that

$$n \geq c' \cdot B \log \left(\frac{B}{\delta} \right) \quad \implies \quad P[\mathcal{B} \text{ is 2-well-balanced with respect to } P_Z] \geq 1 - \delta.$$

While the value of the universal constant c was not specified in the original reference [162], we remark that one may set, for example, $c' = 2420$, which can be verified by following their proof with c' kept explicit.

The proof of Lemma 6.3 in [162] relies on a discretization argument that considers a fine-grained cover of $\mathcal{Z} = [0, 1]$ consisting of disjoint intervals—namely, $\{I'_j : j \in [10B]\}$ such that $P[Z \in I'_j] = \frac{1}{10B}$ for all $j \in [10B]$ —and then approximates each I_b by a subset of the cover. As the authors of [162] remarked, this argument provides a tighter sample complexity upper bound than naïvely applying Chernoff bounds or a standard VC dimension argument, which would yield an upper bound of order $O\left(B^2 \log\left(\frac{B}{\delta}\right)\right)$. We omit the proof of Lemma 6.3 and refer interested readers to the referenced paper [162] for more details.

Uniform concentration of bin-wise means. Next, we argue that for the uniform-mass binning scheme $\mathcal{B}$ induced by an IID sample, the conditional empirical means of each bin concentrates to the population conditional expectation, uniformly for all bins in $\mathcal{B}$. Here we restate a result from [110].

Lemma 6.4 ([110, Corollary 1]). *Let P_Z be an absolutely continuous probability measure on $\mathcal{Z} = [0, 1]$, and $S = \{Z_i : i \in [n]\}$ be an IID sample drawn from P_Z . Let $B \in \mathbb{N}$ such that $B \leq \frac{n}{2}$ and $\mathcal{B}$ be the uniform-mass binning scheme of size B induced by S . Then for any $\delta \in (0, 1)$,*

$$P[\Phi_{\text{approx}}(\mathcal{B}; \varepsilon_\delta) = 1] \geq 1 - \delta \quad \text{where} \quad \varepsilon_\delta = \sqrt{\frac{1}{2(\lfloor n/B \rfloor - 1)} \log \left(\frac{2B}{\delta} \right)} + \frac{1}{\lfloor n/B \rfloor}. \quad (6.24)$$

Lemma 6.4 states that under the mild regularity condition of P_Z being absolutely continuous, the uniform-mass binning accurately approximates all bin-wise conditional means as long as there are at least two samples per bin in the sense that

$$n \geq 2B \quad \implies \quad P \left[\sup_{b \in [B]} |\hat{\mu}_b - \mu_b| \leq \sqrt{\frac{1}{2(\lfloor n/B \rfloor - 1)} \log \left(\frac{2B}{\delta} \right)} + \frac{1}{\lfloor n/B \rfloor} \right] \geq 1 - \delta.$$

6.9.2 Conditional upper bounds on reliability risk and sharpness risk

In this section, we establish upper bounds on the reliability risk R^{cal} and the sharpness risk R^{sha} for $\hat{h}$ under the premise that $\Phi_{\text{balance}}(\mathcal{B}; \alpha) = 1$ and $\Phi_{\text{approx}}(\mathcal{B}; \varepsilon) = 1$ for appropriate parameters $\alpha, \varepsilon \in \mathbb{R}$.

Preparation. To avoid clutter in the lemma statements to follow, here we recall our problem setting and set several notation that will be used throughout this section. Recall that $P = P_{X,Y}$ is a joint distribution on $\mathcal{X} \times \mathcal{Y}$ and let $f : \mathcal{X} \rightarrow \mathcal{Z}$ is a measurable function. In addition, we let $\tilde{S} = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} : i \in [n]\}$ be an IID sample drawn from P , and let $S = \{(z, y) \in \mathcal{Z} \times \mathcal{Y} : (x, y) \in \tilde{S} \text{ and } z = f(x)\}$. Let $\mathcal{B}$ be the uniform-mass binning scheme induced by $(z$'s in S , and let $\hat{h} = \hat{\mathcal{B}} : \mathcal{Z} \rightarrow \mathcal{Z}$ be the recalibration function derived from $\mathcal{B}$ as we described in Section 6.4.1; see (6.9). The dependence among $P, f, \tilde{S}, S, \mathcal{B}$, and $\hat{h}$ are summarized by a diagram in Figure 6.4.

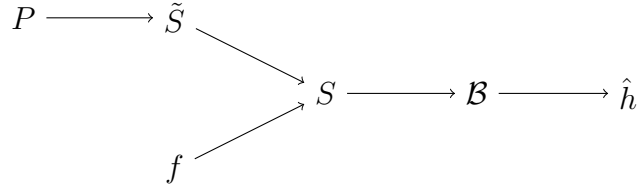


Figure 6.4: Stochastic dependence among $P, f, \tilde{S}, S, \mathcal{B}$, and $\hat{h}$.

Furthermore, we define the index function for a binning scheme to facilitate our analysis.

Definition 6.9. Let $\mathcal{B}$ be a binning scheme. The index function for $\mathcal{B}$ is the function $\beta : \mathcal{Z} \rightarrow [|\mathcal{B}|]$ such that

$$\beta(z) = \sum_{I \in \mathcal{B}} \mathbf{1}_{(0, \sup I]}(z). \quad (6.25)$$

Remark 6.7. Note that β is a measurable function and defines an index function that identifies which bin of $\mathcal{B}$ the argument $z \in [0, 1]$ belongs to. Specifically, suppose that $\mathcal{B} = \{I_1, \dots, I_B\}$ for some $B \in \mathbb{N}$ and there exists $u_0, u_1, \dots, u_B \in [0, 1]$ such that (i) $0 = u_0 < u_1 < \dots < u_B = 1$ and (ii) $I_b = (u_{b-1}, u_b]$ for all $b \in [B] \setminus \{1\}$ and $I_1 = [u_0, u_1]$. Then $\beta(z) = b$ if and only if $z \in I_b$.

6.9.2.1 Calibration risk upper bound

We observe that if a binning scheme $\mathcal{B}$ produces empirical means $\hat{\mu}_I$ that approximate the true means μ_I with error at most ε , then the calibration risk is upper bounded by ε^2 .

Lemma 6.5 (Calibration risk bound). *For any $\varepsilon \geq 0$, if $\Phi_{\text{approx}}(\mathcal{B}; \varepsilon) = 1$, then*

$$R^{\text{cal}}(\hat{h}; f, P) \leq \varepsilon^2.$$

Proof of Lemma 6.5. To begin with, we recall the definition of the calibration risk (Definition 6.3), and let $Z = f(X)$. Then we may write

$$\begin{aligned} R^{\text{cal}}(\hat{h}; f, P) &= \mathbb{E} \left[\left(\hat{h}(Z) - \mathbb{E}[Y \mid \hat{h}(Z)] \right)^2 \right] \\ &= \mathbb{E} \left[\mathbb{E} \left[\left(\hat{h}(Z) - \mathbb{E}[Y \mid \hat{h}(Z)] \right)^2 \mid \beta(Z) \right] \right] \quad \because \text{the law of total expectation} \\ &= \mathbb{E} \left[\left(\hat{\mu}_{I_{\beta(Z)}} - \mu_{I_{\beta(Z)}} \right)^2 \right] \quad \text{cf. (6.23)} \\ &\leq \max_{I \in \mathcal{B}} \left(\hat{\mu}_I - \mu_I \right)^2. \end{aligned}$$

Note that if $\Phi_{\text{approx}}(\mathcal{B}; \varepsilon) = 1$, then $\max_{I \in \mathcal{B}} \left(\hat{\mu}_I - \mu_I \right)^2 \leq \varepsilon^2$. $\square$

We remark that the proof of Lemma 6.5 is a simple application of applying Hölder's inequality. Also, we note that a similar argument was considered in [110, Proposition 1] to establish the inequalities between the L^p -counterparts of the calibration risk, which they call the ℓ_p -expected calibration error (ECE). In this work, we focus on the case $p = 2$.

6.9.2.2 Sharpness risk upper bound

Next, we present an upper bound for the sharpness risk that diminishes as the binning scheme $\mathcal{B}$ becomes more balanced.

Lemma 6.6 (Sharpness risk bound). *Suppose that the optimal post-hoc recalibration function $h_{f,P}^*$, cf. (6.7), is monotonically non-decreasing. Let $\alpha \in \mathbb{R}$ such that $\alpha \geq 1$. If $\Phi_{\text{balance}}(\mathcal{B}, \alpha) = 1$, then*

$$R^{\text{sha}}(\hat{h}; f, P) \leq \frac{\alpha}{|\mathcal{B}|}.$$

Proof of Lemma 6.6. Letting $Z = f(X)$, we can write the sharpness risk of $\hat{h}$ over f with respect to P as

$$R^{\text{sha}}(\hat{h}; f, P) := \mathbb{E} \left[\left(\mathbb{E}[Y \mid \hat{h}(Z)] - \mathbb{E}[Y \mid Z] \right)^2 \right].$$

We recall the definition of the index function β for $\mathcal{B}$ (Definition 6.9) and observe that

$$\begin{aligned}
& \mathbb{E} \left[\left(\mathbb{E} \left[Y \mid \hat{h}(Z) \right] - \mathbb{E} \left[Y \mid Z \right] \right)^2 \right] \\
& \leq \mathbb{E} \left[\left| \mathbb{E} \left[Y \mid \hat{h}(Z) \right] - \mathbb{E} \left[Y \mid Z \right] \right| \right] \quad \because \left| \mathbb{E} \left[Y \mid \hat{h}(Z) \right] - \mathbb{E} \left[Y \mid Z \right] \right| \leq 1 \\
& = \sum_{I \in \mathcal{B}} \mathbb{E} \left[\left| \mathbb{E} \left[Y \mid \hat{h}(Z) \right] - \mathbb{E} \left[Y \mid Z \right] \right| \cdot \mathbf{1}_I(Z) \right] \\
& = \sum_{I \in \mathcal{B}} \mathbb{E} \left[\mathbb{E} \left[\left| \mathbb{E} \left[Y \mid \hat{h}(Z) \right] - \mathbb{E} \left[Y \mid Z \right] \right| \cdot \mathbf{1}_I(Z) \mid \beta(Z) \right] \right] \quad \because \text{the law of total expectation} \\
& = \sum_{I \in \mathcal{B}} P[Z \in I] \mathbb{E} \left[\mathbb{E} \left[\left| \mathbb{E} \left[Y \mid \hat{h}(Z) \right] - \mathbb{E} \left[Y \mid Z \right] \right| \mid Z \in I \right] \right] \quad \because \text{Remark 6.7} \\
& \leq \sum_{I \in \mathcal{B}} P[Z \in I] \left(\sup_{z \in I} h_{f,P}^*(z) - \inf_{z \in I} h_{f,P}^*(z) \right) \quad \because \text{by definition of } h_{f,P}^*; \text{ cf. (6.7)} \\
& \leq \sum_{I \in \mathcal{B}} \frac{\alpha}{|\mathcal{B}|} \left(\sup_{z \in I} h_{f,P}^*(z) - \inf_{z \in I} h_{f,P}^*(z) \right) \quad \because \Phi_{\text{balance}}(\mathcal{B}, \alpha) = 1 \\
& \leq \frac{\alpha}{|\mathcal{B}|}.
\end{aligned}$$

The inequality in the last line follows from the facts that (i) $I \in \mathcal{B}$ are mutually exclusive and (ii) $h_{f,P}^*(z) \in [0, 1]$ and $h_{f,P}^*$ is monotone non-decreasing. $\square$

Our proof of Lemma 6.6 relies on similar techniques that are used in [162, Lemmas D.5 and D.6]. However, we note that we obtain an improved constant — 1 as opposed to 2 in [162, Lemma D.6] — with a more refined analysis.

An improved rate with additional assumptions. It is possible to improve the rate of the sharpness risk upper bound from $O(|\mathcal{B}|^{-1})$ to $O(|\mathcal{B}|^{-2})$ with an additional regularity assumption on $h_{f,P}^*$.

Recall that we assumed in (A3) that there exists $K > 0$ such that if $z_1 \leq z_2$, then $h_{f,P}^*(z_2) - h_{f,P}^*(z_1) \leq K \cdot (F_Z(z_2) - F_Z(z_1))$, that is, $h_{f,P}^*$ is K -smooth with respect to F_Z . This posits that the conditional probability $P[Y = 1 | Z = z]$ of the target variable Y given a forecast variable Z cannot vary too much in regions where the density of Z is low, or where the forecast is rarely issued. This is a reasonable assumption because if $P[Y = 1 | Z]$ changes too rapidly with respect to Z , then it suggests that we need additional information about Y beyond what Z can provide in order to improve the quality of forecasts. We remark that (A3) is indeed a fairly mild assumption to impose on, however, is not a trivial one.

Remark 6.8 (Mildness of (A3)). *Suppose that $Z = f(X)$ has a density p_Z that is uniformly lower bounded by ϵ on the support of Z . If $h_{f,P}^*$ is L -Lipschitz, then $h_{f,P}^*$ is (L/ϵ) -smooth*

with respect to F_Z . This also provides a sufficient condition to verify (A3) in practice.

Remark 6.9 (Non-triviality of (A3)). Notice that even if F_Z is absolutely continuous and $h_{f,P}^*$ is continuous, the smoothness constant K could become large if the prediction Z is heavily miscalibrated. For instance, in Figure 6.6, $h_{f,P}^*(z)$ is changing fast in the interval $[0.5, 0.75]$ where $p_Z(z)$ is small, which results in a larger value of K that can even diverge if $p_Z(z) \rightarrow 0$.

Here we define the notion of ψ -smoothness to formalize Assumption (A3), and then present an improved upper bound for the sharpness risk.

Definition 6.10 (ψ -smoothness). Let $K \in \mathbb{R}_+$ and $\psi : [0, 1] \rightarrow [0, 1]$ be a monotone non-decreasing function. A function $\phi : [0, 1] \rightarrow [0, 1]$ is K -smooth with respect to ψ if for any $z_1, z_2 \in [0, 1]$ such that $z_1 \leq z_2$,

$$|\phi(z_2) - \phi(z_1)| \leq K \cdot (\psi(z_2) - \psi(z_1)). \quad (6.26)$$

Lemma 6.7 (Improved sharpness risk bound). Suppose that the function $h_{f,P}^*(z)$ defined in (6.7) is monotonically non-decreasing and K -smooth with respect to F_Z for some $K \geq 0$, where F_Z is the cumulative distribution function of $Z = f(X)$. If $\Phi_{\text{balance}}(\mathcal{B}, \alpha) = 1$, then

$$R^{\text{sha}} \leq \frac{K^2 \alpha^3}{B^2}.$$

Proof of Lemma 6.7. Let $Z = f(X)$ and $B = |\mathcal{B}|$. For each $b \in [B]$, we let $z_{b,\max} := \sup I_b$ and $z_{b,\min} := \inf I_b$. Then we have

$$\begin{aligned} & R^{\text{sha}}(\hat{h}; f, P) \\ &= \mathbb{E} \left[\left(\mathbb{E}[Y \mid \hat{h}(Z)] - \mathbb{E}[Y \mid Z] \right)^2 \right] \\ &= \mathbb{E} \left[\mathbb{E} \left[\left(\mathbb{E}[Y \mid \hat{h}(Z)] - \mathbb{E}[Y \mid Z] \right)^2 \mid \beta(Z) \right] \right] \\ &= \sum_{b=1}^B P[Z \in I_b] \cdot \mathbb{E} \left[\left(\mathbb{E}[Y \mid \hat{h}(Z)] - \mathbb{E}[Y \mid Z] \right)^2 \mid \beta(Z) = b \right] \\ &\leq \sum_{b=1}^B P[Z \in I_b] \cdot \left(h_{f,P}^*(z_{b,\max}) - h_{f,P}^*(z_{b,\min}) \right)^2 && \because h_{f,P}^* \text{ is non-decreasing} \\ &\leq \sum_{b=1}^B P[Z \in I_b] \cdot \left(K \cdot (F_Z(z_{b,\max}) - F_Z(z_{b,\min})) \right)^2 && \because h_{f,P}^* \text{ is } K\text{-smooth w.r.t. } F_Z \\ &= \sum_{b=1}^B K^2 \cdot P[Z \in I_b]^3 \end{aligned}$$

$$\begin{aligned}
&\leq K^2 \sum_{b=1}^B \left(\frac{\alpha}{B}\right)^3 && \because \Phi_{\text{balance}}(\mathcal{B}, \alpha) = 1 \\
&= \frac{K^2 \alpha^3}{B^2}.
\end{aligned}$$

□

Remark 6.10 (Tightness of the rate $O(B^{-2})$). *The asymptotic rate $R^{\text{sha}} = O(B^{-2})$ is tight and cannot be further improved without additional assumptions. For instance, let's consider a uniform-mass binning of size B on $Z \sim \text{Uniform}[0, 1]$. In the population limit, each bin has width $1/B$ and within-bin variance $1/(12B^2)$. Thus, the sharpness risk, obtained by taking expectation of the conditional variance (per each bin), is $1/(12B^2)$, attaining the rate B^{-2} .*

6.9.3 Completing the proof of Theorem 6.1

Proof of Theorem 6.1. For given $\delta \in (0, 1)$, let $\delta_1 = \delta_2 = \delta/2$. Then we observe that

$$\begin{aligned}
n \geq c' \cdot |\mathcal{B}| \log \left(\frac{|\mathcal{B}|}{\delta_1} \right) &\implies P[\Phi_{\text{balance}}(\mathcal{B}, 2) = 1] \geq 1 - \delta_1 && \text{by Lemma 6.3} \\
n \geq 2|\mathcal{B}| &\implies P[\Phi_{\text{approx}}(\mathcal{B}, \varepsilon_{\delta_2}) = 1] \geq 1 - \delta_2 && \text{by Lemma 6.4}
\end{aligned}$$

where $c' > 0$ is the universal constant that appears in Lemma 6.3 and

$$\varepsilon_{\delta_2} = \sqrt{\frac{1}{2(\lfloor n/|\mathcal{B}| \rfloor - 1)} \log \left(\frac{2|\mathcal{B}|}{\delta_2} \right)} + \frac{1}{\lfloor n/|\mathcal{B}| \rfloor}.$$

Observe that $\delta_1 = \frac{\delta}{2} < \frac{1}{2}$ and $|\mathcal{B}| \geq 1$, and thus, $\log \left(\frac{|\mathcal{B}|}{\delta_1} \right) \geq \log 2$. Letting $c := \max\{c', \frac{2}{\log 2}\}$ and applying the union bound, we have

$$n \geq c \cdot |\mathcal{B}| \log \left(\frac{2|\mathcal{B}|}{\delta} \right) \implies P[\Phi_{\text{balance}}(\mathcal{B}, 2) = 1 \text{ and } \Phi_{\text{approx}}(\mathcal{B}, \varepsilon_{\delta/2}) = 1] \geq 1 - \delta.$$

Next, we observe that if $\Phi_{\text{balance}}(\mathcal{B}, 2) = 1$ and $\Phi_{\text{approx}}(\mathcal{B}, \varepsilon_{\delta_2}) = 1$, then

$$\begin{aligned}
R^{\text{cal}}(\hat{h}; f; P) &\leq (\varepsilon_{\delta_2})^2 && \text{by Lemma 6.5,} \\
R^{\text{sha}}(\hat{h}; f, P) &\leq \frac{2}{|\mathcal{B}|}, && \text{by Lemma 6.6.}
\end{aligned}$$

Additionally, if the Assumption (A3) also holds, then we obtain a stronger upper bound on

$R^{\text{sha}}(\hat{h}; f, P)$ by Lemma 6.7:

$$R^{\text{sha}}(\hat{h}; f, P) \leq \frac{8K^2}{|\mathcal{B}|^2}.$$

□

6.10 Appendix: Proof of Theorem 6.2

This section contains a proof of Theorem 6.2. Prior to the proof, in Section 6.10.1, we provide several lemmas that will be useful in our proof. Thereafter, we present a proof of Theorem 6.2 in its entirety in Section 6.10.2.

6.10.1 Useful lemmas

6.10.1.1 Concentration of $\hat{w}_k$ to w_k^*

First of all, we recall the binomial Chernoff bound, which is a classical result about the concentration of measures that can be found in standard textbooks on probability theory.

Lemma 6.8 (Binomial Chernoff bound). *Let X_i be IID Bernoulli random variables with parameters $p \in (0, 1)$, and let $S_n := \frac{1}{n} \sum_{i=1}^n X_i$. Then for any $\delta \in \mathbb{R}$ such that $0 < \varepsilon < 1$,*

$$\begin{aligned} P[S_n \geq (1 + \varepsilon)p] &\leq \exp\left(-\frac{\varepsilon^2 p}{3} n\right), \\ P[S_n \leq (1 - \varepsilon)p] &\leq \exp\left(-\frac{\varepsilon^2 p}{2} n\right). \end{aligned}$$

It follows from Lemma 6.8 that for any $\varepsilon, \delta \in (0, 1)$,

$$n \geq \frac{3}{\varepsilon^2 p} \log\left(\frac{2}{\delta}\right) \implies P\left(\frac{|S_n - p|}{p} > \varepsilon\right) \leq \delta. \quad (6.27)$$

Let P, Q be two distributions on $\mathcal{Y} = \{0, 1\}$, and let $\mathcal{D}_P \sim P, \mathcal{D}_Q \sim Q$ denote IID samples of size n_P, n_Q , respectively. Recall from (6.13) and (6.16) that for each $k \in \{0, 1\}$, we define

$$w_k^* = \frac{P_Q[Y = k]}{P_P[Y = k]}, \quad \text{and} \quad \hat{w}_k = \frac{P_{\mathcal{D}_Q}[Y = k]}{P_{\mathcal{D}_P}[Y = k]}.$$

Then, we let

$$\rho_0 := \frac{\hat{w}_0}{w_0^*} \quad \text{and} \quad \rho_1 := \frac{\hat{w}_1}{w_1^*}. \quad (6.28)$$

Now we define another parameterized family of Boolean-valued functions $\Phi_{\text{ratio}}(\mathcal{D}_P, \mathcal{D}_Q; \beta)$ as follows. Given $\mathcal{D}_P \sim P$, $\mathcal{D}_Q \sim Q$, and $\beta \in \mathbb{R}$ such that $1 < \beta \leq 2$,

$$\Phi_{\text{ratio}}(\mathcal{D}_P, \mathcal{D}_Q; \beta) := \mathbb{1} \left\{ \frac{1}{\beta} \leq \rho_k \leq \beta, \quad \forall k \in \{0, 1\} \right\}. \quad (6.29)$$

Corollary 6.9. *Let P, Q be two distributions on $\mathcal{Y} = \{0, 1\}$, and let $\mathcal{D}_P \sim P$, $\mathcal{D}_Q \sim Q$ denote IID samples of size n_P , n_Q , respectively. For each $k \in \{0, 1\}$, let $p_k := P_P[Y = k]$ and $q_k := P_Q[Y = k]$. Likewise, we let $\hat{p}_k = \frac{1}{n_P} \sum_{y_i \in \mathcal{D}_P} \mathbb{1}\{y_i = k\}$ and $\hat{q}_k = \frac{1}{n_Q} \sum_{y_i \in \mathcal{D}_Q} \mathbb{1}\{y_i = k\}$. For any $\delta \in (0, 1)$ and any $\beta \in (1, 2]$, if*

$$n_P \geq \frac{27}{(\beta - 1)^2 \min\{p_0, p_1\}} \log \left(\frac{8}{\delta} \right) \quad \text{and} \quad n_Q \geq \frac{27}{(\beta - 1)^2 \min\{q_0, q_1\}} \log \left(\frac{8}{\delta} \right),$$

then

$$P(\Phi_{\text{ratio}}(\mathcal{D}_P, \mathcal{D}_Q; \beta) = 1) \geq 1 - \delta.$$

Proof of Corollary 6.9. Let $\varepsilon = \frac{\beta-1}{3}$. Since $\frac{1+x}{1-x} \leq 1 + 3x$ for all $x \in [0, 1/3]$, we have $\frac{1}{\beta} \leq \frac{1-\varepsilon}{1+\varepsilon} < \frac{1+\varepsilon}{1-\varepsilon} \leq \beta$. Then it follows from (6.27) that for each $k \in \{0, 1\}$,

$$\begin{aligned} n_P \geq \frac{3}{\varepsilon^2 p_k} \log \left(\frac{8}{\delta} \right) &\implies P \left(\frac{|\hat{p}_k - p_k|}{p_k} > \varepsilon \right) \leq \frac{\delta}{4}, \\ n_Q \geq \frac{3}{\varepsilon^2 q_k} \log \left(\frac{8}{\delta} \right) &\implies P \left(\frac{|\hat{q}_k - q_k|}{q_k} > \varepsilon \right) \leq \frac{\delta}{4}. \end{aligned}$$

Applying the union bound, we obtain the following implication:

$$\begin{aligned} n_P \geq \frac{3}{\varepsilon^2 \min\{p_0, p_1\}} \log \left(\frac{8}{\delta} \right) \quad \text{and} \quad n_Q \geq \frac{3}{\varepsilon^2 \min\{q_0, q_1\}} \log \left(\frac{8}{\delta} \right) \\ \implies P \left(\max_{k \in \{0, 1\}} \frac{|\hat{p}_k - p_k|}{p_k} > \varepsilon \quad \text{or} \quad \max_{k \in \{0, 1\}} \frac{|\hat{q}_k - q_k|}{q_k} > \varepsilon \right) \leq \delta \\ \implies P \left(\max_{k \in \{0, 1\}} \rho_k > \frac{1+\varepsilon}{1-\varepsilon} \quad \text{or} \quad \min_{k \in \{0, 1\}} \rho_k < \frac{1-\varepsilon}{1+\varepsilon} \right) \leq \delta \\ \implies P \left(\max_{k \in \{0, 1\}} \rho_k > \beta \quad \text{or} \quad \min_{k \in \{0, 1\}} \rho_k < \frac{1}{\beta} \right) \leq \delta. \end{aligned}$$

□

6.10.1.2 Regularity of the Shift Correction Function

Lemma 6.10. *Let $w = (w_0, w_1) \in \mathbb{R}^2$ such that $w_0, w_1 > 0$ and $w_0 + w_1 = 1$. The function $g_w : [0, 1] \rightarrow [0, 1]$ such that $g_w(z) = \frac{w_1 z}{w_1 z + w_0(1-z)}$ is L -Lipschitz where $L = \max \left\{ \frac{w_1}{w_0}, \frac{w_0}{w_1} \right\}$.*

Proof of Lemma 6.10. First of all, consider the first-order derivative of g_w :

$$\frac{d}{dz}g_w(z) = \frac{w_1 \cdot [w_1z + w_0(1-z)] - w_1z \cdot (w_1 - w_0)}{[w_1z + w_0(1-z)]^2} = \frac{w_1w_0}{[w_1z + w_0(1-z)]^2}.$$

We observe that g_w is monotone increasing as $\frac{d}{dz}g_w(z) > 0$ for all $z \in [0, 1]$. Next, we consider the second-order derivative of g_w :

$$\frac{d^2}{dz^2}g_w(z) = \frac{2w_0w_1 \cdot (w_0 - w_1)}{[w_1z + w_0(1-z)]^3} \begin{cases} > 0, & \forall z \in [0, 1] & \text{if } w_0 > w_1, \\ = 0, & \forall z \in [0, 1] & \text{if } w_0 = w_1, \\ < 0, & \forall z \in [0, 1] & \text{if } w_0 < w_1. \end{cases}$$

Therefore,

$$\sup_{z \in [0,1]} \frac{d}{dz}g_w(z) = \begin{cases} \left. \frac{d}{dz}g_w(z) \right|_{z=1} = \frac{w_0}{w_1} & \text{if } w_0 > w_1, \\ \left. \frac{d}{dz}g_w(z) \right|_{z=0} = \frac{w_1}{w_0} & \text{if } w_0 \leq w_1. \end{cases}$$

□

Lemma 6.11. *Let P, Q be joint distributions of $(X, Y) \in \mathcal{X} \times \{0, 1\}$, and let $w_k = \frac{P[Y=k]}{Q[Y=k]}$ for $k \in \{0, 1\}$. If P, Q satisfy the label shift assumption (Definition 6.7), i.e., if Assumptions (B1) and (B2) hold, then for any measurable function $f : \mathcal{X} \rightarrow \mathbb{R}$, the following two-sided inequality holds:*

$$\min_{k \in \{0,1\}} w_k \leq \frac{\mathbb{E}_Q[f(X)]}{\mathbb{E}_P[f(X)]} \leq \max_{k \in \{0,1\}} w_k. \quad (6.30)$$

Proof of Lemma 6.11. First of all, we observe that

$$\begin{aligned} \mathbb{E}_Q[f(X)] &= \mathbb{E}_Q[\mathbb{E}_Q[f(X) | Y]] && \text{by the law of total expectation} \\ &= \sum_{k=0}^1 P_Q[Y = k] \cdot \mathbb{E}_Q[f(X) | Y] \\ &= \sum_{k=0}^1 (w_k \cdot P_P[Y = k]) \cdot \mathbb{E}_P[f(X) | Y]. && \text{by definition of } w_k \text{ \& Assumption (B1)} \end{aligned}$$

Thus, it follows that $\min_k w_k \cdot \mathbb{E}_P[f(X)] \leq \mathbb{E}_Q[f(X)] \leq \max_k w_k \cdot \mathbb{E}_P[f(X)]$. □

6.10.2 Completing the proof of Theorem 6.2

Proof of Theorem 6.2. This proof is presented in four steps. In Step 1, we establish a simple upper bound for the risk $R_Q(\hat{h}_Q; f)$ that consists of two error terms: the first term quantifies the error introduced by the estimated label shift correction, $\hat{g}$, while the second term quantifies the error due to the estimated recalibration function, $\hat{h}_P$. In Steps 2 and 3, we derive separate upper bounds for these two error terms. Finally, in Step 4, we combine the results from Steps 1-3 to obtain a comprehensive upper bound for R_Q , which concludes the proof.

Step 1. Decomposition of R_Q . Recalling the definition of the risk R_Q , cf. (6.5), we obtain the following inequality:

$$\begin{aligned} R_Q(\hat{h}_Q; f) &= \mathbb{E}_Q \left[\left(\hat{h}_Q \circ f(X) - \mathbb{E}_Q[Y|f(X)] \right)^2 \right] \\ &= \mathbb{E}_Q \left[\left(\hat{g} \circ \hat{h}_P \circ f(X) - g^* \circ \hat{h}_P \circ f(X) + g^* \circ \hat{h}_P \circ f(X) - \mathbb{E}_Q[Y|f(X)] \right)^2 \right] \\ &\stackrel{(a)}{\leq} 2 \cdot \left\{ \underbrace{\mathbb{E}_Q \left[\left(\hat{g} \circ \hat{h}_P \circ f(X) - g^* \circ \hat{h}_P \circ f(X) \right)^2 \right]}_{=:T_1} \right. \end{aligned} \quad (6.31)$$

$$\left. + \underbrace{\mathbb{E}_Q \left[\left(g^* \circ \hat{h}_P \circ f(X) - \mathbb{E}_Q[Y|f(X)] \right)^2 \right]}_{=:T_2} \right\}, \quad (6.32)$$

where (a) follows from the simple inequality $(a + b)^2 \leq 2(a^2 + b^2)$ for all $a, b \in \mathbb{R}$.

In Step 2 and Step 3 of this proof, we establish separate upper bounds for the two terms, T_1, T_2 .

Step 2. An upper bound for T_1 . Recall from (6.13) and (6.16) that

$$\begin{aligned} g^*(z) &= \frac{w_1^* z}{w_1^* z + w_0^* (1 - z)} & \text{where} & & w_k^* &= \frac{Q[Y = k]}{P[Y = k]}, \quad \forall k \in \{0, 1\}, \\ \hat{g}(z) &= \frac{\hat{w}_1 z}{\hat{w}_1 z + \hat{w}_0 (1 - z)} & \text{where} & & \hat{w}_k &= \frac{\hat{Q}[Y = k]}{\hat{P}[Y = k]}, \quad \forall k \in \{0, 1\}. \end{aligned}$$

Let

$$\rho_0 := \frac{\hat{w}_0}{w_0^*} \quad \text{and} \quad \rho_1 := \frac{\hat{w}_1}{w_1^*}. \quad (6.33)$$

Then we observe that for any $z \in (0, 1)$,

$$\begin{aligned}
|\hat{g}(z) - g^*(z)| &= \left| \frac{\hat{w}_1 z}{\hat{w}_1 z + \hat{w}_0(1-z)} - \frac{w_1^* z}{w_1^* z + w_0^*(1-z)} \right| \\
&= \left| \frac{(\hat{w}_1 w_0^* - w_1^* \hat{w}_0) \cdot z(1-z)}{[\hat{w}_1 z + \hat{w}_0(1-z)] \cdot [w_1^* z + w_0^*(1-z)]} \right| \\
&\leq \left| \frac{(\hat{w}_1 w_0^* - w_1^* \hat{w}_0) \cdot z(1-z)}{(\hat{w}_1 w_0^* + w_1^* \hat{w}_0) \cdot z(1-z)} \right| \\
&= \left| \frac{\hat{w}_1 w_0^* - w_1^* \hat{w}_0}{\hat{w}_1 w_0^* + w_1^* \hat{w}_0} \right| \\
&= \frac{|\rho_0 - \rho_1|}{\rho_0 + \rho_1}.
\end{aligned}$$

Moreover, $\hat{g}(0) = g^*(0) = 0$ and $\hat{g}(1) = g^*(1) = 1$. Letting $Z_{\hat{h}} := \hat{h}_P \circ f(X)$, we obtain

$$T_1 = \mathbb{E}_Q \left[\left(\hat{g}(Z_{\hat{h}}) - g^*(Z_{\hat{h}}) \right)^2 \right] \leq \left(\frac{\rho_0 - \rho_1}{\rho_0 + \rho_1} \right)^2. \quad (6.34)$$

It remains to establish probabilistic tail bounds for ρ_0, ρ_1 , which we will accomplish in Step 4 of this proof.

Step 3. An upper bound for T_2 . We observe that

$$\begin{aligned}
T_2 &= \mathbb{E}_Q \left[\left(g^* \circ \hat{h}_P \circ f(X) - \mathbb{E}_Q[Y \mid f(X)] \right)^2 \right] \\
&= \mathbb{E}_Q \left[\left(g^* \circ \hat{h}_P \circ f(X) - g^*(\mathbb{E}_P[Y \mid f(X)]) \right)^2 \right] && \because \text{by Equation (6.14)} \\
&\leq \left(\frac{w_{\max}^*}{w_{\min}^*} \right)^2 \cdot \mathbb{E}_Q \left[\left(\hat{h}_P \circ f(X) - \mathbb{E}_P[Y \mid f(X)] \right)^2 \right] && \because \text{Lemma 6.10} \\
&\leq \left(\frac{w_{\max}^*}{w_{\min}^*} \right)^2 w_{\max}^* \mathbb{E}_P \left[\left(\hat{h}_P \circ f(X) - \mathbb{E}_P[Y \mid f(X)] \right)^2 \right] && \because \text{by Lemma 6.11} \\
&= \frac{w_{\max}^*{}^3}{w_{\min}^*{}^2} \cdot R_P(\hat{h}_P; f).
\end{aligned}$$

Step 4. Concluding the proof. For given $\delta \in (0, 1)$, let¹ $\delta_1 = \delta_2 = \delta/4$ and $\delta_3 = \delta/2$. We observe that

$$n_P \geq c' \cdot |\mathcal{B}| \log \left(\frac{|\mathcal{B}|}{\delta_1} \right) \implies P[\Phi_{\text{balance}}(\mathcal{B}, 2) = 1] \geq 1 - \delta_1 \quad \text{by Lemma 6.3}$$

$$n_P \geq 2|\mathcal{B}| \implies P[\Phi_{\text{approx}}(\mathcal{B}, \varepsilon_{\delta_2}) = 1] \geq 1 - \delta_2 \quad \text{by Lemma 6.4}$$

where $c' > 0$ is the universal constant that appears in Lemma 6.3 and

$$\varepsilon_{\delta_2} = \sqrt{\frac{1}{2(\lfloor n/|\mathcal{B} \rfloor - 1)} \log \left(\frac{2|\mathcal{B}|}{\delta_2} \right)} + \frac{1}{\lfloor n/|\mathcal{B} \rfloor}.$$

Furthermore, assuming

$$n_P \geq \frac{27}{\min\{p_0, p_1\}} \log \left(\frac{8}{\delta_3} \right) \quad \text{and} \quad n_Q \geq \frac{27}{\min\{q_0, q_1\}} \log \left(\frac{8}{\delta_3} \right),$$

we may define β_{δ_3} as a function of n_P, n_Q and δ_3 such that

$$\beta_{\delta_3} = \beta_{\delta_3}(n_P, n_Q) := 1 + \sqrt{\max \left\{ \frac{1}{n_P \cdot \min\{p_0, p_1\}}, \frac{1}{n_Q \cdot \min\{q_0, q_1\}} \right\} \cdot 27 \log \left(\frac{8}{\delta_3} \right)}. \quad (6.35)$$

Then it follows from Corollary 6.9 that

$$P(\Phi_{\text{ratio}}(\mathcal{D}_P, \mathcal{D}_Q; \beta_0) = 1) \geq 1 - \delta_3.$$

Observe that $\delta_1 = \frac{\delta}{4} < \frac{1}{4}$ and $|\mathcal{B}| \geq 4$, and thus, $\log \left(\frac{|\mathcal{B}|}{\delta_1} \right) \geq \log 16 \geq 2$. Let $c = c'$. Since $c' \geq 1$ and $\log \left(\frac{|\mathcal{B}|}{\delta_1} \right) \geq \log \left(\frac{16}{\delta} \right) = \log \left(\frac{8}{\delta_3} \right)$, we notice that

$$\begin{aligned} n_P &\geq \max \left\{ c, \frac{27}{\min\{p_0, p_1\}} \right\} \cdot |\mathcal{B}| \log \left(\frac{4|\mathcal{B}|}{\delta} \right) \\ \implies n_P &\geq \max \left\{ c' \cdot |\mathcal{B}| \log \left(\frac{|\mathcal{B}|}{\delta_1} \right), 2|\mathcal{B}|, \frac{27}{\min\{p_0, p_1\}} \log \left(\frac{8}{\delta_3} \right) \right\}. \end{aligned}$$

¹We remark that our decomposition of δ into $\delta_1, \delta_2, \delta_3$ is arbitrary, and is intended to simplify the subsequent analysis.

In summary, we obtain that for any given $\delta \in (0, 1)$,

$$\begin{aligned} n_P &\geq \max \left\{ c, \frac{27}{\min\{p_0, p_1\}} \right\} \cdot |\mathcal{B}| \log \left(\frac{4|\mathcal{B}|}{\delta} \right) \quad \text{and} \quad n_Q \geq \frac{27}{\min\{q_0, q_1\}} \log \left(\frac{16}{\delta} \right) \\ \implies P \left[\Phi_{\text{balance}}(\mathcal{B}, 2) = 1 \ \& \ \Phi_{\text{approx}}(\mathcal{B}, \varepsilon_{\delta/4}) = 1 \ \& \ \Phi_{\text{ratio}}(\mathcal{D}_P, \mathcal{D}_Q; \beta_{\delta/2}) = 1 \right] &\geq 1 - \delta. \end{aligned} \quad (6.36)$$

Conditioned on the event $\Phi_{\text{balance}}(\mathcal{B}, 2) = 1 \ \& \ \Phi_{\text{approx}}(\mathcal{B}, \varepsilon_{\delta/4}) = 1 \ \& \ \Phi_{\text{ratio}}(\mathcal{D}_P, \mathcal{D}_Q; \beta_{\delta/2}) = 1$,

$$\begin{aligned} T_1 &\leq \left(\frac{|\rho_0 - \rho_1|}{\rho_0 + \rho_1} \right)^2 \leq \left(\frac{\beta_{\delta/2} - \frac{1}{\beta_{\delta/2}}}{\beta_{\delta/2} + \frac{1}{\beta_{\delta/2}}} \right)^2 \leq (\beta_{\delta/2} - 1)^2, \quad \because (6.34); \text{ also, see (6.29)} \\ T_2 &\leq \frac{w_{\max}^*}{w_{\min}^*} \frac{3}{2} \cdot R_P(\hat{h}_P; f) \\ &\leq \frac{w_{\max}^*}{w_{\min}^*} \frac{3}{2} \cdot \left(\varepsilon_{\delta/4}^2 + \frac{2}{|\mathcal{B}|} \right). \quad \because \text{proof of Theorem 6.1; Lemmas 6.5 \& 6.6} \end{aligned}$$

Note that if Assumption (A3) holds, then we additionally have

$$T_2 \leq \frac{w_{\max}^*}{w_{\min}^*} \frac{3}{2} \cdot \left(\varepsilon_{\delta/4}^2 + \frac{8K^2}{|\mathcal{B}|^2} \right).$$

Inserting these upper bounds for T_1 and T_2 into (6.31), (6.32) and recalling the expression for β in (6.35), we complete the proof. $\square$

6.11 Appendix: Proof of the argument in Remark 6.6

Recall our composite recalibration function,

$$\hat{h}_Q = \hat{g} \circ \hat{h}_P, \quad (6.37)$$

does not use the features in $\mathcal{D}_Q$. Specifically, $\hat{g}$ is parameterized by $\hat{w} = (\hat{w}_0, \hat{w}_1)$, which can be estimated using only labels in $\mathcal{D}_P$ and $\mathcal{D}_Q$, cf. (6.16). According to Theorem 6.2, $R_Q(\hat{h}_Q) = O(n_Q^{-1})$ with high probability for sufficiently large n_P .

Now suppose we are given an unlabeled target sample with unknown label shift. We can estimate w using the target features via a maximum likelihood label shift estimation approach [90], yielding $\hat{w}^{\text{ML}} = (\hat{w}_0^{\text{ML}}, \hat{w}_1^{\text{ML}})$. and the calibrated classifier $\hat{h} \circ f$. This results

in a different composite recalibration function than Equation (6.37),

$$\hat{h}_Q^{\text{ML}} = \hat{g}^{\text{ML}} \circ \hat{h}_P, \quad (6.38)$$

where $\hat{g}^{\text{ML}} : [0, 1] \rightarrow [0, 1]$ is defined as $\hat{g}^{\text{ML}}(z) = \hat{w}_1^{\text{ML}} z / (\hat{w}_1^{\text{ML}} z + \hat{w}_0^{\text{ML}}(1 - z))$. We claim in Remark 6.6 that, for sufficiently large n_P , the composite recalibration function in Equation (6.38) achieves $R_Q(\hat{h}_Q^{\text{ML}}) = O(n_Q^{-1})$ with high probability, enjoying the same convergence rate as $\hat{h}_Q$ (Equation 6.37). Here we provide a proof.

Proof. Suppose we use the maximum likelihood approach in [90] to estimate w . We want to show $R_Q(\hat{h}_Q^{\text{ML}}) = O(n_Q^{-1})$ with high probability for sufficiently large n_P . Recall $R_Q(\hat{h}_Q) \leq 2(T_1 + T_2)$ according to Equation (6.31) and (6.32), and label shift estimation error only affects T_1 , so it is sufficient to show $T_1 = O(n_Q^{-1})$ with high probability.

For sufficiently large n_P , $\hat{h}_P \circ f$ is sufficient calibrated, so $\|\hat{w} - w\|_2^2 = O(n_Q^{-1})$ by Theorem 3 in [90]. Since

$$\|\hat{w} - w\|_2^2 = \sum_{k \in \{0,1\}} (\rho_k - 1)^2 w_k^2 \geq w_{\min}^{*2} \sum_{k \in \{0,1\}} (\rho_k - 1)^2 \geq w_{\min}^{*2} \max_{k \in \{0,1\}} (\rho_k - 1)^2, \quad (6.39)$$

we have $\rho_k \in [1 - \alpha, 1 + \alpha]$ for $k \in \{0, 1\}$, where $\alpha = \frac{\|\hat{w} - w\|_2}{w_{\min}^*} > 0$. For sufficiently small $\|\hat{w} - w\|_2^2$, we can control $\alpha < 0.5$, which bounds T_1 in (6.34):

$$T_1 \leq \left(\frac{\rho_0 - \rho_1}{\rho_0 + \rho_1} \right)^2 \leq \frac{2\alpha}{2 - 2\alpha} \leq 2\alpha = 2 \frac{\|\hat{w} - w\|_2}{w_{\min}^*} = O(n_Q^{-1}). \quad (6.40)$$

The rest of the proof are the same with Appendix 6.10.2. $\square$

6.12 Appendix: Details on the experiments

In Section 6.12.1 and 6.12.3, we consider a family of joint distributions $\mathcal{D}(\pi)$ of X and Y , where $Y \sim \text{Bernoulli}(\pi)$, $X \mid Y = 0 \sim N(-2, 1)$, and $X \mid Y = 1 \sim N(2, 1)$. Suppose we are given $f(x) = \sigma(x) := 1/(1 + e^{-x})$, for $x \in \mathbb{R}$, as a probabilistic classifier. The optimal recalibration function can be derived as

$$h_{f,P}^*(z) = P[Y = 1 \mid f(X) = z] = \sigma(4\sigma^{-1}(z)). \quad (6.41)$$

In Section 6.12.2, we consider a parametric family of recalibration functions called beta calibration [160]: $\mathcal{H}_{\text{beta}} = \{h_{\text{beta}}(\cdot; a, b, c) : a \geq 0, b \geq 0, c \in \mathbb{R}\}$, where $h_{\text{beta}}(\cdot; a, b, c) :$

$[0, 1] \rightarrow [0, 1]$ is defined as

$$h_{\text{beta}}(z; a, b, c) = \frac{1}{1 + 1/\left(e^c \frac{z^a}{(1-z)^b}\right)}. \quad (6.42)$$

In addition, consider a subfamily $\mathcal{H}_{\text{logit-normal}} \subset \mathcal{H}_{\text{beta}}$ defined as $\mathcal{H}_{\text{logit-normal}} = \{h_{\text{logit-normal}}(\cdot; a, c) := h_{\text{beta}}(\cdot; a, a, c) = \sigma(a\sigma^{-1}(\cdot) + c) : a \geq 0, c \in \mathbb{R}\}^2$. Apparently, the optimal recalibration function in Equation (6.41), $h_{f,P}^* \in \mathcal{H}_{\text{logit-normal}}$.

6.12.1 Verifying results for UMB

First, we recalibrate f on data distributed as $\mathcal{D}(0.5)$ using UMB.

Verifying the risk convergence in Theorem 6.1 We vary $n \in [10^2, 10^7]$ and $B \in [6, 10^3]$ in the log scale. For each combination of (n, B) , we use UMB to recalibrate f on data generated from $\mathcal{D}(0.5)$, and compute quadrature estimates of population $R^{\text{cal}}(\hat{h})$, $R^{\text{sha}}(\hat{h})$, and $R(\hat{h})$, as well as their high probability bounds based on Theorem 6.1. The constant K in Assumption (A3) is selected by numerical maximization as

$$K = \max_{0 \leq z_1 < z_2 \leq 1} \frac{h^*(z_2) - h^*(z_1)}{P[Z \in [z_1, z_2]]}.$$

Figure 6.1 shows the bounds follow the same trends as their associated population quantities, providing valid upper bounds in all cases.

Verifying the optimal choice of the number of bins. We find empirically optimal $B^{\text{experiment}}$ that achieves the minimal risk for each choice of n . We compute the theoretically optimal choice of the number of bins, B^{theory} , by minimizing the finite-sample upper bounds. Figure 6.2 shows $B^{\text{experiment}}$ follows the same trend with B^{theory} , both scales in $O(n^{1/3})$.

6.12.2 Comparing recalibration methods

To highlight the benefits and drawbacks of UMB's nonparametric nature, we compare UMB with (semi-)parametric recalibration methods in scenarios where the parametric assumption is correct and where it is misspecified. We compare the method under study, uniform-mass

²We say $Z \sim \text{Logit-Normal}(\mu, \tau^2)$ if $\sigma^{-1}(Z) \sim N(\mu, \tau^2)$ [5]. Similar to beta calibration [160], we adopt the name “logit-normal calibration” after a simple example: if $Y = \text{Bernoulli}(0.5)$, $Z \mid Y = i \sim \text{Logit-Normal}(\mu_i, \tau_i^2)$ for $i \in \{0, 1\}$, then the optimal recalibration function $\mathbb{E}[Y \mid Z = z] = h_{\text{logit-normal}}(z; a, c)$ for some a, c depending on μ_i 's and τ_i 's.

binning (UMB), with 3 other recalibration methods: uniform-width binning (UWB) [109], Platt scaling [229]³, and a hybrid parametric-binning method [162]. Note that Platt scaling and the hybrid method adopt the parametric assumption $h^* \in \mathcal{H}_{\text{logit-normal}}$.

For the first setting, we construct optimal recalibration function $h^* \in \mathcal{H}_{\text{logit-normal}}$ so that the parametric assumption of Platt scaling and the hybrid method holds. In particular, we consider the distribution $Z \in \text{Uniform}[0, 1]$ and $Y \mid Z \sim \text{Bernoulli}(h_{\text{logit-normal}}(Z; a, c))$ with $a = 4$ and $c = 0$. For the second setting, we construct $h^* \in \mathcal{H}_{\text{beta}}$ but $h^* \notin \mathcal{H}_{\text{logit-normal}}$ so that the parametric assumption fails. In particular, we consider the distribution $Z \sim \text{Uniform}[0, 1]$ and $Y \mid Z \sim \text{Bernoulli}(h_{\text{beta}}(z; a, b, c))$ with $a = 0.1$, $b = 4$, and $c = 0$. For each setting, we fix calibration sample size to be $n = 5000$.

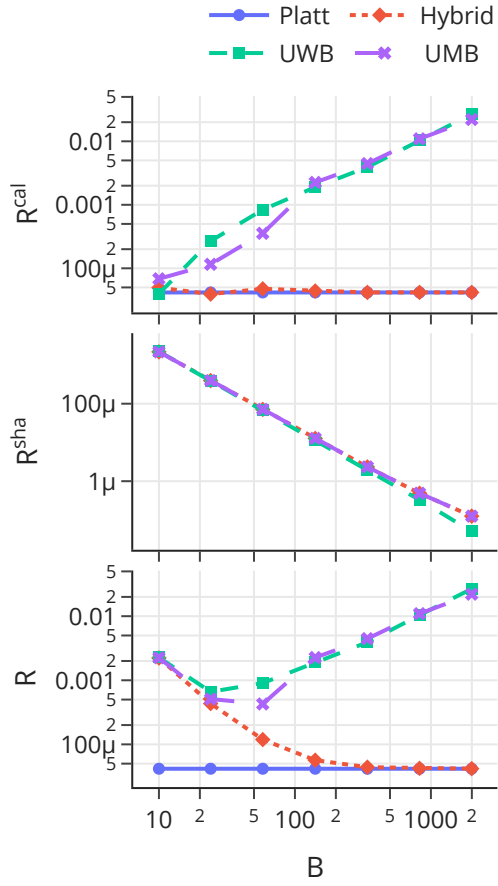
Risks as functions of the number of bins B We traverse the number of bins $B \in [10, 2000]$ in the log scale and compare how each method behaves as B changes. When the parametric assumption is correct, the hybrid method achieves significantly lower R^{cal} and overall R than UMB and UWB for sufficiently large number of bins (Figure 6.5a), an advantage highlighted in [162]. In contrast, when the parametric assumption fails, the binning methods UMB and UWB has better performance with the optimal number of bins (Figure 6.5b). This is because Platt scaling and hybrid methods are intrinsically biased when $h^* \notin \mathcal{H}_{\text{logit-normal}}$, as noted in Section 6.4.2.

Quantitative results of risks under optimal B For each setting, we fix B that achieves low recalibration risk for UWB and UMB in Figure 6.5. Specifically, we choose $B = 2 \lfloor n^{1/3} \rfloor = 34$ for the correct parametric assumption setting, and $B = \lfloor n^{1/3} \rfloor = 17$ for the misspecified parametric assumption setting. Then, for each setting, we compare the 90% quantiles of risks of each recalibration method fitted on 100 random replicates of calibration datasets of size $n = 5000$.

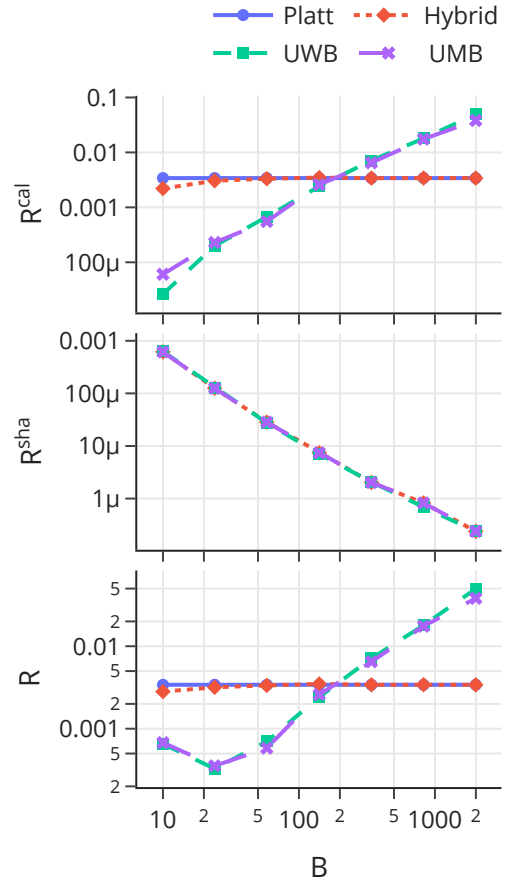
Table 6.1 quantitatively verifies that Platt and the Hybrid method achieves lower R^{cal} and overall R if the parametric assumption is correct, and UWB and UMB achieves lower R^{cal} and overall R when the parametric assumption fails.

Visualization of calibration curves We fix the calibration dataset and visualize the calibration curves for all methods under the two settings. Figure 6.3 shows that the binning methods (UWB and UMB) closely track the optimal recalibration function h^* in both

³The original Platt scaling operates on outputs of real-valued SVM outputs [229]. For probabilistic classifiers, we follow [222, 162, 112] and implement Platt scaling by first transforming probabilities onto the real line via the logit transform σ^{-1} .



(a) Correct parametric assumption



(b) Misspecified parametric assumption

Figure 6.5: Risks vs. number of bins B .

settings. In contrast, the hybrid approach follows the Platt scaling estimates, leading to an inherent bias from h^* when the parametric assumption is invalid (Figure 6.3b).

6.12.3 Comparing recalibration schemes under label shift

We consider the label shift with source distribution $\mathcal{D}(0.5)$ and target distribution $\mathcal{D}(\pi_Q)$, where π_Q varies in $\{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$. The results where $\pi_Q > 0.5$ can be inferred by symmetry and hence not experimented. We vary n_P in $\{10, 10^3, 10^5, 10^7\}$ and n_Q in $\{10, 10^3, 10^5\}$.

Aside from our proposed recalibration function $\hat{h}_Q = \hat{g} \circ \hat{h}_P$ (6.17), referred to as COMPOSITE, we consider three other calibration approaches as baselines: (1) SOURCE, denoted as $\hat{h}_P$, which is only calibrated on the source data, (2) LABEL-SHIFT, denoted as $\hat{g}$, which performs label shift correction without calibration, and (3) TARGET, denoted as $\hat{g}_Q^{\text{target}}$, which is only calibrated on the target data. The number of bins B are chosen to be $n_P^{1/3}$ for COMPOSITE and SOURCE, and $n_Q^{1/3}$ for TARGET.

Table 6.2 shows the risks for different approaches with $\pi_Q = 0.1$, $n_P = 10^3$, and $n_Q = 10^2$. In terms of R^{cal} , COMPOSITE performs the best, as it is calibrated to the target distribution by taking advantage of the abundant source data. In terms of R^{sha} , LABEL-SHIFT achieves $R^{\text{sha}} = 0$ due to the strictly increasing $\hat{g}$, but it suffers from high R^{cal} . COMPOSITE and SOURCE achieve smaller R^{sha} than TARGET, as a result of using more bins on a larger sample. Considering the combined impact of calibration and sharpness, our approach COMPOSITE attains the lowest overall recalibration risk R as well as MSE.

Figure 6.6 shows the optimal recalibration function h^* and the recalibration functions for the four approaches. It can be seen that COMPOSITE best estimates h^* with the highest resolution.

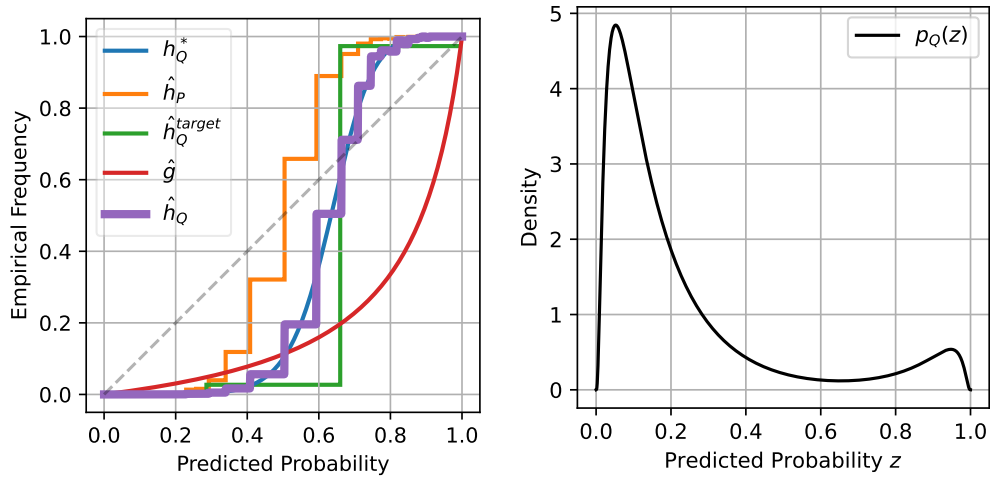


Figure 6.6: *Left*: calibration curves of for COMPOSITE $\hat{h}_Q$, SOURCE $\hat{h}_P$, TARGET $\hat{h}_Q^{\text{target}}$, and LABEL-SHIFT $\hat{g}$. *Right*: the marginal density of $Z = f(X)$ under Q .

CHAPTER 7

Solar Flare Prediction under Event Rate Shift

7.1 Introduction

Solar flares are electrical magnetic explosions on the sun that last in the time scale of minutes. Strong solar flares and accompanying highly energetic solar events can expose astronauts and high-altitude flights to radiation, disrupt terrestrial communications, and causing power grid failures. Therefore, predicting solar flares is an important task and has been researched for decades.

Recently, the increasing quality and quantity of data have spurred the use of machine learning methods in predicting solar flares. A machine learning-based solar flare forecasting system typically begins with training models on historical data, aiming to identify correlation between strong flare events and preflare features. These trained models are then deployed operationally to continuously make predictions, which are subsequently verified against actual observations.

However, the successful generalization of machine learning models depends on the similarity between the distributions of the training and test data. In the context of solar flare prediction, achieving such similarity is challenging, particularly due to event rate disparities between the training and test data. Two common causes contribute to this event rate shift, which we detail below.

The first cause is the naturally varying event rate due to the solar cycle progression. Solar activities follows an approximately 11-year periodic solar cycle, during which the frequency and intensity of many solar events, including solar flares, exhibit a synchronized increase to the solar maxima, followed by a decrease to the solar minima. During solar maxima, multiple strong solar flares can occur within a single day, whereas, during solar minima, months may pass without a significant event. Moreover, the overall solar flare event rate can differ from one solar cycle to another [201]. These natural variations in event rates may cause machine learning models to systematically over- or under-forecast in operational solar

flare prediction.

The second cause of the event rate shift is artificial. Strong solar flares, particularly catastrophic ones with severe terrestrial impact, are rare events [171]. Consequently, historical data typically contain an overwhelming number of negatively-labeled, non-event samples, with relatively few positively-labeled event samples. To prevent machine learning models from being dominated by the negative class, rebalancing techniques, such as oversampling the positive class and/or undersampling the negative class, are commonly employed [4]. In Chapter 2, we address this issue by incorporating data from a previous solar cycle with more strong solar flares and by undersampling the negative class. However, this rebalancing artificially inflate the event rate, making machine learning models tend to overforecast and produce many more false alarms when deployed into the real world [174, 175].

There are three types of approaches to consider the influence of the event rate in flare forecasting.

The first type of approaches uses the event rate in the past as a direct estimate of the flaring probability in the future. It is known that the event rate in the near past, or persistence, is a good predictor for the event rate in the future. The event rate in a longer time window, or climatology, is also useful in providing a baseline of the flare occurrence. Murphy [208] discussed climatology, persistence, and their linear combination as reference forecasts in computing skill scores. Wheatland [316, 317, 318] proposed Bayesian approaches that only uses the event statistics of flares. The basic assumptions are two folds: (1) the peak flux in soft X-rays of flares follows a power-law distribution, and (2) the flaring time in short timescales follow a Poisson process. The author found that this approach provides a strong baseline for the whole-Sun prediction of flares. Lim et al. [182] predicts flare occurrence rates as a weighted combination of the long-term (several years), the mid-term (several days), and the short-term (1 day) flaring rates. They performed linear regression on the log-scale of SHARP parameters and the event rates. The results show that the long-term event rate contributes the most towards the Brier Skill Score (BSS), and taking the short- and mid-term event rates into consideration further improves the performance.

The second type of approaches use the past event rate as one of many features and let machine learning models to find their correlation to future flaring probabilities. To characterize the flare event rate in the past, many works use the *flare index*, which is measured by the sum of the peak intensities of flares over a specified time period [1, 246, 48, 168]. For example, [1] defined the daily soft X-ray flare index as $(100S^{(X)} + 10S^{(M)} + 1.0S^{(C)} + 0.1S^{(B)})/\tau$, where $S^{(j)} = \sum_{i=1}^{N_j} I_i^{(j)}$ denotes the sum of GOES peak flare intensities of all N_j flares of class j that occurred in the time interval τ . Flare index is shown to be correlated with magnetic power spectrum index [1], Pyonting flux [313], total magnetic flux [47], etc, and is

used in operational flare forecasts [168]. Falconer et al. [82] used prior flaring as a complement to free magnetic energy parameters and found it greatly changes the flaring probability prediction. Jonas et al. [145] constructed flare history features from exponentially decayed peak intensities for flare class C, M and X. They found the flare history features combined with the photospheric vector magnetogram data gives the best performance. They also found flare history features 12 hr prior to the flare is the most useful predictor variable across all time windows from 1 hr to 1000 hr. Nishizuka et al. [217] used one-day history and the total history of major flares in the active region and identify them as more effective features than magnetic properties for flare prediction with certain machine learning algorithms. Compared to the first type of approaches, this type of approaches typically use observables as features in addition to past flaring productivity. Moreover, no strong assumption is made on the underlying process of the arriving times of flares of different intensities. However, many machine learning models are black boxes and the relationship between the prediction and the persistence can be unreliable and hard to interpret. In addition, there's no obvious way to construct an event rate feature that's most informative about the flares. Thus, a series of flaring history features with varying looking-back time window and quantification methods are often provided for the machine learning model [145].

The third type of approaches is to absorb the impact of event rate in evaluation schemes and metrics. One common way to evaluate a probabilistic forecast is to convert it to a deterministic forecast via thresholding, and then use deterministic forecast metrics to evaluate the performance. In solar flare prediction, the threshold should be carefully chosen and the deterministic forecast metrics should be insensitive to the event rate. Typically, a threshold of 0.5 is chosen for the posterior probability [e.g. 56]. This threshold corresponds to the minimum probability error decision rule. However, the probability of error is generally considered as a misleading metric in rare event prediction because Some metrics such as accuracy and Heidke Skill Score (HSS) are sensitive to the class ratios, whereas other metrics such as True Skill Score (TSS) is insensitive to the class ratios. It is desired that the deterministic forecast metrics are insensitive to class imbalance and event rate shift, thus TSS is recommended instead of HSS as the evaluation metric for solar flare prediction [29, 173]. However, accuracy is not considered a good metric for rare events prediction because a consistent negative prediction achieves high accuracy. Some methods do not use the 0.5 probability threshold. They adjust the threshold to maximize a given metric such as TSS [201] or to exercise the false positive control [69]. The adjustment of the threshold is highly dependent on the metric to use.

In this chapter, we propose to Bayesian method to correct for event rate shift. We show that a combined procedure of calibration and label shift correction is able to correct for

the label shift due to both artificial or natural causes in solar flare event rate, improving both classification and calibration metrics. Compared to previous methods, our method is theoretically principled and is also flexible to accommodate the artificial change of event rate due to rebalancing.

7.2 Methods

7.2.1 Mathematical setup

First, we set up some basic notations in probability theory. Let P denote a probability distribution of random variables, with random variables inferred from its argument. Let p denote the probability density function for a continuous random variable, or the probability mass function for a discrete random variable. Let $\mathbb{E}_P$ denotes the expectation of random variables under distribution P . For an event $A \in \mathcal{A}$, let $\mathbb{1}_A$ be the indicator function, which evaluates 1 when A occurs and 0 otherwise.

We consider the task of forecasting a binary event, represented as a Bernoulli random variable Y taking values in $\{0, 1\}$, with value 1 indicating the event of interest occurs. The marginal probability masses of Y are denoted as $\pi = (\pi_0, \pi_1)$, where $\pi_1 = P[Y = 1] = \mathbb{E}[Y]$ known as the *event rate*. We model a probabilistic forecast as a random variable Z that takes values in the interval $[0, 1]$ and is jointly distributed with Y . The probabilistic forecast Z provides an estimated probability for the event to occur. For a machine learning-based probabilistic forecast, Z is produced by a probabilistic classifier f using predictor variables known as features. We represent the features as a random vector X taking values in a Euclidean space $\mathbb{R}^p$. Using a threshold $\zeta \in [0, 1]$, we can discretize the probabilistic forecast Z into deterministic forecast $\hat{Y} = \mathbb{1}_{[Z \geq \zeta]}$, which is a Bernoulli random variable with success probability $P[Z \geq \zeta]$.

The joint distribution of the forecast-observation random variable pair (Z, Y) plays a central role in evaluating the quality of a probabilistic forecast. By conditioning on the observation Y or the forecast Z , we obtain two factorizations of the joint distribution that are useful in characterizing forecast verification [211]: (1) likelihood-base rate factorization: $p(z, y) = p(z|y)p(y)$, and (2) calibration-refinement factorization: $p(z, y) = p(y|z)p(z)$. In a Bayesian perspective, the event rate $\pi_1 = P[Y = 1]$ is often interpreted as the *prior probability* of the event in the sense that it is the probability of $Y = 1$ before making the forecast Z , and the random variable $\Phi = P[Y = 1 | Z]$ is the *posterior probability* of the event after making the forecast Z .

7.2.2 Forecast verification methods

A *deterministic forecast* produces binary label whether the flare event of interest actually occurs, whereas a *probabilistic forecast* produces a continuous score indicating how likely the flare event occurs. Probabilistic forecasts quantify the uncertainty and are increasingly popular, and deterministic forecasts are usually converted from probabilistic forecasts by applying a threshold.

Skill scores is a type of metric that are frequently used to measure the relative skill of a forecast to a reference forecast. Given a metric A (e.g., accuracy, squared error), the skill score can be computed as

$$\text{Skill} = \frac{A_{\text{forecast}} - A_{\text{reference}}}{A_{\text{perfect}} - A_{\text{reference}}}, \quad (7.1)$$

where A_{forecast} , A_{perfect} , and $A_{\text{reference}}$ denotes the metric for the forecast to be evaluated, the perfect forecast, and the reference forecast, respectively. As we will see below, Brier Skill Score (BSS), True Skill Score (TSS), and Heidke Skill Score (HSS) are different forms of skill scores.

7.2.2.1 Probabilistic forecast evaluation

For probabilistic prediction methods, the goal is to obtain an estimate $Z(X)$ to the posterior $\Phi = P[Y = 1|X]$. Some desired attributes of a probabilistic forecast include [320]:

- *Reliability*, or *calibration*, measures how well the probabilistic forecast matches the posterior probability of the event. A probability forecast is known as perfectly *reliable* or *calibrated* if $P[Y = 1|Z = z] = z$ for any z . By definition, given any probabilistic forecast Z , the posterior $\Phi = P[Y = 1|Z]$ is a perfectly reliable forecast. In particular, if Z is independent of Y , the unconditional probability $P[Y]$ is reliable. Therefore, reliability alone does not make a useful forecast [144].
- *Refinement*, or *sharpness*, measures the dispersion of the refinement distribution $p(z)$ and is an attribute for the forecast alone, without regarding to the corresponding observation. A forecast rarely deviates from the event rate exhibit low sharpness, whereas a forecast predicting 1 or 0 exhibit high sharpness.
- *Resolution* measures the degree to which the forecasts sort the observed events into groups that are different from each other.

The quality of the estimation can be quantified using a scoring rule. We assume scoring rules have negatively orientation (the smaller the better) and interpret them as the loss

incurred by forecasting a different distribution than the actual distribution. Perhaps the most commonly used scoring rule is Brier score¹, which defined as

$$\text{BS}(Z) = \mathbb{E}[(Y - Z)^2] \quad (7.2)$$

The Brier score ranges from 0 to 1, with 0 indicting the perfect forecast. As a strictly proper [96, 64] scoring rule², the Brier score can be decomposed into a calibration loss (REL, standing for reliability) and a refinement loss (REF) [282]:

$$\text{BS}(Z) = \mathbb{E}[(Z - Y)^2] \quad (7.3)$$

$$= \mathbb{E}[(Z - \Phi)^2] + \mathbb{E}[(\Phi - Y)^2] + 2\mathbb{E}[(Z - \Phi)(\Phi - Y)] \quad (7.4)$$

$$= \mathbb{E}[(Z - \Phi)^2] + \mathbb{E}[\mathbb{E}[(\Phi - Y)^2|Z]] + 2\mathbb{E}[\mathbb{E}[(Z - \Phi)(\Phi - Y)|Z]] \quad (7.5)$$

$$= \mathbb{E}[(Z - \Phi)^2] + \mathbb{E}[(\Phi - 1)^2\Phi + \Phi^2(1 - \Phi)] + 2\mathbb{E}[(Z - \Phi)(\Phi - \Phi)] \quad (7.6)$$

$$= \underbrace{\mathbb{E}[(Z - \Phi)^2]}_{\text{REL}} + \underbrace{\mathbb{E}[\Phi(1 - \Phi)]}_{\text{REF}} \quad (7.7)$$

REL describes the deviation of Z from the reliable forecast $\Phi = P[Y = 1|Z]$. We can see REL is related to the covariance between Z and Φ :

$$\text{REL}(Z) = \mathbb{E}[(Z - \mathbb{E}[Y|Z])^2] \quad (7.8)$$

$$= \mathbb{E}[(Z - \mathbb{E}Z)^2 + (\mathbb{E}Z - \mathbb{E}Y)^2 + (\mathbb{E}Y - \mathbb{E}[Y|Z])^2] - 2\mathbb{E}[(Z - \mathbb{E}Z)(\mathbb{E}[Y|Z] - \mathbb{E}Y)] \quad (7.9)$$

$$= (\mathbb{E}Z - \mathbb{E}Y)^2 + \text{Var}[Z] + \text{Var}[\Phi] - 2\text{Cov}(Z, \Phi) \quad (7.10)$$

REF can be seen as the irreducible error that cannot be explained by Z . To see this, we first find the optimal estimator of Y based on Z that minimizes $\text{BS}(\delta(Z))$. For an estimator

¹Strictly speaking, the Brier score defined here is a scoring function [144]. A scoring rule maps a probabilistic forecast (a distribution on the range of Y) and an observation (a realization of Y) to a number on the extended real line, whereas a scoring function is a function of two distributions, defined as the expected score with the observation coming from a distribution.

²A scoring rule is *proper* if it achieves the minimum when the predicted probability perfectly matches the actual probability, that is, $S(z, z) \leq S(s, z)$ for all $s, z \in [0, 1]$. It is *strictly proper* if the minimum is unique. See Gneiting and Raftery [96], Dawid and Musio [64] for details.

$\delta = \delta(Z)$, the risk function is

$$\text{BS}(\delta(Z)) = \mathbb{E}[(Y - \delta)^2] \quad (7.11)$$

$$= \mathbb{E}[\mathbb{E}[(Y - \delta)^2 | Z]] \quad (7.12)$$

$$= \mathbb{E}[\mathbb{E}[(Y - \Phi)^2 + (\Phi - \delta)^2 + 2(Y - \Phi)(\Phi - \delta) | Z]] \quad (7.13)$$

$$= \underbrace{\mathbb{E}[\text{Var}(Y | Z)]}_{\text{REF}} + \mathbb{E}[(\Phi - \delta)^2] \quad (7.14)$$

The first term REF is independent of δ , so the optimal estimator is $\delta^* = \Phi$, for which the residual risk is REF. We can see the Brier score can be interpreted as the sum of the cost of recalibration, REL, and unexplained error by the recalibrated forecast, REF.

Another interesting decomposition reveals the resolution component in the Brier score. Consider a special case for Equation (7.14) where δ is the climatology forecast, i.e., $\delta = \mathbb{E}[Y] = \mathbb{E}[\Phi]$. We have

$$\text{BS}(s) = \mathbb{E}[\text{Var}[Y | Z]] + \mathbb{E}[(\Phi - s)^2] \quad (7.15)$$

$$\underbrace{\text{Var}[Y]}_{\text{UNC}} = \underbrace{\mathbb{E}[\Phi(1 - \Phi)]}_{\text{REF}} + \underbrace{\text{Var}[\Phi]}_{\text{RES}} \quad (7.16)$$

where UNC is the uncertainty term, and RES is the resolution term. RES is the average squared difference between Φ and the event rate s . A high value of RES means the forecast tends to deviate from the event rate and towards 0 or 1. Equation (7.16) can also be obtained by applying the law of total variance. It essentially decomposes the unconditional variance of Y into the explained variance (variance of the regression function), RES, and unexplained variance, REF. With resolution characterized, we can rewrite the BS as

$$\text{BS} = \text{REL} - \text{RES} + \text{UNC} \quad (7.17)$$

Note that only REL and RES depend on the forecast.

The Brier skill score is defined with the climatological forecast being the reference. The Brier skill for the climatological forecast is simply $\text{BS}(s) = \text{Var}[Y] = \text{UNC}$. The Brier skill score is

$$\text{BSS}(Z) = \frac{\text{BS}(Z) - \text{BS}(s)}{0 - \text{BS}(s)} \quad (7.18)$$

$$= \underbrace{\frac{\text{RES}}{\text{UNC}}}_{R^2} - \frac{\text{REL}}{\text{UNC}}. \quad (7.19)$$

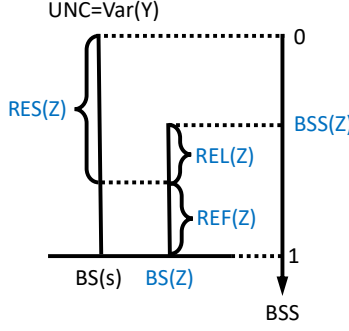


Figure 7.1: Brier score decomposition for forecast Z .

Abbreviation	Name	Definition
RES	Resolution	$\frac{1}{N} \sum_{i=1}^B N_i (\langle o_i \rangle - \bar{o})^2$
REL	Reliability	$\frac{1}{N} \sum_{i=1}^B N_i (\langle f_i \rangle - \langle o_i \rangle)^2$
ECE	Expected Calibration Error	$\frac{1}{N} \sum_{i=1}^B N_i \langle f_i \rangle - \langle o_i \rangle $
UNC	Uncertainty	$\bar{o}(1 - \bar{o})$
BS	Brier Score	$REL - RES + UNC = \frac{1}{N} \sum_{i=1}^B (\langle f_i \rangle - \langle o_i \rangle)$
BSS	Brier skill score	$\frac{BS - BS_{\text{clim}}}{0 - BS_{\text{clim}}} = \frac{RES - REL}{UNC}$

Table 7.1: Probabilistic metrics. Suppose N probability forecasts are binned into B groups. The sample size, center predicted probability, and the observed frequency of the i -th bin are denoted as N , $\langle f_i \rangle$, and $\langle o_i \rangle$, respectively. The average observed frequency (i.e., climatology) is computed as $\bar{o} = (\sum_{i=1}^B N_i \langle o_i \rangle) / N$.

where $\frac{RES}{UNC}$ represents the proportion of the variance of Y explained by Φ . We refer to this proportion as R^2 to recognize the similarity of its interpretation to the *coefficient of determination* in linear regression. A general definition of $R^2 = \text{Var}[\mathbb{E}[Y|X]] / \text{Var}[Y]$. In linear regression, the regression function $\mathbb{E}[Y|X]$ is linear w.r.t. X . Here, the only difference is that we no longer restrict the regression function $\mathbb{E}[Y|X]$ to be linear; instead we assume knowledge of underlying distribution, so $\mathbb{E}[Y|X]$ is considered as the “ground truth” regression function with squared loss. Thus, the Brier skill can be seen as R^2 with a scaled reliability cost subtracted. Figure 7.1 shows the Brier score decomposition for the forecast Z .

Table 7.1 shows the binning-based estimates for the metrics. The number of bins B is chosen as $O(n^{1/3})$ to balance the reliability and the sharpness [271].

At the end of this section, we introduce the reliability diagram, also known as the calibration curve, which is a graphical tool to verify the probabilistic forecasts. To generate a reliability diagram, we first group the probability into bins, and then plot the the observed event rate within each group against the the centers of the probability bins. On the reliability diagram, a reliable forecast has all the points close to the diagonal, and a high-resolution forecast has all the points away from the horizontal climatology line and close to 0 or 1. Notably, there is a region on the reliability diagram in which the points contributes positively to the the Brier Skill Score, or equivalently, the resolution outweighs the calibration loss. The reliability diagram is usually accompanied with a histogram showing the number of samples in each bin. The histogram can be used to evaluate the sharpness of the probabilistic forecast.

7.2.2.2 Deterministic forecast evaluation

Deterministic forecast generate a single forecast decision without a confidence score. One way to evaluate the probabilistic forecast Z is through its deterministic classification performance. To do that, we need to convert Z into discrete predictions via thresholding. For threshold ζ , the converted binary discrete prediction $\hat{Y} = \mathbf{1}_{[Z \geq \zeta]}$.

For binary forecasting problems, both the observation Y and the deterministic forecast take values as 0 or 1. A set of forecast-observation pairs $\{(y_i, \hat{y}_i)\}_{i=1}^N$ can be summarized into a contingency table shown in Table 7.2. Deterministic forecast evaluation metrics can be computed using the 4 elements a , b , c , and d in the contingency table. Only 3 elements are independent since the metrics are only determined by the joint distribution of Y and $\hat{Y}$ regardless of the total sample size. One way of choosing 3 independent elements is based on the likelihood-base rate factorization $p(\hat{y}, y) = p(\hat{y}|y)p(y)$: the event rate $s = P[Y = 1]$, the hit rate (probability of detection) $P_D = P[\hat{Y} = 1|Y = 1]$, and the false alarm rate (probability of false detection) $P_F = P[\hat{Y} = 1|Y = 0]$. Since the event rate s is out of the forecaster's control, the forecast evaluation can be reduced to a 2 dimension problem. Given a event rate $\pi_1 = s$, various measures can be plotted on a graph of P_D vs P_F . Alternatively, they can be plotted on a skill-bias diagram [144], with two axes being True Skill Statistics and frequency bias (the ratio the class-1 proportion in prediction r and the true class-1 proportion s):

$$\text{TSS} = P_D - P_F \quad (7.20)$$

$$\text{FB} = \frac{r}{s} = P_D + \frac{1-s}{s}P_F \quad (7.21)$$

	Observed		Total
	Positive	Negative	
Forecast			
Positive	a (TP)	b (FP)	$p' = a + b$
Negative	c (FN)	d (TN)	$n' = c + d$
Total	$p = a + c$	$n = b + d$	$n + p$

Table 7.2: A contingency table of $n + p$ forecast-observation pairs categorized into TP (true positive), FP (false positive), FN (false negative), and TN (true negative). The number of in each category is denoted as a, b, c, d , respectively.

7.2.3 Recalibration under event rate shift

We first characterize the label shift for a general K -class classification. We work on the same sample space Ω and event space $\mathcal{A}$, but with different distributions: a *source distribution* P from which the training data are sampled, and a *target distribution* P' from which the test data are sampled. We adopt the label shift assumption [264], which states that the class priors may change between P and P' , whereas class conditional feature distributions are kept the same:

$$\begin{aligned} \pi_k &\neq \pi'_k \\ p(x|Y = k) &= p'(x|Y = k) \end{aligned} \tag{7.22}$$

where $\pi'_k = P'[Y = k]$ denotes the class priors for the target distribution.

For class k , we have Bayes rule on both the source and the target distribution:

$$P[Y = k|x] = \frac{\pi_k}{p(x)} p(x|Y = k) \tag{7.23}$$

$$P'[Y = k|x] = \frac{\pi'_k}{p'(x)} p'(x|Y = k) \tag{7.24}$$

Noticing $p(x|Y = k) = p'(x|Y = k)$ under the label shift assumption, we have

$$P'[Y = k|x] = \frac{\pi'_k}{\pi_k} \frac{p(x)}{p'(x)} P[Y = k|x] \tag{7.25}$$

$$= \frac{w_k P[Y = k|x]}{\sum_{j=1}^K w_j P[Y = j|x]} \tag{7.26}$$

where $w_k = \frac{\pi'_k}{\pi_k}$ is known as the importance weight for class k .

In our binary classification application, k takes value 0 or 1 and π_1 can be interpreted

as the event rate. The posterior adjustment can be seen as applying a function g to the posterior, where $g : [0, 1] \rightarrow [0, 1]$ is defined as

$$g(p) = \frac{w_1 p}{w_1 p + w_0 (1 - p)} \quad (7.27)$$

For an estimate $Z(X)$ to the source posterior $P[Y = 1|X]$, we can get an estimate to the target posterior $P[Y = 1|X]$ by applying the same function g :

$$Z'(X) = g(Z(X)) \quad (7.28)$$

Then $Z'(X)$ is compared with the threshold 0.5 to make a decision. We note that the threshold 0.5 is a natural choice for many classifiers because it is based on the maximum a posteriori (MAP) decision rule: we decide the class with the larger posterior. The threshold can be changed to maximize a certain metric [29]. We review the optimal probability thresholds for accuracy, TSS, and HSS in Appendix 7.5.2.

In the remainder of this section, we theoretically analyze the effects of event-rate-based recalibration in both probabilistic and deterministic metrics. These are characterized by the following theorems.

First we notice that, under the label shift assumption, we can change the expectation under the target distribution back to the source distribution by introducing the importance weights $w(k) = \frac{\pi'_k}{\pi_k}$. This is stated as the following Lemma 7.1, which is proved in Appendix 7.5.3.1.

Lemma 7.1 (Importance reweighting). *Under label shift assumption (Equation 7.22), for any random variable X ,*

$$\mathbb{E}'[X] = \mathbb{E}[w(Y)\mathbb{E}[X|Y]]$$

Then we show that the recalibration loss on the test set is reduced by applying g to the probabilistic forecast Z . This is stated as the following theorem and proved in Appendix 7.5.3.2.

Theorem 7.2 (The recalibration g). *Suppose*

$$\mathbb{E}'[(\Phi - Z)^2] \leq \frac{(w_1 - w_0)^2}{4 \max\left(\frac{w_0^2}{w_1}, \frac{w_1^2}{w_0}\right)} \mathbb{E}'[(Z(1 - Z))^2] ,$$

then

$$\mathbb{E}'[(g(\Phi) - g(Z))^2] \leq \mathbb{E}'[(g(\Phi) - Z)^2]$$

Next we show the conditional forecast is not changed by g . Denote $\Phi'_g = \mathbb{E}[Y|g(Z)]$, we have the following theorem:

Theorem 7.3 (The recalibration does not change the conditional forecast).

$$\Phi'_g = \Phi'$$

This is because g is an invertible function so the sigma-algebra induced by $g(Z)$ and Z is the same. Therefore, $\Phi'_g = \mathbb{E}[Y|g(Z)] = \mathbb{E}[Y|Z] = \Phi'$.

A natural consequence for Theorem 7.3 is that the refinement loss, as the variance of Φ' , is not changed by the recalibration g .

Corollary 7.4 (The recalibration does not change the refinement loss).

$$REF(g(Z)) = REF(Z)$$

Since BS is the sum of REL and REF, and BSS is the proportion of decreased BS relative to BS of the climatology, we have the following corollary:

Corollary 7.5 (Recalibration g decreases BS and increases BSS).

$$\begin{aligned} BS(g(Z)) &\leq BS(Z) \\ BSS(g(Z)) &\geq BSS(Z) \end{aligned}$$

As for deterministic metrics, the maximum value by varying the threshold (i.e., hedging) is not changed. Therefore, the recalibration may change the deterministic metrics, but does not change the maximum values.

However, hedging is not always desirable as argued by many researchers. Often used is the threshold 0.5, which corresponds to the minimum classification error, or equivalently, maximum a-posteriori (MAP) decision rule. The optimal thresholds for different metrics and how the recalibration g changes those metrics under the MAP rule are discussed in the following section.

7.3 Experiments

7.3.1 Data preparation

To construct features of solar active regions, we use Space-Weather HMI Active Region Patches [SHARPs, 31], a data product derived from full-disk observations of the Helioseismic

Parameter	Definition	Unit
AREA_ACR	De-projected area of active pixels on sphere in micro-hemisphere	mH
CRVAL1	Longitude at center of patch	degree
CRVAL2	Latitude at center of patch	degree
TOTUSJH	Total unsigned current helicity	G ² /m
TOTUSJZ	Total unsigned vertical current	Amperes
SAVNCP	Sum of the Absolute Value of the Net Currents Per Polarity (SAVNCP)	Amperes
USFLUX	Total unsigned flux	Maxwell
ABSJZH	Absolute value of the net current helicity	G ² /m
TOTPOT	Total photospheric magnetic energy density	Erg/cm ³
MEANPOT	Mean photospheric excess magnetic energy density	Erg/cm ³
SHRGT45	Area with shear angle greater than 45 as a percent of total area	Percentage
R.VALUE	Unsigned Flux R [252]	Maxwell
MEANJZH	Mean current helicity	G ² /m
MEANSHR	Mean shear angle for B.total	Degrees
MEANJZD	Mean vertical current density	mA/m ²
MEANALP	Mean twist parameter, alpha	1/Mm
MEANGBT	Mean value of the total field gradient	Gauss/Mm
MEANGAM	Mean inclination angle, gamma	Degrees
MEANGBZ	Mean value of the vertical field gradient	Gauss/Mm
MEANGBH	Mean value of the horizontal field gradient	Gauss/Mm

Table 7.3: SHARP summary parameters that are used as active region features in this chapter. See Bobra and Couvidat [30] for details.

and Magnetic Imager [HMI, 251] aboard the *Solar Dynamics Observatory* (SDO). SHARPs contain maps and summary parameters of automatically tracked active regions, called HMI Active Region Patches (HARPs), from May 2010 to the present day. Table 7.3 shows the 20 summary parameters in the definitive SHARP records in Cylindrical Equal Area (CEA) coordinates that are used as active region features in this chapter. We consider HARPs associated with at least one NOAA active region number in a 12-year time window between May 1, 2010 and May 1, 2022. Only good quality summary parameters when the active region is within $\pm 70^\circ$ of the central meridian are queried.

Observational features are labeled using the *Geostationary Operational Environmental Satellites* (GOES) catalog of X-ray solar flare events. Based on the peak magnitude of 1–8 Å soft X-ray flux, solar flare events are classified into five increasingly intense classes: A, B, C, M, and X. We only consider GOES solar flare events with at least one associated NOAA active region that can be used to cross-reference the NOAA_ARS keyword in SHARP databases to associate the flare with a HARP.

We follow the procedure in Sun et al. [270] to extract feature-label pairs from the SHARP

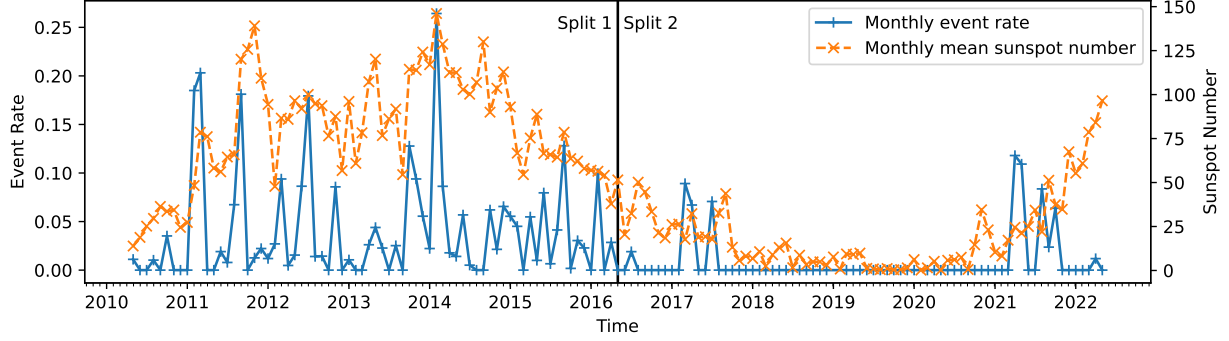


Figure 7.2: Monthly event rates vs mean sun spot numbers from May 2010 to May 2022. The sun spot numbers is an indicator of the solar activity and the phase in a solar cycle. The event rate correlates well with the trend of the sunspot number. Source of sunspot number data: WDC-SILSO, Royal Observatory of Belgium, Brussels.

	Period	Number of HARPs	Sample size	Event rate
Split 1	2010.5–2016.5	979	444055	0.041303
Split 2	2016.5–2022.5	332	158604	0.008625

Table 7.4: Statistics of the two periods.

database and the GOES flare event catalog. Specifically, we consider the prediction task of “strong-versus-quiet” flare prediction. For an active region, we sample the SHARP parameters at a 96-min interval. We label a sample as positive if the active region produces an M- or X-class flare in the next 24 hours, and as negative if there is no flare of any class in ± 24 hours. Samples are not labeled are considered irrelevant and discarded for our task.

The event rate can be higher or lower on the test set than the training set. To investigate how the proposed recalibration method works in both cases, we split the 12-year data into two 6-year periods, and perform two sets of experiments—training one set of models on either split and testing on the other. The former split (2010.5–2016.6), referred to as Split 1, covers the solar maximum for Solar Cycle 24, whereas the latter split (2016.6–2022.6), referred to as Split 2, covers the solar minimum of Solar Cycle 24 and the beginning of Solar Cycle 25 (Figure 7.2). The strong flare event rate is 0.041 on Split 1, about 4.8 times higher than on Split 2 (Table 7.4).

7.3.2 Experiment setup

In this section, we show how flare prediction performance of Logistic Regression (Logistic), Quadratic Discriminant Analysis (QDA), Decision Tree (DT), Random Forest (RF),

Histogram-based Gradient Boosting Trees (HGBT), and Multilayer Perceptrons (MLP) can be improved by applying our label shift recalibration method. We also include a climatology baseline, “Clim”, which always predicts the prior probabilities and decides the most frequent class in the training set without exploiting the features. A brief review of the algorithms are given in Appendix 7.5.1.

We consider two types of training set: the original “unbalanced” data without resampling, and the “rebalanced” data with the majority class undersampled until the sample sizes of the two classes are equalized. For the “rebalanced” training data, the event rate is 1. When the training data is “unbalanced”, we handle the class imbalance if there is a well-established way (e.g., cost-sensitive learning, reweighting samples, etc.) For Logistic regression and MLP, we assign weights to the loss terms of the two classes in the cross entropy loss function. The weights are set to be inversely proportional to the sample size³. Similarly for Decision Trees, we reweight the samples so that they are inversely proportional to the sample size. The weights for the samples are used in the computation of the impurity measure. For QDA, the class priors are estimated as the class proportions in the training set. For HGBT, we do not perform any treatment to the class imbalance as `scikit-learn` does not offer relevant options and there is no standard algorithmic modification to handle class imbalance for HGBT.

We tuned the hyperparameters using Bayesian optimization [260]. The Bayesian optimization is performed 10 steps. At each step, Bayes optimization queries 4 sets of parameters to run in parallel. The hyperparameter configuration that achieves the highest mean HSS produced by 5-fold cross-validation is selected. The model is then refit on the entire training set.

When presenting of the evaluation results, we use $1 - \text{REL}$ as opposed to REL so that all metrics has positive orientation, i.e., a higher value in the metric indicates closer association between forecasts and the observations. We include two the hedged metrics for TSS and HSS, denoted as `TSS_MAX` and `HSS_MAX`, respectively.

7.3.3 Recalibration performance on Split 2

We train the machine learning algorithms on Split 1 (2010.5–2016.5) and evaluate them on Split 2 (2016.5–2022.5). The event rate is much smaller on the test set than the training set. The calibration function (Equation 7.26) for the unbalanced and rebalanced training set is shown in Figure 7.3. Both unbalanced and rebalanced training set has an event rate smaller than the test set. The classifiers trained on those datasets are likely to overforecast

³In `scikit-learn` [222], this is achieved by setting `class_weights` to be `balanced`

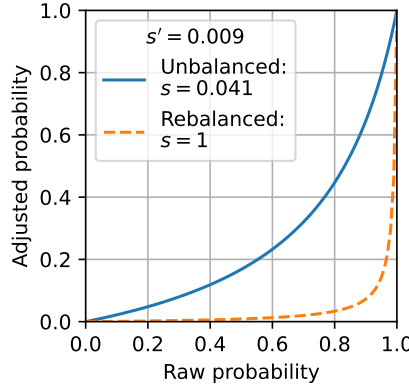


Figure 7.3: The probability calibration on the test set (Split 2) with event rate $s' = 0.009$. The blue solid (oranged dashed) curve shows the recalibration for classifiers trained on unbalanced (rebalanced) training data where $s = 0.041$ ($s = 1$).

Metric		ACC		TSS		HSS		RES		1-REL		BSS	
		Adj	p-val	Adj	p-val	Adj	p-val	Adj	p-val	Adj	p-val	Adj	p-val
Untreated	Logistic	0.990	0.006	0.408	0.984	0.401	0.139	0.002	0.124	0.996	0.009	0.001	0.008
	QDA	0.964	0.001	0.649	1.000	0.234	0.001	0.001	0.001	0.974	0.001	0.001	0.001
	DT	0.988	0.002	0.271	0.985	0.275	0.002	0.001	0.997	0.995	0.001	0.001	0.001
	RF	0.991	0.001	0.238	1.000	0.316	0.999	0.002	0.005	0.996	0.001	0.103	0.001
	HGBT	0.989	0.001	0.201	0.999	0.244	0.998	0.002	0.007	0.996	0.001	0.001	0.001
	MLP	0.984	0.006	0.393	0.972	0.298	0.012	0.001	0.007	0.992	0.001	0.001	0.001
Reweighted	Logistic	0.978	0.010	0.513	0.998	0.288	0.003	0.001	0.068	0.987	0.010	0.001	0.010
	QDA	0.964	0.001	0.649	1.000	0.234	0.001	0.001	0.001	0.974	0.001	0.001	0.001
	DT	0.979	0.035	0.355	0.870	0.222	0.005	0.001	0.007	0.988	0.009	0.001	0.008
	RF	0.991	0.001	0.217	1.000	0.281	0.995	0.002	0.001	0.996	0.001	0.084	0.001
	HGBT	0.989	0.001	0.476	0.930	0.415	0.015	0.002	0.002	0.995	0.001	0.001	0.001
	MLP	0.984	0.006	0.393	0.972	0.298	0.012	0.001	0.007	0.992	0.001	0.001	0.001
Rebalanced	Logistic	0.988	0.001	0.415	1.000	0.367	0.001	0.001	0.001	0.995	0.001	0.001	0.001
	QDA	0.957	0.001	0.691	1.000	0.216	0.001	0.001	0.001	0.968	0.001	0.001	0.001
	DT	0.986	0.001	0.231	1.000	0.218	0.011	0.001	0.928	0.993	0.001	0.001	0.001
	RF	0.990	0.001	0.020	1.000	0.032	1.000	0.002	0.295	0.996	0.001	0.001	0.001
	HGBT	0.990	0.001	0.004	1.000	0.006	1.000	0.002	0.013	0.997	0.001	0.103	0.001
	MLP	0.983	0.001	0.444	0.961	0.312	0.005	0.001	0.015	0.990	0.001	0.001	0.001

Table 7.5: Mean adjusted (Adj) performance on Split 2 and p -values from one-tailed paired t -tests for significant improvement over the uncalibrated models. For each imbalance treatment and each machine learning algorithm, 5 models are independently tuned, trained and tested, producing 5 metric values computed from the uncalibrated predicted probabilities, and 5 metric values computed from the recalibrated probabilities. A one-tailed paired t -test is performed to test if the recalibrated metric is significantly larger than the raw metric. p -value smaller than 0.05 is treated as significant. p -values smaller than 0.001 is reported as 0.001.

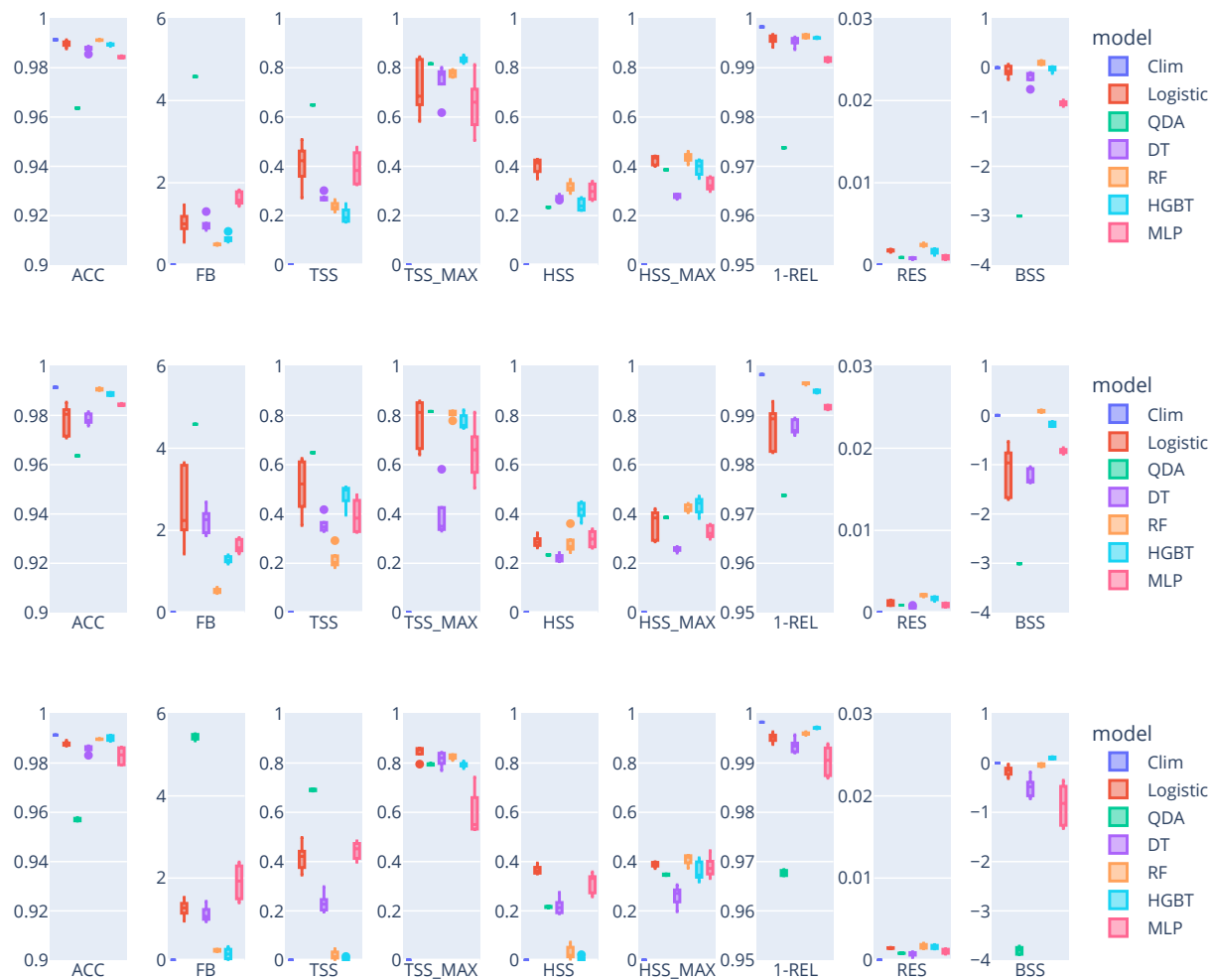


Figure 7.4: Recalibrated performance of models on Split 2 (2016.5 – 2022.5) with different imbalance treatment in training: no treatment (up), loss is reweighted (middle), and data is rebalanced (bottom).

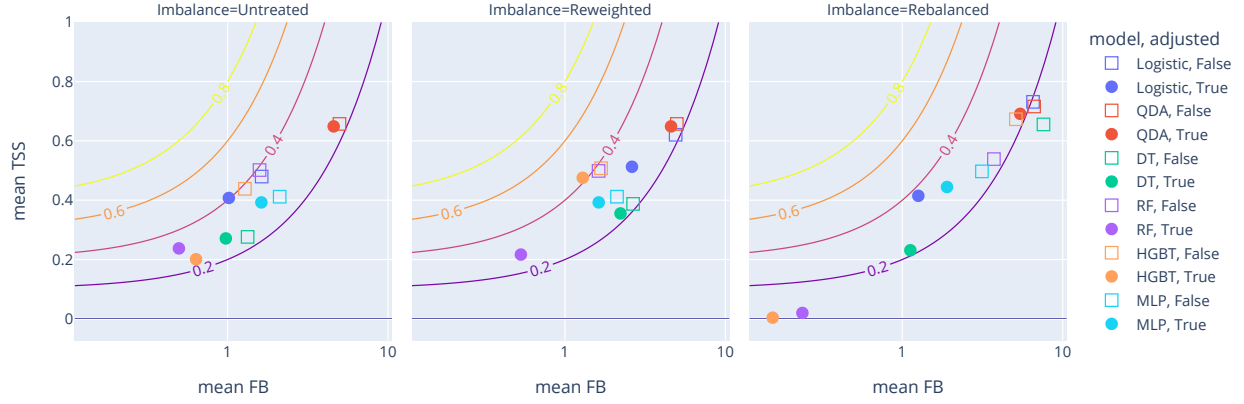


Figure 7.5: Skill bias diagram. The mean TSS and mean FB on Split 2 are plotted for each algorithm with unadjusted predictions (shown as open squares) and recalibrated predictions (shown as solid circles). The contours show the HSS on the skill-bias diagram with the Split 2 event rate $s = 0.009$. The left panel corresponds to models trained on the unbalanced data, whereas the right panel corresponds to models trained on the rebalanced data. The FB is displayed in log scale.

on the test set where the positive event is less frequent. The bias correction scheme shrinks the probability to correct the overforecast. The shrinkage can be drastic for the rebalanced training data which induces a significant shift of the event rate.

Figure 7.4 shows the recalibrated performance of different machine learning models trained on the unbalanced (panel a) and rebalanced datasets (panel b). It is observed different methods can perform differently and there is no method that consistently outperform others in all metrics. The baseline “Clim” has the highest accuracy but virtually no skills. The HGBT performs similar to “Clim” in terms of accuracy and reliability, but has positive TSS. In fact, the TSS_MAX achieved by HGBT is actually among the highest in all methods. Logistic regression and QDA also has good TSS and TSS_MAX. The HSS is generally low, with the best HSS in all experiments barely exceeds 0.4. The reliability of QDA is the lowest in all methods, which is because the posteriors are estimated with the samples in the predicted leaf node. When the sample size in the leaf node is small, the posterior estimates has large uncertainties. The low reliability of the Decision Tree contributes to its low BSS. The HGBT is an ensemble of decision trees and the reliability is significantly improved.

By comparing the two panels of Figure 7.4, we found no significant difference between using unbalanced and using rebalanced training data. This shows random undersampling is perhaps preferred because it does not decrease the generalization performance while significantly reducing the training data and time.

Table 7.5 shows the one-sided t -test results for the improvement of the recalibration

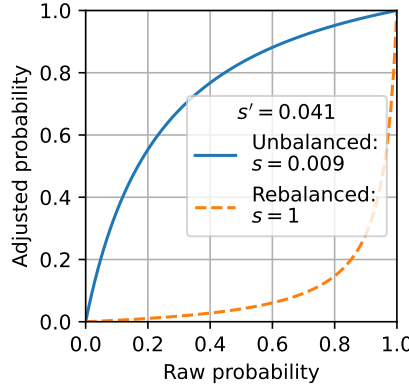


Figure 7.6: The probability calibration on the test set (Split 1) with event rate $s' = 0.041$. The blue solid (orange dashed) curve shows the recalibration for classifiers trained on unbalanced (rebalanced) training data where $s = 0.009$ ($s = 1$).

method in different metrics. It is observed that the BSS are improved in all cases and achieves statistical significance in many cases. Other metrics are improved in general except for TSS. The TSS performance decreased because the TSS achieves the maximum when the probability threshold is equal to the event rate which is much smaller than 0.5. At the threshold 0.5, the trend of TSS is decreasing as the threshold increases. With the bias correction, the predicted probability gets shrunk further towards 0, so the TSS at 0.5 will be smaller.

Figure 7.5 shows the skill bias diagram of the unadjusted and recalibrated predictions for each algorithm. It can be seen that the points corresponding to the recalibrated forecasts all shifted left in both panels, more so for the rebalanced training data than the unbalanced training data. For the unbalanced training data, the bias is somewhat corrected for QDA, MLP, and DT, and over-corrected for Logistic regression. HGBT is underforecasting before and after the correction. The QDA, MLP, and DT are observed to have higher HSS after recalibration. For the rebalanced training data, all points are more close to the $FB = 1$ unbiased line after recalibration. The TSS is decreased after the recalibration for all methods, whereas all HSS increase except for HGBT.

7.3.4 Recalibration performance on Split 1

Figure 7.6 shows the probability adjustment if we are training on Split 2 and testing on Split 1. If the training data is rebalanced, the event rate on the test set will decrease, and the predicted probability will be shrunk just as in the previous section. If the training data is not rebalanced, the event rate on the test set will increase, and the predicted probability

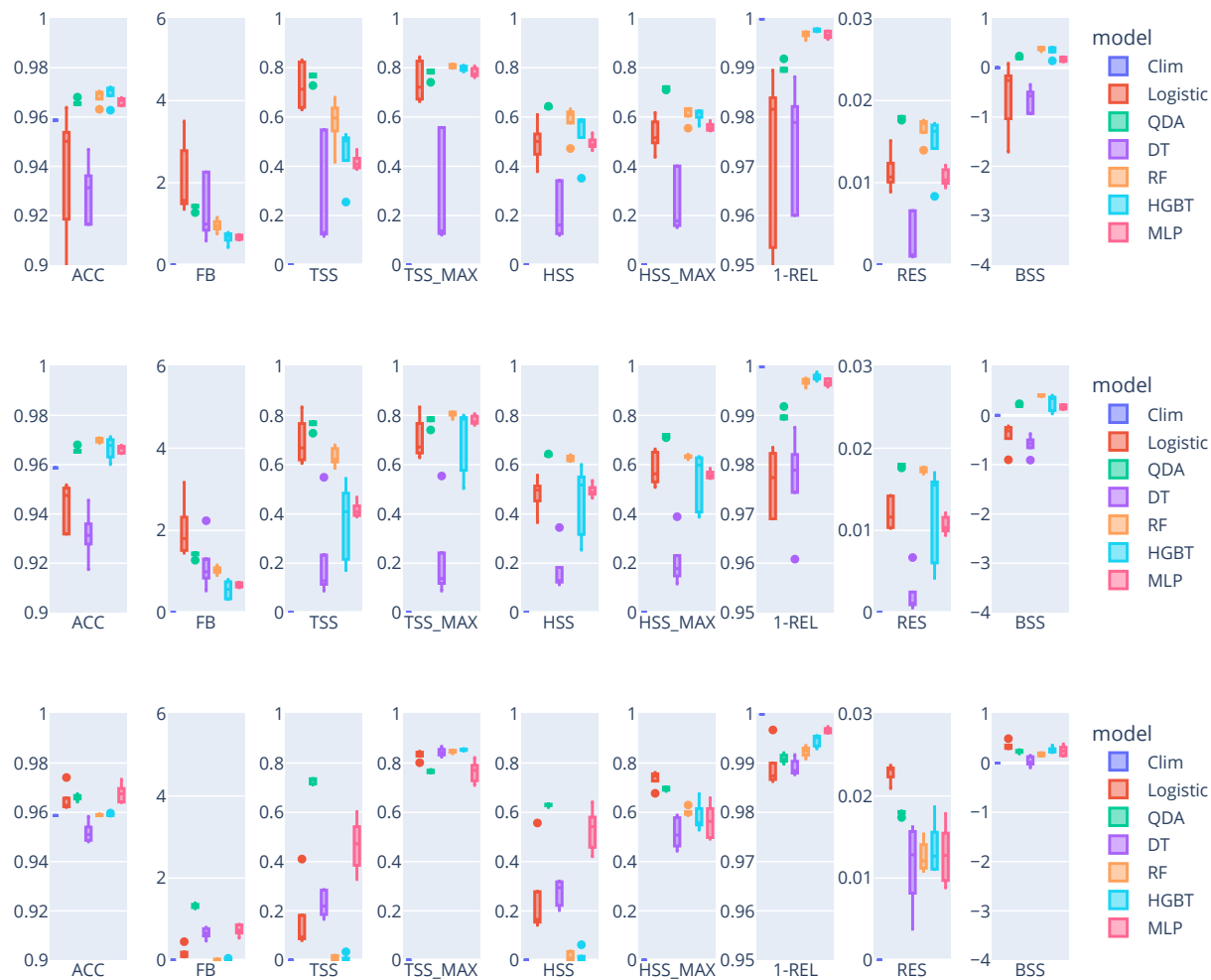


Figure 7.7: Recalibrated performance of models on Split 1 (2010.5 – 2016.5) with different imbalance treatment in training: no treatment (up), loss is reweighted (middle), and data is rebalanced (bottom).

Metric		ACC		TSS		HSS		RES		1-REL		BSS	
		Adj	p-val	Adj	p-val	Adj	p-val	Adj	p-val	Adj	p-val	Adj	p-val
Imbalance Model treat- ment	Logistic	0.937	0.942	0.727	0.675	0.493	0.945	0.011	0.942	0.969	0.939	-0.587	0.940
	QDA	0.966	1.000	0.765	0.001	0.638	1.000	0.018	1.000	0.990	1.000	0.196	1.000
	DT	0.929	0.992	0.294	0.992	0.219	0.990	0.003	-	0.973	0.992	-0.663	0.990
	RF	0.968	0.001	0.580	0.001	0.589	0.001	0.017	0.020	0.997	0.005	0.386	0.003
	HGBT	0.970	0.063	0.456	0.004	0.537	0.005	0.015	0.001	0.998	0.437	0.346	0.009
	MLP	0.966	0.879	0.416	0.010	0.494	0.015	0.011	0.003	0.997	1.000	0.174	0.660
Reweighed	Logistic	0.938	0.847	0.695	0.173	0.480	0.861	0.012	0.873	0.973	0.850	-0.402	0.855
	QDA	0.966	1.000	0.765	0.001	0.638	1.000	0.018	1.000	0.990	1.000	0.196	1.000
	DT	0.932	0.992	0.203	0.992	0.169	0.991	0.002	-	0.977	0.992	-0.592	0.992
	RF	0.970	0.001	0.634	0.001	0.625	0.001	0.017	0.001	0.997	0.004	0.405	0.002
	HGBT	0.967	0.113	0.364	0.030	0.448	0.035	0.012	0.039	0.998	0.206	0.255	0.066
	MLP	0.966	0.879	0.416	0.010	0.494	0.015	0.011	0.003	0.997	1.000	0.174	0.660
Rebalanced	Logistic	0.965	0.001	0.156	1.000	0.242	0.781	0.023	0.996	0.989	0.001	0.341	0.001
	QDA	0.966	0.001	0.727	1.000	0.629	0.001	0.018	0.001	0.991	0.001	0.234	0.001
	DT	0.952	0.002	0.230	0.999	0.273	0.981	0.012	0.863	0.989	0.001	0.040	0.001
	RF	0.959	0.005	0.009	1.000	0.016	1.000	0.013	0.999	0.992	0.001	0.178	0.007
	HGBT	0.959	0.001	0.007	1.000	0.013	1.000	0.014	0.977	0.995	0.001	0.262	0.001
	MLP	0.968	0.009	0.466	0.997	0.527	0.995	0.013	0.992	0.997	0.007	0.245	0.015

Table 7.6: Similar to Table 7.5, except that the models are trained on Split 2 and tested on Split 1.

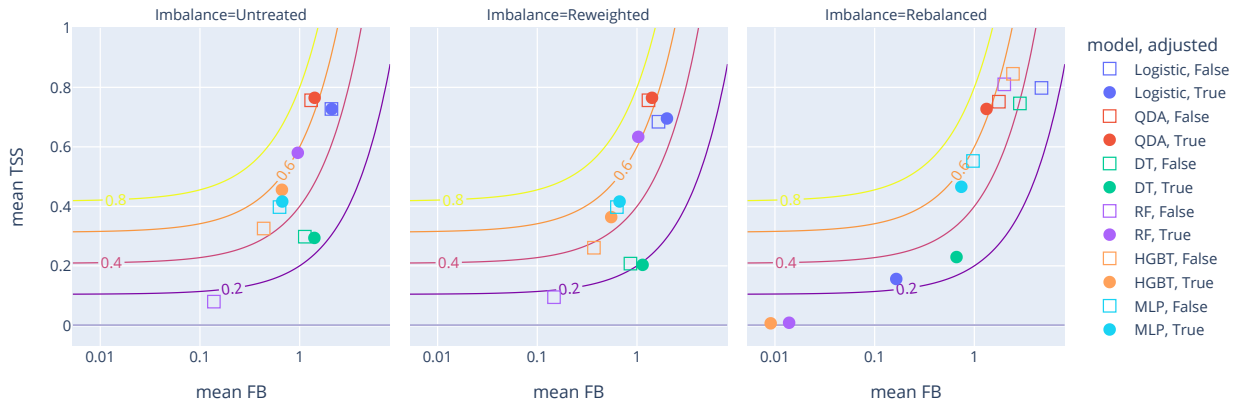


Figure 7.8: Skill bias diagram similar to Figure 7.5, except that models are trained on Split 2 and tested on Split 1.

will be shifted towards 1.

Figure 7.7 shows the recalibrated performance on Split 2. Compared to Figure 7.4, BSS for all models except for DT is positive due to increased resolution. The maximum HSS increases, and HSS increases for most models. The models are less biased except for HGBT according to FB. The accuracy for all models except for DT has better accuracy than the climatology baseline.

Table 7.6 shows the one-sided t -test results for the improvement of the recalibration method in different metrics. It is observed that the BSS are improved in all cases and achieves statistical significance in many cases. Other metrics are improved in general except for TSS. The TSS performance decreased because the TSS achieves the maximum when the probability threshold is equal to the event rate which is much smaller than 0.5. At the threshold 0.5, the trend of TSS is decreasing as the threshold increases. With the bias correction, the predicted probability gets shrunk further towards 0, so the TSS at 0.5 will be smaller.

Figure 7.8 shows the skill bias diagram for forecasts trained on Split 2 and tested on Split 1. For the unbalanced training data, almost all methods are underforecasting before the recalibration and this trend is corrected by the recalibration. We also see a universal increase of TSS and HSS for all methods. For rebalanced data, points are moved from right of the unbiased line $FB = 1$ towards left, with HGBT getting overcorrected.

7.3.5 Case study: Logistic regression

Logistic regression is one of the most easily interpretable machine learning models. In this section, we perform a case study of a logistic regression model trained on the rebalanced Split 1 and tested on Split 2. The best model selected by the Bayes optimization has an L1 penalty with coefficient $\alpha = 5.7 \times 10^{-5}$.

Since the recalibration is a post-processing procedure and does not change the model itself, we first inspect the logistic model. In logistic regression, the coefficients can be interpreted as how much the logit score will change if that parameter changes one unit. When the features are normalized, the magnitude indicates the significance of the parameter on the prediction and the sign indicates whether the correlation is positive or negative. Table 7.7 shows the logistic regression coefficients (include the intercept) ordered by their absolute value. We can see that TOTUSJH is regarded as the most important feature by this model. The higher the value of TOTUSJH, the more likely the active region will produce a strong flare. The AREA is the second important feature, negatively correlated with flare probability. A possible reason is that the flux density could be lower for large active regions if the total amount of

Parameter	Coefficient	Parameter	Coefficient	Parameter	Coefficient
TOTUSJH	13.1504	SAVNCPP	1.5245	SHRGT45	0.0000
AREA_ACR	-4.2948	USFLUX	-0.5075	TOTUSJZ	0.0000
TOTPOT	-2.3057	MEANPOT	-0.3466	MEANJZH	0.0000
R_VALUE	2.1356	MEANGBZ	0.1475	CRVAL2	0.0000
(Intercept)	2.0977	ABSNJZH	0.1371	MEANALP	0.0000
MEANGAM	1.8897	CRVAL1	-0.1367	MEANGBT	0.0000
MEANSHR	-1.5525	MEANJZD	-0.0808	MEANGBH	0.0000

Table 7.7: Logistic regression coefficients (include the intercept) ordered by absolute value from top to bottom and then from left to right.

	FB	ACC	TSS	TSS_MAX	HSS	HSS_MAX	RES	1-REL	BSS
Raw	6.1579	0.9511	0.6963	0.8506	0.1958	0.3994	0.0011	0.9673	-3.5729
Recalibrated	1.2719	0.9878	0.4210	0.8506	0.3710	0.3994	0.0015	0.9953	-0.1480

Table 7.8: Metrics of the selected Logistic regression model. The bold fonts shows the value is better (closer to 1 for FB and higher for other metrics).

unsigned flux is fixed. The sparsity inducing L1 regularization has set the the coefficients in the last column of Table 7.7 are all 0, indicating they are not so informative of the strong flares.

To study the effect of recalibration, we first collect the probability directly predicted by the model and refer to it as the “raw” probability. Then, we correct the raw probability using the proposed recalibration method, and produce the “recalibrated” probability. The two sets of probabilities are used to make predictions with a threshold 0.5. Both the probabilistic and the deterministic aspects of the forecast are examined. Note that since we are training on the higher event rate Split 1 and testing on the lower event rate Split 2, the recalibration essentially shrinks the posterior, the same way as the orange dashed line in Figure 7.3.

Table 7.8 shows the deterministic and probabilistic metrics for the selected logistic regression model. We can see that recalibration improved the probabilistic metrics (RES, 1-REL, and BSS). The improvement in 1-REL and BSS is justified by the theorem, whereas the improvement in RES ostensibly violates the theorems. We argue that RES increases due to the binning effect in the computation, which will be explained later using the reliability diagram. In terms of deterministic metrics, the tendency to overforecast is mostly corrected, as shown by the frequency bias decreasing from 6.16 to 1.27. The ACC and HSS increases, whereas TSS decreases. However, the maximum values of those metrics are left unchanged by the label shift correction.

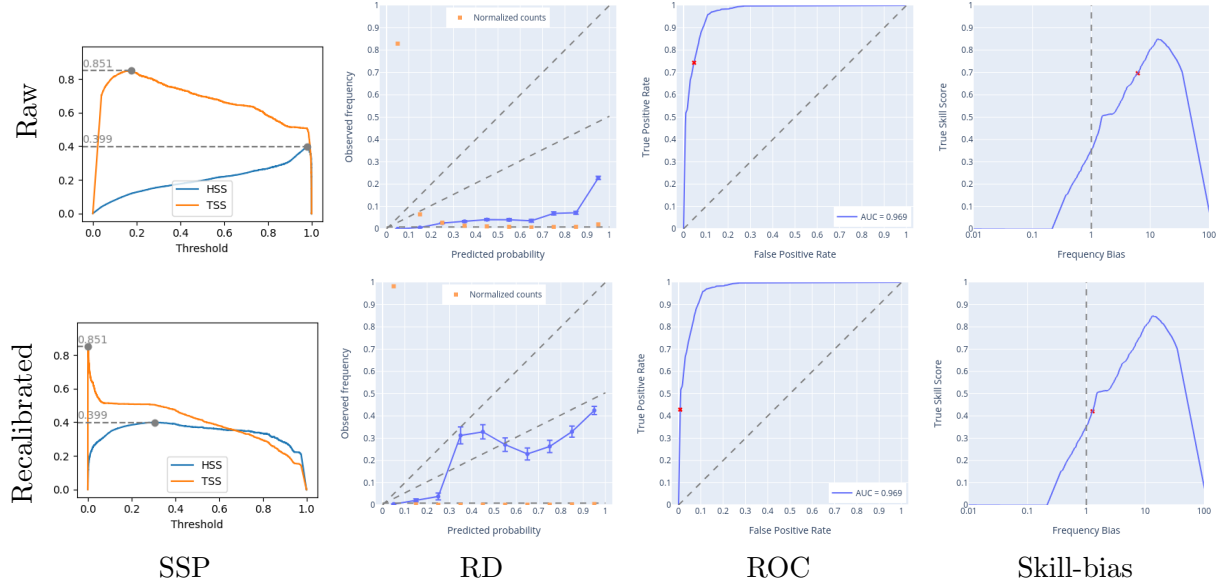


Figure 7.9: Graphical tools for a logistic regression model trained with rebalanced data and tested on Split 2. The first row shows skill score profiles (SSP), reliability diagram (RD), ROC, and Skill-bias diagrams for the raw predicted probabilities. The second row shows the same figures for the recalibrated probabilities. The red cross on the ROC and Skill-bias corresponds to the 0.5 threshold.

Figure 7.9 shows the graphical evaluation of the raw and the recalibrated logistic regression model. The effect of recalibration can be visualized by those graphs.

The skill score profiles (SSP) shows how TSS and HSS changes with the probability threshold. The recalibration essentially moves the points on the skill score profile along the horizontal axis towards 0, so the maximum values of TSS and HSS are not changed. To see why TSS decreases with recalibration in this case. Let us assume the recalibrated posterior is close to the true posterior, then the empirical TSS curve in the recalibrated SSP is closed to the population curve. Therefore, the empirical TSS is also maximized at around the event rate, which is very close to 0. Consequently, the threshold 0.5 lies in the decreasing interval of TSS profile. The uncalibrated TSS profile are moved horizontally towards right, hence having a larger TSS at threshold 0.5. As we will see later, the decrease in TSS is general if the recalibration shrinks the predicted probability. Conversely, TSS tends to increase if the recalibration inflates the predicted probability.

The reliability diagrams (RD) also have predicted probabilities shrunk towards 0. Since the points of the original model are below the diagonal (a sign of overforecasting), moving them left makes the curve closer to the diagonal, indicating better calibration. Of note, due to the binning error introduced in estimating $P[Y|Z]$, the points are not moving *precisely* along the horizontal axis. The binning error also affect the computation of RES, the weighted

squared deviation of the points from the horizontal climatology line. This is why RES is seemingly improving in Table 7.8. Later in the text, we will see RES is increased after recalibration in many cases. This shows the binning effect introduces a bias that increases the theoretical version of resolution.

Both the ROC and the skill-bias diagram are not changed in shape. In fact, the points are moving along the curves, as indicated by the operating point at 0.5 threshold. The Skill bias diagram shows the recalibration moved the operating point towards the unbiased line, at a price of decreasing the “true” skill as indicated by TSS.

7.4 Conclusions and discussion

In this chapter, we introduced an event rate based calibration method to solar flare forecasting. Experiments of multiple machine learning algorithms showed that the calibration method can generally make solar flare forecasts more calibrated, less biased, and more skillful in terms of HSS. The TSS is generally decreased and we discussed the reasons that lead to this decrease.

For the future work, we consider the case where the event rate is unknown. In this work, we have assumed knowledge of the test set class proportions. However, when the machine learning models is deployed to the real world, we will not have this information. One possibility is to use the EM algorithm can be used to infer the unknown event rate under the label shift [248]. The estimating may require many unlabeled test samples [184]. However, there are only a few active regions on the sun at the same time, and those active regions comprise of the unlabeled test set. Another possibility is to model the event rate as a continuously changing process and predict the current event rate [194].

7.5 Appendix

7.5.1 Classification algorithms

This section gives a brief introduction to the classification algorithms used in this chapter. Since the recalibration method deals with posterior probability, we focus on the algorithms that can produce an estimate of the posterior probability.

7.5.1.1 Quadratic Discriminant Analysis

Quadratic Discriminant Analysis (QDA) is a generative model that assumes binary labels are from a Bernoulli distribution with success probability π_1 (also known as the class prior

for the positive class) and the class conditional probability are Gaussians with mean μ_k and covariance Σ_k for class k :

$$Y \sim \text{Ber}(\pi_1), \quad (7.29)$$

$$X|Y = k \sim N_p(\mu_k, \Sigma_k), \quad k = 1, 2 \quad (7.30)$$

A sample x is classified into class 1 if and only if

$$\log \frac{p(y = 1|x)}{p(y = 0|x)} = \delta_1(x) - \delta_0(x) > 0, \quad (7.31)$$

where

$$\delta_k(x) = \log \pi_k - \frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k), \quad k = 1, 2 \quad (7.32)$$

are known as discriminant functions. The decision boundary turns out to be quadratic in x .

QDA estimate the class priors π_k with the class proportion in the training data. In flare forecasting where there are much more negative samples than the positive samples, we have $\pi_1 \ll \pi_0$.

7.5.1.2 Logistic regression

Logistic regression is a linear model for classification. In the binary case where the target $y \in \{0, 1\}$, logistic regression models the log odds as a linear function with respect to the features $x \in \mathbb{R}^p$:

$$f(x) = \beta^T x + \beta_0 \quad (7.33)$$

where $\beta \in \mathbb{R}^p$ and $\beta_0 \in \mathbb{R}$ are linear coefficients. The probability that x belongs to class 1 is then expressed as $p(y = 1|x) = \sigma(f(x))$, where $\sigma(z) := (1 + \exp(-z))^{-1}$ is the sigmoid function. The class-weighted logistic loss for (x_i, y_i) is expressed as

$$\ell(y_i, x_i) = -w_1 y_i \log \sigma(f(x_i)) - w_0 (1 - y_i) \log(1 - \sigma(f(x_i))), \quad (7.34)$$

where w_0 and w_1 are pre-specified weights associated with the loss for class 0 and class 1, respectively. To tackle class imbalance, the class weights w_0 and w_1 can be set to be inversely proportional to class frequencies.

We use the scikit-learn implementation of `SGDClassifier` with logistic loss and regular-

ization. The regularized loss function on the training set $\{x_i, y_i\}_{i=1}^n$ is

$$L(\beta, \beta_0) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, x_i) + \alpha R(\beta) \quad (7.35)$$

where R can be the L1 norm, L2 norm, or Elastic Net. This loss function is convex and can be minimized using gradient descent or Newton's method.

7.5.1.3 Decision tree classifiers

A decision tree classifier classifies an instance using a sequence of decision rules. The decision rules are structured as a binary tree, where each node is a decision rule associated with a variable in the feature space. The decision tree partitions the feature space into hypercubes and assign a class to each hypercube. We define the impurity of a node, denoted as H , as a measure of heterogeneity of the labels in the node. For example, H can be chosen to be the Gini index. In the binary case, the Gini index for a node is expressed as

$$H = 2p(1 - p) \quad (7.36)$$

where p is the proportion of the positive samples in the node. To grow a decision tree, we first assign all the training samples to the root node, and then recursively split the samples into child nodes in a way that minimizes the weighted sum of impurity in the child nodes:

$$\theta_{\text{parent}}^* = \arg \min_{\theta} \frac{n_{\text{left}}(\theta)}{n_{\text{parent}}} H_{\text{left}}(\theta) + \frac{n_{\text{right}}(\theta)}{n_{\text{parent}}} H_{\text{right}}(\theta), \quad (7.37)$$

where θ parameterize the way we split the parent node, and n denotes the sample size. If the training data have class imbalance, we can set class weights to adjust the count for each class. In this chapter, set the class weights inversely proportional to class frequencies in the training data.

7.5.1.4 Random Forest

Random forest [RF; 36] is an ensemble classifier that fits multiple decision trees on bootstrap samples with random subsets of predictor variables and averages the outputs of individual decision trees. The randomness from sampling data points and predictor variables reduces the variance of the ensemble.

7.5.1.5 Histogram-based Gradient Boosting Classification Tree

Gradient Boosted Decision Trees [85] is an additive model built in a greedy forward stage-wise fashion. Histogram-based Gradient Boosting Classification Tree (HGBT) implemented in Scikit-learn [222] is inspired by LightGBM [148]. It significantly improves the runtime when the sample size is large.

7.5.1.6 Multilayer perceptron

Multilayer perceptron (MLP) is a simple neural network formed by stacking parameterized linear layers appended with nonlinear functions called activation functions. Commonly used activation functions include Rectified Linear Unit (ReLU) and hyperbolic tangent (tanh). A loss function takes the output of the MLP and the ground truth labels and computes a scalar loss. The trainable parameters are updated via stochastic gradient descent or its variants [e.g., Adam; 151], with gradients computed via back-propagation [247]. An appropriately parameterized MLP is able to represent a wide variety of nonlinear functions according to the universal approximation theorem [132].

7.5.2 Optimal probability thresholds

A probabilistic forecast $Z(X)$ of a binary event can be converted to a deterministic forecast by applying a probability threshold $\eta \in [0, 1]$: $\hat{Y} = 1_{[Z(X) \geq \eta]}$. When we change the threshold η , the distribution of the contingency table is also changed, leading to different metric values. We say a probabilistic threshold η^* is *optimal* for a specific metric, if the metric achieves the best value at that threshold.

In that case, the elements in the contingency tables and the metrics deterministic metrics are functions of the probability threshold. Given the cost of each observation-decision pair, an optimal threshold can be derived to minimize the expected cost (also known as the Bayes risk), or equivalently, maximize a performance measure that is related to the Bayes risk.

To derive the optimal probability thresholds using the Bayes detection theory, we first define some notations. Denote the cost matrix as C , with $C_{i,j}$ indicating the cost of deciding class i when the true class is j . It is reasonable to assume a detection error yields higher cost than a correct detection, i.e., $C_{i,j} \geq C_{j,j}$ for all i and j . Denote the decision region for class i as $R_i = \{x : \hat{y} = i\}$. Let $\pi_i = p(y = i)$ denote the priors, $f_i(x) = p(x|y = i)$ denote the likelihood, and $\bar{p}_i = p(y = i|x) = \mathbb{E}[y|x]$ denote the posteriors. The Bayes risk is defined

as the expected cost

$$\mathbb{E}[C] = \sum_{i,j=0}^1 C_{i,j} \pi_j p(\hat{y} = i | y = j) \quad (7.38)$$

$$= \sum_{i,j} C_{i,j} \pi_j \int_{R_i} f_i(x) dx \quad (7.39)$$

$$= \int [1_{R_0}(x)(C_{0,0}\pi_0 f_0(x) + C_{0,1}\pi_1 f_1(x)) \quad (7.40)$$

$$1_{R_1}(x)(C_{1,0}\pi_0 f_0(x) + C_{1,1}\pi_1 f_1(x))] dx \quad (7.41)$$

Note that, for a given x , only one term in the integrand can be nonzero. To minimize the integral, we only need to select the term which is smaller. Therefore, the Bayes risk detector can be derived as follows:

$$C_{0,0}\pi_0 f_0(x) + C_{0,1}\pi_1 f_1(x) \underset{H_0}{\overset{H_1}{\geq}} C_{1,0}\pi_0 f_0(x) + C_{1,1}\pi_1 f_1(x) \quad (7.42)$$

$$\frac{\pi_1 f_1(x)}{\pi_0 f_0(x)} \underset{H_0}{\overset{H_1}{\geq}} \frac{C_{1,0} - C_{0,0}}{C_{0,1} - C_{1,1}} \quad (7.43)$$

$$\frac{\bar{p}}{1 - \bar{p}} \underset{H_0}{\overset{H_1}{\geq}} \frac{C_{1,0} - C_{0,0}}{C_{0,1} - C_{1,1}} \quad (7.44)$$

$$\bar{p} \underset{H_0}{\overset{H_1}{\geq}} \eta^* \quad (7.45)$$

where the optimal threshold

$$\eta^* = \frac{C_{1,0} - C_{0,0}}{C_{1,0} - C_{0,0} + C_{0,1} - C_{1,1}} \quad (7.46)$$

Below we derive the optimal thresholds for several metrics that can be expressed in the Bayes risk with a properly designed cost matrix.

Accuracy When $C_{1,0} = C_{0,1} = 1$ and $C_{0,0} = C_{1,1} = 0$, the Bayes risk is

$$\mathbb{E}[C] = \pi_0 \underbrace{p(\hat{y} = 1 | y = 0)}_{P_F} + \pi_1 \underbrace{p(\hat{y} = 0 | y = 1)}_{P_M} \quad (7.47)$$

$$= P_e = 1 - \text{ACC}, \quad (7.48)$$

where P_F , P_M , and P_e denotes false alarm rate, miss rate, and probability of error, respectively. Therefore, the Bayes detector maximizes the accuracy. From Equation 7.46 we get

the optimal threshold that maximize the accuracy is

$$\eta^* = 0.5 \quad (7.49)$$

TSS When $C_{1,0} = 1/\pi_0$, $C_{0,1} = 1/\pi_1$, $C_{0,0} = C_{1,1} = 0$, the Bayes risk is

$$\mathbb{E}[C] = \frac{1}{\pi_0}\pi_0 P_F + \frac{1}{\pi_1}\pi_1 P_M \quad (7.50)$$

$$= P_F + P_M \quad (7.51)$$

$$= 1 - (P_D - P_F) = 1 - \text{TSS}, \quad (7.52)$$

where $P_D = 1 - P_M$ denotes the probability of detection. Therefore, the Bayes detector with this cost matrix maximizes the TSS. From Equation 7.46 we get

$$\eta^* = \pi_1 \quad (7.53)$$

This verifies the well-known fact that TSS is maximized at the climatology rate π_1 [198, 196, 158]. The cost of incorrect decision for an instance belonging to a class is inversely proportional to the sample size of that class. For solar flare forecast, this means TSS treats a miss π_1/π_0 more costly than a false alarm.

HSS We first show the skill-bias decomposition of HSS.

$$\text{HSS} = \frac{2(a \times d - c \times b)}{pn' + p'n} \quad (7.54)$$

$$= \frac{2}{\underbrace{\frac{n'}{n} + \frac{p'}{p}}_{\alpha}} \underbrace{(P_D - P_F)}_{\text{TSS}} \quad (7.55)$$

$$= \alpha \text{TSS} \quad (7.56)$$

where

$$\alpha = \frac{2s(1-s)}{s + (1-2s)r} \quad (7.57)$$

$$r = sP_D + (1-s)P_F \quad (7.58)$$

On the P_D -vs- P_F axes, both α and TSS depends on P_D and P_F . The HSS depends on P_D and P_F nonlinearly. Thus, it is difficult to express HSS as a Bayes risk. However, on the the skill-bias diagram, we can clearly see that TSS is the skill and α depends only on the bias through the forecast rate r (an example is shown in Figure 7.10).

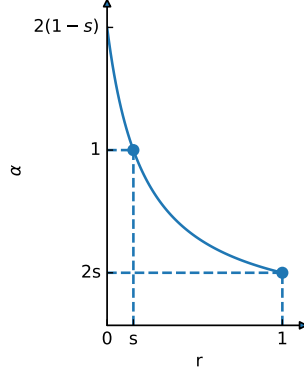


Figure 7.10: The factor α as a function of forecast rate r when the base rate $s = 0.15$. When $r = s$, the forecast frequency is unbiased and $\text{HSS} = \text{TSS}$. When $r > s$, there is a tendency to overforecast and $\text{HSS} < \text{TSS}$. When $r < s$, there is a tendency to underforecast and $\text{HSS} > \text{TSS}$.

Now we derive the optimal threshold for HSS. We first note P_D and P_F depend on the threshold η via

$$P_D(\eta) = \int_{\eta}^{\infty} f_1(s) ds \quad (7.59)$$

$$P_F(\eta) = \int_{\eta}^{\infty} f_0(s) ds \quad (7.60)$$

where s denotes the predicted score or probability of x and f_k denotes the density of s if x belongs to class k . Taking the derivatives, we have $P'_D(\eta) = -f_1(\eta)$ and $P'_F(\eta) = -f_0(\eta)$. Setting the derivative of HSS to zero

$$\text{HSS}'(\eta) = \alpha'(\eta)\text{TSS}(\eta) + \alpha(\eta)\text{TSS}'(\eta) = 0 \quad (7.61)$$

and simplifying using

$$\frac{\eta}{1-\eta} = \frac{f_1(\eta)}{f_0(\eta)} \times \frac{s}{1-s}, \quad (7.62)$$

we have

$$\eta^* = s + \left(\frac{1}{2} - s\right) \text{HSS}(\eta^*) \quad (7.63)$$

When $s \ll 0.5$ and $\text{HSS}(\eta^*) > 0$, we have η^* is larger than the optimal threshold for TSS. So under mild conditions, HSS is decreasing on the interval $\bar{p} \in [\eta^*, 1]$.

7.5.3 Proofs

7.5.3.1 Proof of Lemma 7.1

Proof.

$$\mathbb{E}'[X] = \sum_{k=1}^K \pi'_k \mathbb{E}'[X|Y = k] \quad (7.64)$$

$$= \sum_{k=1}^K \pi'_k \mathbb{E}[X|Y = k] \quad (\text{label shift assumption})$$

$$= \sum_{k=1}^K \pi_k w_k \mathbb{E}[X|Y = k] \quad (7.65)$$

$$= \mathbb{E}[w(Y) \mathbb{E}[X|Y]] \quad (7.66)$$

$$(7.67)$$

□

The importance weights w can be seen as the Radon-Nikodym density of P' w.r.t. P .

7.5.3.2 Proof of Theorem 7.2

Theorem 7.2 (The recalibration g). *Suppose*

$$\mathbb{E}'[(\Phi - Z)^2] \leq \frac{(w_1 - w_0)^2}{4 \max\left(\frac{w_0^2}{w_1}, \frac{w_1^2}{w_0}\right)^2} \mathbb{E}'[(Z(1 - Z))^2],$$

then

$$\mathbb{E}'[(g(\Phi) - g(Z))^2] \leq \mathbb{E}'[(g(\Phi) - Z)^2]$$

Proof. Some useful facts:

$$\begin{aligned} g(\Phi) - g(Z) &= g'(\tilde{Z})(\Phi - Z) && (\tilde{Z} \text{ bounded by } Z \text{ and } \Phi) \\ &= (g'(Z) + g''(\tilde{Z})(\tilde{Z} - Z))(\Phi - Z) && (\tilde{Z} \text{ bounded by } Z \text{ and } \tilde{Z}) \end{aligned} \quad (7.68)$$

To prove

$$\mathbb{E}'[(g(\Phi) - g(Z))^2] \leq \mathbb{E}'[(g(\Phi) - Z)^2]$$

we subtract left hand side (LHS) from the right hand side (RHS):

$$\text{RHS} - \text{LHS} \tag{7.69}$$

$$= \mathbb{E}'[(g(\Phi) - Z)^2] - \mathbb{E}'[(g(\Phi) - g(Z))^2] \tag{7.70}$$

$$= \mathbb{E}'[(g(Z) - Z)^2] + 2\mathbb{E}'[(g(Z) - Z)(g(\Phi) - g(Z))] \tag{7.71}$$

$$= \mathbb{E}'[(g(Z) - Z)^2] + 2\mathbb{E}'[(g(Z) - Z)g'(Z_1)(\Phi - Z)] \quad (Z_1 \text{ bounded by } \Phi \text{ and } Z) \tag{7.72}$$

$$\geq \mathbb{E}'[(g(Z) - Z)^2] - 2\sqrt{\mathbb{E}'[(g(Z) - Z)^2 g'(Z_1)^2]} \sqrt{\mathbb{E}'[(\Phi - Z)^2]} \tag{7.72}$$

It suffices to show

$$\mathbb{E}'[(g(Z) - Z)^2] - 2\sqrt{\mathbb{E}'[(g(Z) - Z)^2 g'(Z_1)^2]} \sqrt{\mathbb{E}'[(\Phi - Z)^2]} \geq 0 \tag{7.73}$$

$$\Leftrightarrow \mathbb{E}'[(\Phi - Z)^2] \leq \frac{\mathbb{E}'[(g(Z) - Z)^2]^2}{4\mathbb{E}'[(g(Z) - Z)g'(Z_1)]} \tag{7.74}$$

We want to find a lower bound for RHS of the above inequality.

$$g'(Z_1)^2 = (g'(Z) + g''(Z_2)(Z_1 - Z))^2 \quad (Z_2 \text{ bounded by } Z \text{ and } Z_1) \tag{7.75}$$

$$= g'(Z)^2 + g''(Z_2)^2(Z_2 - Z)^2 + 2g'(Z)g''(Z_2)(Z_2 - Z) \tag{7.75}$$

$$\leq g'(Z)^2 + g''(Z_2)^2(Z_2 - Z)^2 + 2g'(Z)|g''(Z_2)||Z_2 - Z| \tag{7.76}$$

$$= g'(Z)^2 + L^2 \cdot 4Z^2(1 - Z)^2 + 2ML \cdot 2Z(1 - Z) \tag{7.77}$$

$$= g'(Z)^2 + 4L^2Z^2(1 - Z)^2 + 4MLZ(1 - Z) \tag{7.78}$$

where inequality 7.77 is because $g'(Z)$ is bounded by M , $g''(Z)$ is bounded by L , and

$$|Z_2 - Z| \leq |Z_1 - Z| \leq |\Phi - Z| \leq \max(Z, 1 - Z) \geq 1 - 2Z(1 - Z) = 1 - 2V$$

So, to get inequality 7.74, we only need to control

$$\mathbb{E}'[(\Phi - Z)^2] \leq \frac{\mathbb{E}'[(g(Z) - Z)^2]^2}{4(\mathbb{E}'[(g(Z) - Z)^2 g'(Z)^2] + \mathbb{E}'[(g(Z) - Z)^2 g'(Z)^2])} \tag{7.79}$$

□

CHAPTER 8

Conclusion

8.1 Summary

This dissertation has explored different aspects in probabilistic forecasting, with a particular focus on enhancing calibration, sharpness, and interpretability, in the context of solar flare prediction applications. The research described was divided into three parts: improving solar flare forecasting using multisource multimodal data, structural modeling of multiway covariance for spatio-temporal processes, and recalibration theory and techniques for probabilistic classifiers.

In the first part, comprised of Chapter 2 and Chapter 3, the investigation into solar flare forecasting demonstrated the effectiveness of leveraging diverse data sources and advanced machine learning models. The studies highlighted how combining data from multiple solar cycles and integrating global features such as polar field strengths can significantly improve forecasting accuracy. The use of visual attribution methods also contributed to the interpretability of these models, providing insights into their decision-making processes and uncovering potential limitations.

The second part, comprised of Chapter 4 and Chapter 5, delved into the challenges of modeling multiway covariance and inverse covariance matrices, with particular application in the multichannel spatio-temporal data. The development of Kronecker-structured (inverse-)covariance models and their application to space weather prediction showcased the utility of these methods in handling complex data structures. Furthermore, the introduction of a Julia software package for multiway covariance modeling bridged the gap between theoretical advances and practical implementation, offering tools for broader application in various scientific domains.

In the third part, comprised of Chapter 6 and Chapter 7, the focus shifted to the calibration of probabilistic forecasting models, addressing the crucial aspect of ensuring the reliability of predictions. The research advanced the theoretical understanding of recalibra-

tion methods, particularly in the context of binary classification. The proposed minimum-risk recalibration framework provided a principled approach to balancing calibration and sharpness, with significant implications for the robustness of probabilistic forecasts. The application of these calibration techniques to solar flare forecasting further demonstrated their practical relevance, especially in scenarios involving shifts in event rates.

Overall, this dissertation contributes to the field of probabilistic forecasting by offering both theoretical insights and practical tools for improving the performance and interpretability of predictive models. The findings underscore the importance of integrating diverse data sources, sophisticated modeling techniques, and rigorous calibration methods to address the challenges of forecasting in complex and dynamic environments.

8.2 Future Work

Forecasting solar flares with state-of-the-art deep learning models Since publication of our paper [270], significant advances have been made in deep learning for modeling sequential data and multimodal data. The transformer architectures [297] have revolutionized natural language processing with applications in Large Language Models (LLMs), and have also been applied to computer vision using the vision transformer [73]. In particular, multimodal large language models such as GPT-4, a large-scale, multimodal model which can accept image and text inputs and produce text outputs, achieves human-level performance on various professional and academic benchmarks [2]. It would be interesting to explore the potential of these recent advances in deep learning for solar flare forecasting, where data are naturally multimodal and sequential. The recalibration methods introduced in Chapter 6-7 would be especially worthwhile given challenges in retraining foundation machine learning models.

Combining multiway covariance modeling with deep learning This dissertation has proven the power of the using Kronecker covariance matrices to model multiway data such as solar active region videos. It would be interesting to see if the combination of Kronecker covariance matrix estimation with deep learning models might be valuable in applications. For example, the variational autoencoder (VAE) [152] is a powerful generative deep learning model that can be used to generate data and learn meaningful latent representations. The prior of VAE is typically modeled as a standard Gaussian distribution. Using Kronecker covariance priors and multi-dimensional convolutional layers could be an extension of the latent VAE model given the naturally multiway tensor data. It would be interesting to see if such a Kronecker enabled VAE can perform better in terms of learning latent representations

of solar active regions, as well as the applications for downstream tasks such as solar flare forecasting.

Multiclass extension of binary calibration We have studied in Chapter 6 the binary calibration. We note that the $O(n^{-2/3})$ rate for the recalibration risk $R(\hat{h})$ for $\hat{h}$ fitted by UMB with the number of bins $B \asymp n^{1/3}$ (c.f. Equation (6.11)) matches the minimax rate of univariate nonparametric regression under Lipschitz continuity assumption [291]. It would be interesting to discuss whether the UMB estimator is minimax optimal under our proposed smoothness assumption, Assumption (A3).

In real world, many problems have more than two classes. There are many ways to generalize from binary calibration to multiclass calibration, with the strongest notion being canonical calibration [161]. It is generally believed that the canonical calibration is intractable with exponentially increasing sample complexity as the number of classes K increases. This has been verified in the context of hypothesis testing of canonical calibration [166, 98]. However, the precise form of the minimax lower bound has yet to be developed for canonical calibration function estimation, which is an interesting direction for future work.

One way to mitigate the sample complexity of learning canonical calibration functions is to adopt weaker but more efficient notion of calibration such as weaker notions such as classwise calibration [337] and confidence calibration [109]. It is known that these notions are strictly weaker [105], but there lacks a useful characterization of how much weaker these notions are. In particular, there seems to be a strong connection between classwise calibration and separable functions [236]. It would be interesting to quantify the conditions under which classwise calibration implies canonical calibration. Moreover, the sharpness risk has not been characterized in multiclass calibration. It would be interesting to extend our theory of calibration-sharpness tradeoff in binary calibration in Chapter 6 to the multiclass case. Such a trade-off might be able to provide insights in terms of choices of hyperparameters.

Another way to address the sample complexity of learning canonical calibration function is to impose additional assumptions on the recalibration function. Here we list some examples:

- Since recalibration typically deals with predictions that are already reasonably accurate, e.g., those generated by powerful neural networks. Margin assumptions might be useful to achieve fast rates in learning, similarly to [16, 72].
- One popular assumption in binary recalibration is that the recalibration function is monotonically increasing, reflecting the belief that the probabilistic predictions do not need to be absolutely accurate, but they should capture the general trend and preserve the relative ranking. This assumption drove early work on recalibration and inspired

the approach of isotonic regression [338], as well as its multiclass extensions such as intra- and inter-order preserving [236], or accuracy-preserving transforms [342].

- The union of subspace assumption could be used as a low rank assumption to model the distribution of the probabilistic prediction $f(X)$. Specifically, we could be modeled $f(X)$ as a Mixture of Low-Rank Gaussians (MoLRG) [299]. Specifically, it is observed that for any class, a well-trained model confuses it with only a few other classes. This implies that the distribution of $f(X)$ given a specific class Y is supported on a low-dimensional subspace, and that the distribution of $f(X)$ lies in a union of subspaces.

These assumptions might also be fruitful to consider in our recalibration theory framework, potentially allowing better or more robust approximations to the canonical recalibration function defined in Chapter 6.

Continuous label shift This dissertation does not model the dynamics of label shift. In many cases, label shift occurs as a continuous process with gradual changes in class proportions. Certain smoothness assumptions can be leveraged to predict future class proportions or serve as a prior to reduce noise from small test sample sizes. In solar flare forecasting, the monthly event rate is observed to be noisy (see Figure 7.2). It would be interesting to explore whether modeling the continuous changes in event rates—potentially using sunspot number data—could improve event rate estimates in the target domain.

BIBLIOGRAPHY

- [1] VI Abramenko. Relationship between magnetic power spectrum and flare productivity in solar active regions. *The Astrophysical Journal*, 629(2):1141, 2005.
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Floren-
cia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anad-
kat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [3] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt,
and Been Kim. Sanity checks for saliency maps. In S. Bengio, H. Wal-
lach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors,
Advances in Neural Information Processing Systems, volume 31. Curran Asso-
ciates, Inc., 2018. URL <https://proceedings.neurips.cc/paper/2018/file/294a8ed24b1ad22ec2e7efea049b8737-Paper.pdf>.
- [4] Azim Ahmadzadeh, Berkay Aydin, Manolis K Georgoulis, Dustin J Kempton,
Sushant S Mahajan, and Rafal A Angryk. How to train your flare prediction model:
Revisiting robust sampling of rare events. *The Astrophysical Journal Supplement Se-
ries*, 254(2):23, 2021. doi: 10.3847/1538-4365/abec88.
- [5] John Aitchison and Sheng M Shen. Logistic-normal distributions: Some properties
and uses. *Biometrika*, 67(2):261–272, 1980.
- [6] Deniz Akdemir and Arjun K Gupta. Array variate random variables with multiway
kronecker delta covariance matrix structure. *Journal of algebraic statistics*, 2(1), 2011.
- [7] Mehmet Aktukmak, Zeyu Sun, Monica Bobra, Tamas Gombosi, Ward B Manch-
ester IV, Yang Chen, and Alfred Hero. Incorporating polar field data for improved
solar flare prediction. *Frontiers in Astronomy and Space Sciences*, 9:1040107, 2022.
doi: 10.3389/fspas.2022.1040107.
- [8] Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum likelihood with
bias-corrected calibration is hard-to-beat at label shift adaptation. In *International
Conference on Machine Learning*, pages 222–232. PMLR, 2020.
- [9] Aida Ali, Siti Mariyam Shamsuddin, and Anca L Ralescu. Classification with class
imbalance problem. *Int. J. Advance Soft Compu. Appl*, 5(3), 2013.

- [10] Genevera I Allen and Robert Tibshirani. Transposable regularized covariance models with an application to missing data imputation. *The Annals of Applied Statistics*, 4(2):764, 2010.
- [11] Genevera I Allen and Robert Tibshirani. Inference with transposable data: modelling the effects of row and column correlations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(4):721–743, 2012.
- [12] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*, 2018.
- [13] Rafal A Angryk, Petrus C Martens, Berkay Aydin, Dustin Kempton, Sushant S Mahajan, Sunitha Basodi, Azim Ahmadzadeh, Xumin Cai, Soukaina Filali Boubrahimi, Shah Muhammad Hamdi, et al. Multivariate time series dataset for space weather data analytics. *Scientific data*, 7(1):1–13, 2020.
- [14] C Nick Arge, Carl J Henney, Josef Koller, C Rich Compeau, Shawn Young, David MacKenzie, Alex Fay, and John W Harvey. Air force data assimilative photospheric flux transport (ADAPT) model. In *AIP conference proceedings*, volume 1216, pages 343–346. American Institute of Physics, 2010.
- [15] J Scott Armstrong. Evaluating forecasting methods. *Principles of forecasting: A handbook for researchers and practitioners*, pages 443–472, 2001.
- [16] Jean-Yves Audibert and Alexandre B Tsybakov. Fast learning rates for plug-in classifiers. 2007.
- [17] Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts. In *International Conference on Learning Representations*, 2018.
- [18] Harold D Babcock. The sun’s polar magnetic field. *The Astrophysical Journal*, 130:364, 1959.
- [19] Liang Bai, Yi Bi, Bo Yang, Jun-Chao Hong, Zhe Xu, Zhen-Hong Shang, Hui Liu, Hai-Sheng Ji, and Kai-Fan Ji. Predicting the evolution of photospheric magnetic field in solar active regions using deep learning. *Research in Astronomy and Astrophysics*, 21(5):113, 2021.
- [20] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d’Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008.
- [21] Graham Barnes, KD Leka, CJ Schrijver, Tufan Colak, Rami Qahwaji, OW Ashamari, Y Yuan, J Zhang, RTJ McAteer, DS Bloomfield, et al. A comparison of flare forecasting methods. i. results from the “all-clear” workshop. *The Astrophysical Journal*, 829(2):89, 2016.

- [22] Julian Besag. Efficiency of pseudolikelihood estimation for simple gaussian fields. *Biometrika*, pages 616–618, 1977.
- [23] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017.
- [24] Shamik Bhattacharjee, Rasha Alshehhi, Dattaraj B Dhuri, and Shravan M Hanasoge. Supervised convolutional neural networks for classification of flaring and nonflaring active regions using line-of-sight magnetograms. *The Astrophysical Journal*, 898(2):98, 2020.
- [25] Peter J Bickel and Kjell A Doksum. *Mathematical statistics: basic ideas and selected topics, volumes I-II package*. CRC Press, 2015.
- [26] Fetsje Bijma, Jan C De Munck, and Rob M Heethaar. The spatiotemporal meg covariance matrix modeled as a sum of kronecker products. *NeuroImage*, 27(2):402–415, 2005.
- [27] Christopher M Bishop. Mixture density networks. 1994.
- [28] Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. A unifying theory of distance from calibration. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1727–1740, 2023.
- [29] D Shaun Bloomfield, Paul A Higgins, RT James McAteer, and Peter T Gallagher. Toward reliable benchmarking of solar flare forecasting methods. *The Astrophysical Journal Letters*, 747(2):L41, 2012.
- [30] Monica G Bobra and Sebastien Couvidat. Solar flare prediction using sdo/hmi vector magnetic field data with a machine-learning algorithm. *The Astrophysical Journal*, 798(2):135, 2015.
- [31] Monica G Bobra, Xudong Sun, J Todd Hoeksema, M Turmon, Yang Liu, Keiji Hayashi, Graham Barnes, and KD Leka. The helioseismic and magnetic imager (hmi) vector magnetic field pipeline: Sharps-space-weather hmi active region patches. *Solar Physics*, 289(9):3549–3578, 2014.
- [32] Monica G. Bobra, Paul J. Wright, Xudong Sun, and Michael J. Turmon. SMARPs and SHARPs: Two solar cycles of active region data. *The Astrophysical Journal Supplement Series*, 256(2):26, sep 2021. doi: 10.3847/1538-4365/ac1f1d. URL <https://doi.org/10.3847/1538-4365/ac1f1d>.
- [33] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.
- [34] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.
- [35] Leo Breiman. Stacked regressions. *Machine learning*, 24(1):49–64, 1996.

- [36] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [37] Glenn W Brier et al. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.
- [38] Jochen Bröcker. Estimating reliability and resolution of probability forecasts through decomposition of the empirical score. *Climate dynamics*, 39:655–667, 2012.
- [39] Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, 113(15):3932–3937, 2016.
- [40] Cristina Campi, Federico Benvenuto, Anna Maria Massone, D Shaun Bloomfield, Manolis K Georgoulis, and Michele Piana. Feature ranking of active region source properties in solar flare forecasting and the uncompromised stochasticity of flare occurrence. *The Astrophysical Journal*, 883(2):150, 2019.
- [41] Enrico Camporeale. The challenge of machine learning in space weather: Nowcasting and forecasting. *Space Weather*, 17(8):1166–1207, 2019.
- [42] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.
- [43] Dallas Card and Noah A Smith. The importance of calibration for estimating proportions from annotations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1636–1646, 2018.
- [44] Venkat Chandrasekaran and Michael I Jordan. Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences*, 110(13):E1181–E1190, 2013.
- [45] Venkat Chandrasekaran, Sujay Sanghavi, Pablo A Parrilo, and Alan S Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.
- [46] Subrahmanyan Chandrasekhar. Stochastic problems in physics and astronomy. *Reviews of modern physics*, 15(1):1, 1943.
- [47] AnQin Chen and JingXiu Wang. On the possibility of predicting flare index and cme velocity using vector magnetograms. *Science China Physics, Mechanics & Astronomy*, 63:1–10, 2020.
- [48] Anqin Chen, Qian Ye, and Jingxiu Wang. Flare index prediction with machine learning algorithms. *Solar Physics*, 296(10):1–17, 2021.
- [49] Yang Chen, Ward B Manchester, Alfred O Hero, Gabor Toth, Benoit DuFumier, Tian Zhou, Xiantong Wang, Haonan Zhu, Zeyu Sun, and Tamas I Gombosi. Identifying solar flare precursors using time series of sdo/hmi images and sharp parameters. *Space Weather*, 17(10):1404–1426, 2019. doi: 10.1029/2019SW002214.

- [50] Yilun Chen, Ami Wiesel, and Alfred O Hero. Robust shrinkage estimation of high-dimensional covariance matrices. *IEEE Transactions on Signal Processing*, 59(9):4097–4107, 2011.
- [51] MCM Cheung, Matthias Rempel, M Schüssler, et al. Simulation of the formation of a solar active region. *The Astrophysical Journal*, 720(1):233, 2010.
- [52] Moody T Chu and Matthew M Lin. Nonlinear power-like and svd-like iterative schemes with applications to entangled bipartite rank-1 approximation. *SIAM Journal on Scientific Computing*, (0):S448–S474, 2021.
- [53] Tiago Cinto, André Leon Sampaio Gradvohl, Guilherme Palermo Coelho, and Ana Estela Antunes da Silva. A framework for designing and evaluating solar flare forecasting systems. *Monthly Notices of the Royal Astronomical Society*, 495(3):3332–3349, 2020.
- [54] Todd Clark and Michael McCracken. Advances in forecast evaluation. *Handbook of economic forecasting*, 2:1107–1201, 2013.
- [55] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.
- [56] Tufan Colak and R Qahwaji. Automated solar activity prediction: A hybrid computer platform using machine learning and solar imaging for automated prediction of solar flares. *Space Weather*, 7(6), 2009.
- [57] Eurico Covas. Transfer learning in spatial–temporal forecasting of the solar magnetic field. *Astronomische Nachrichten*, 341(4):384–394, 2020.
- [58] Noel Cressie. *Statistics for spatial data*. John Wiley & Sons, 2015.
- [59] Misty D Crown. Validation of the noaa space weather prediction center’s solar flare forecasting look-up table and forecaster-issued probabilities. *Space Weather*, 10(6), 2012.
- [60] Bin Dai, Yu Wang, John Aston, Gang Hua, and David Wipf. Connections with robust pca and the role of emergent sparsity in variational autoencoder models. *The Journal of Machine Learning Research*, 19(1):1573–1614, 2018.
- [61] Cássio F Dantas, Jérémy E Cohen, and Rémi Gribonval. Learning tensor-structured dictionaries with application to hyperspectral image denoising. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5. IEEE, 2019.
- [62] A Philip Dawid. Some matrix-variate distribution theory: notational considerations and a bayesian application. *Biometrika*, 68(1):265–274, 1981.
- [63] A Philip Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.
- [64] Alexander Philip Dawid and Monica Musio. Theory and applications of proper scoring rules. *Metron*, 72(2):169–183, 2014.

- [65] Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2): 12–22, 1983.
- [66] Yujia Deng, Xiwei Tang, and Annie Qu. Correlation tensor decomposition and its application in spatial imaging data. *Journal of the American Statistical Association*, (just-accepted):1–35, 2021.
- [67] Zheng Deng, Feng Wang, Hui Deng, Lei Tan, Linhua Deng, and Song Feng. Fine-grained solar flare forecasting based on the hybrid convolutional neural networks. *The Astrophysical Journal*, 922(2):232, 2021.
- [68] Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, 2020.
- [69] Varad Deshmukh, Natasha Flyer, Kiera Van der Sande, and Thomas Berger. Decreasing false-alarm rates in cnn-based solar flare prediction using sdo/hmi data. *The Astrophysical Journal Supplement Series*, 260(1):9, 2022.
- [70] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [71] Mausumi Dikpati and Peter A Gilman. Simulating and predicting solar cycles using a flux-transport dynamo. *The Astrophysical Journal*, 649(1):498, 2006.
- [72] Vu Dinh, Lam Si Tung Ho, Nguyen Viet Cuong, Duy Nguyen, and Binh T Nguyen. Learning from non-iid data: Fast rates for the one-vs-all multiclass plug-in classifiers. In *Theory and Applications of Models of Computation: 12th Annual Conference, TAMC 2015, Singapore, May 18-20, 2015, Proceedings 12*, pages 375–387. Springer, 2015.
- [73] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [74] Jiewen Du, Xin Yan, Zhihong Liu, Lu Cui, Peng Ding, Xiaoqing Tan, Xiuming Li, Huihao Zhou, Qiong Gu, and Jun Xu. cbinderdb: a covalent binding agent database. *Bioinformatics*, 33(8):1258–1260, 2017.
- [75] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL <http://archive.ics.uci.edu/ml>.
- [76] Pierre Dutilleul. The mle algorithm for the matrix normal distribution. *Journal of statistical computation and simulation*, 64(2):105–123, 1999.

- [77] Soumya Dutta, Natalie Klein, Li Tang, Jonathan David Wolfe, Luke Van Roekel, James Joseph Benedict, Ayan Biswas, Earl Lawrence, and Nathan Urban. In situ climate modeling for analyzing extreme weather events. In *ISAV'21: In Situ Infrastructures for Enabling Extreme-Scale Analysis and Visualization*, pages 18–23. 2021.
- [78] Saso Džeroski and Bernard Ženko. Is combining classifiers with stacking better than selecting the best one? *Machine learning*, 54(3):255–273, 2004.
- [79] Charles Elkan. The foundations of cost-sensitive learning. In *In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 973–978, 2001.
- [80] Andrea Esuli, Alessio Molinari, and Fabrizio Sebastiani. A critical reassessment of the Saerens-Latinne-Decaestecker algorithm for posterior probability adjustment. *ACM Transactions on Information Systems (TOIS)*, 39(2):1–34, 2020.
- [81] Geir Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, 99(C5):10143–10162, 1994.
- [82] David A Falconer, Ronald L Moore, Abdalnasser F Barghouty, and Igor Khazanov. Prior flaring as a complement to free magnetic energy for forecasting solar eruptions. *The Astrophysical Journal*, 757(1):32, 2012.
- [83] Kostas Florios, Ioannis Kontogiannis, Sung-Hong Park, Jordan A Guerra, Federico Benvenuto, D Shaun Bloomfield, and Manolis K Georgoulis. Forecasting solar flares using magnetogram-based predictors and machine learning. *Solar Physics*, 293(2):1–42, 2018.
- [84] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.
- [85] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.
- [86] Yun Fu and Thomas S Huang. Image classification using correlation tensor analysis. *IEEE Transactions on Image Processing*, 17(2):226–234, 2008.
- [87] Andrzej T Galecki. General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics-Theory and Methods*, 23(11):3105–3119, 1994.
- [88] Richard Galvez, David F Fouhey, Meng Jin, Alexandre Szenicer, Andrés Muñoz-Jaramillo, Mark CM Cheung, Paul J Wright, Monica G Bobra, Yang Liu, James Mason, et al. A machine-learning data set prepared from the nasa solar dynamics observatory mission. *The Astrophysical Journal Supplement Series*, 242(1):7, 2019.
- [89] Howard A Garcia. Temperature and emission measure from goes soft x-ray measurements. *Solar Physics*, 154(2):275–308, 1994.

- [90] Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary Lipton. A unified view of label shift estimation. *Advances in Neural Information Processing Systems*, 33:3290–3300, 2020.
- [91] Rong Ge and Tengyu Ma. On the optimization landscape of tensor decompositions. *Mathematical Programming*, pages 1–47, 2020.
- [92] Manolis K. Georgoulis, D. Shaun Bloomfield, Michele Piana, Anna Maria Massone, Marco Soldati, Peter T. Gallagher, Etienne Pariat, Nicole Vilmer, Eric Buchlin, Frederic Baudin, Andre Csillaghy, Hanna Sathiapal, David R. Jackson, Pablo Alingery, Federico Benvenuto, Cristina Campi, Konstantinos Florios, Constantinos Gontikakis, Chloe Guennou, Jordan A. Guerra, Ioannis Kontogiannis, Vittorio Latorre, Sophie A. Murray, Sung-Hong Park, Samuel von Stachelski, Aleksandar Torbica, Dario Vischi, and Mark Worsfold. The flare likelihood and region eruption forecasting (FLARE-CAST) project: flare forecasting in the big data & machine learning era. *Journal of Space Weather and Space Climate*, 11:39, May 2021. doi: 10.1051/swsc/2021023.
- [93] Xavier Gitiaux, Shane A Maloney, Anna Jungbluth, Carl Shneider, Paul J Wright, Atılım Güneş Baydin, Michel Deudon, Yarin Gal, Alfredo Kalaitzis, and Andrés Muñoz-Jaramillo. Probabilistic super-resolution of solar magnetograms: Generating many explanations and measuring uncertainties. *arXiv preprint arXiv:1911.01486*, 2019.
- [94] Tilmann Gneiting and Matthias Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151, 2014.
- [95] Tilmann Gneiting and Adrian E Raftery. Weather forecasting with ensemble methods. *Science*, 310(5746):248–249, 2005.
- [96] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.
- [97] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2007.00587.x.
- [98] Parikshit Gopalan, Lunjia Hu, and Guy N. Rothblum. On computationally efficient multi-class calibration. In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 1983–2026. PMLR, 30 Jun–03 Jul 2024. URL <https://proceedings.mlr.press/v247/gopalan24a.html>.
- [99] Lars Grasedyck. Existence and computation of low kronecker-rank approximations for large linear systems of tensor product structure. *Computing*, 72(3-4):247–265, 2004.
- [100] Kristjan Greenewald and Alfred O Hero. Robust kronecker product pca for spatio-temporal covariance estimation. *IEEE Transactions on Signal Processing*, 63(23):6368–6378, 2015.

- [101] Kristjan Greenewald, Edmund Zelnio, and Alfred Hero. Robust sar stap via kronecker decomposition. *IEEE Transactions on Aerospace and Electronic Systems*, 52(6):2612–2625, 2016.
- [102] Kristjan Greenewald, Shuheng Zhou, and Alfred Hero. Tensor graphical lasso (teralasso). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(5):901–931, 2019.
- [103] Kristjan H Greenewald and Alfred O Hero. Kronecker pca based spatio-temporal modeling of video for dismount classification. In *Algorithms for Synthetic Aperture Radar Imagery XXI*, volume 9093, page 90930V. International Society for Optics and Photonics, 2014.
- [104] Steven J Greybush, Eugenia Kalnay, Takemasa Miyoshi, Kayo Ide, and Brian R Hunt. Balance and ensemble Kalman filter localization techniques. *Monthly Weather Review*, 139(2):511–522, 2011.
- [105] Sebastian Gruber and Florian Buettner. Better uncertainty calibration via proper scores for classification and beyond. *Advances in Neural Information Processing Systems*, 35:8618–8632, 2022.
- [106] Jordan A Guerra, Sophie A Murray, D Shaun Bloomfield, and Peter T Gallagher. Ensemble forecasting of major solar flares: methods for combining models. *Journal of Space Weather and Space Climate*, 10:38, 2020.
- [107] Jordan A. Guerra, Sophie A. Murray, D. Shaun Bloomfield, and Peter T. Gallagher. Ensemble forecasting of major solar flares: methods for combining models. *Journal of Space Weather and Space Climate*, 10:38, 2020. doi: 10.1051/swsc/2020042.
- [108] Jordan Armando Guerra, A Pulkkinen, and Vadim M Uritsky. Ensemble forecasting of major solar flares: First results. *Space Weather*, 13(10):626–642, 2015.
- [109] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.
- [110] Chirag Gupta and Aaditya Ramdas. Distribution-free calibration guarantees for histogram binning without sample splitting. In *International Conference on Machine Learning*, pages 3942–3952. PMLR, 2021.
- [111] Chirag Gupta and Aaditya Ramdas. Top-label calibration and multiclass-to-binary reductions. In *International Conference on Learning Representations*, 2022.
- [112] Chirag Gupta and Aaditya Ramdas. Online Platt scaling with calibeating. In *International Conference on Machine Learning*, 2023.
- [113] Chirag Gupta, Aleksandr Podkopaev, and Aaditya Ramdas. Distribution-free binary classification: Prediction sets, confidence intervals and calibration. *Advances in Neural Information Processing Systems*, 33:3711–3723, 2020.

- [114] Kartik Gupta, Amir Rahimi, Thalaiyasingam Ajanthan, Thomas Mensink, Cristian Sminchisescu, and Richard Hartley. Calibration of neural networks using splines. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=eQe8DEWNN2W>.
- [115] Yuko Hada-Muranushi, Takayuki Muranushi, Ayumi Asai, Daisuke Okanohara, Rudy Raymond, Gentaro Watanabe, Shigeru Nemoto, and Kazunari Shibata. A deep-learning approach for operation of an automated realtime flare forecast. *arXiv preprint arXiv:1606.01587*, 2016.
- [116] George E Hale, Ferdinand Ellerman, Seth Barnes Nicholson, and Alfred Harrison Joy. The magnetic polarity of sun-spots. *The Astrophysical Journal*, 49:153, 1919.
- [117] Botao Hao, Boxiang Wang, Pengyuan Wang, Jingfei Zhang, Jian Yang, and Will Wei Sun. Sparse tensor additive regression. *Journal of Machine Learning Research*, 22(64):1–43, 2021.
- [118] Karen L Harvey and Cornelis Zwaan. Properties and emergence patterns of bipolar active regions: I. size distribution and emergence frequency. *Solar physics*, 148:85–118, 1993.
- [119] Laura A Hatfield and Alan M Zaslavsky. Separable covariance models for health care quality measures across years and topics. *Statistics in Medicine*, 37(12):2053–2066, 2018.
- [120] David H Hathaway. The solar cycle. *Living reviews in solar physics*, 12(1):1–87, 2015.
- [121] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [122] Robert Hecht-Nielsen. Theory of the backpropagation neural network. In *Neural networks for perception*, pages 65–93. Elsevier, 1992.
- [123] Kyle S Hickmann, Humberto C Godinez, Carl J Henney, and C Nick Arge. Data assimilation in the ADAPT photospheric flux transport model. *Solar Physics*, 290(4):1105–1118, 2015.
- [124] Christopher J Hillar and Lek-Heng Lim. Most tensor problems are np-hard. *Journal of the ACM (JACM)*, 60(6):1–39, 2013.
- [125] KM Hiremath. Prediction of solar cycle 24 and beyond. *Astrophysics and Space Science*, 314(1):45–49, 2008.
- [126] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

- [127] J. T. Hoeksema, Y. Liu, K. Hayashi, X. Sun, J. Schou, S. Couvidat, A. Norton, M. Bobra, R. Centeno, K. D. Leka, G. Barnes, and M. Turmon. The Helioseismic and Magnetic Imager (HMI) vector magnetic field pipeline: Overview and performance. *Solar Physics*, 289:3483–3530, 2014. doi: 10.1007/s11207-014-0516-8.
- [128] J Todd Hoeksema. The large-scale structure of the heliospheric current sheet during the ulysses epoch. *The High Latitude Heliosphere*, pages 137–148, 1995.
- [129] Peter D Hoff. Separable covariance arrays via the tucker product, with applications to multivariate relational data. *Bayesian Analysis*, 6(2):179–196, 2011.
- [130] Samuel B Hopkins, Jonathan Shi, and David Steurer. Tensor principal component analysis via sum-of-square proofs. In *Conference on Learning Theory*, pages 956–1006. PMLR, 2015.
- [131] Tomoaki Hori, David Montcho, Clement Agbangla, Kaworu Ebana, Koichi Futakuchi, and Hiroyoshi Iwata. Multi-task gaussian process for imputing missing data in multi-trait and multi-environment trials. *Theoretical and Applied Genetics*, 129(11):2101–2115, 2016.
- [132] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.
- [133] Elizabeth Hou, Earl Lawrence, and Alfred O Hero. Penalized ensemble kalman filters for high dimensional non-linear systems. *PloS one*, 16(3):e0248046, 2021.
- [134] Robert F Howard. The magnetic fields of active regions: I. data and first results. *Solar physics*, 123:271–284, 1989.
- [135] Xin Huang, Huaning Wang, Long Xu, Jinfu Liu, Rong Li, and Xinghua Dai. Deep learning based solar flare forecasting model. i. results for line-of-sight magnetograms. *The Astrophysical Journal*, 856(1):7, 2018.
- [136] Yingxiang Huang, Wentao Li, Fima Macheret, Rodney A Gabriel, and Lucila Ohno-Machado. A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association*, 27(4):621–633, 2020.
- [137] Željko Ivezić, Steven M Kahn, J Anthony Tyson, Bob Abel, Emily Acosta, Robyn Allsman, David Alonso, Yusra AlSayyad, Scott F Anderson, John Andrew, et al. Lsst: from science drivers to reference design and anticipated data products. *The Astrophysical Journal*, 873(2):111, 2019. doi: 10.3847/1538-4357/ab042c.
- [138] Anli Ji, Berkay Aydin, Manolis K. Georgoulis, and Rafal Angryk. All-clear flare prediction using interval-based time series classifiers. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 4218–4225, 2020. doi: 10.1109/BigData50022.2020.9377906.

- [139] Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, 19(2):263–274, 2012.
- [140] Yiye Jiang, Jérémie Bigot, and Sofian Maabout. Online graph topology learning from matrix-valued time series. *arXiv preprint arXiv:2107.08020*, 2021.
- [141] Zhenbang Jiao, Hu Sun, Xiantong Wang, Ward Manchester, Tamas Gombosi, Alfred Hero, and Yang Chen. Solar flare intensity prediction with machine learning models. *Space Weather*, 18(7):e2020SW002440, 2020.
- [142] Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.
- [143] Ian T Jolliffe. Principal components in regression analysis. In *Principal component analysis*, pages 129–155. Springer, 1986.
- [144] Ian T Jolliffe and David B Stephenson. *Forecast verification: a practitioner’s guide in atmospheric science*. John Wiley & Sons, 2012.
- [145] Eric Jonas, Monica Bobra, Vaishaal Shankar, J Todd Hoeksema, and Benjamin Recht. Flare prediction using photospheric and coronal image data. *Solar Physics*, 293(3): 1–22, 2018.
- [146] Anna Jungbluth, Xavier Gitiaux, Shane A Maloney, Carl Shneider, Paul J Wright, Alfredo Kalaitzis, Michel Deudon, Atılım Güneş Baydin, Yarin Gal, and Andrés Muñoz-Jaramillo. Single-frame super-resolution of solar magnetograms: Investigating physics-based metrics & losses. *arXiv preprint arXiv:1911.01490*, 2019.
- [147] Alfredo Kalaitzis, John Lafferty, Neil D Lawrence, and Shuheng Zhou. The bigraphical lasso. In *International Conference on Machine Learning*, pages 1229–1237. PMLR, 2013.
- [148] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.
- [149] Yuan Ke, Stanislav Minsker, Zhao Ren, Qiang Sun, and Wen-Xin Zhou. User-friendly covariance estimation for heavy-tailed distributions. *Statistical Science*, 34(3):454–471, 2019.
- [150] Kshitij Khare, Sang-Yun Oh, and Bala Rajaratnam. A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 803–825, 2015.
- [151] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

- [152] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [153] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.
- [154] MB Korsós, R Erdélyi, Jiajia Liu, and H Morgan. Testing and validating two morphological flare predictors by logistic regression machine learning. *Frontiers in Astronomy and Space Sciences*, 7:571186, 2021.
- [155] Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.
- [156] Daniel Kressner and Christine Tobler. Krylov subspace methods for linear systems with tensor product structure. *SIAM journal on matrix analysis and applications*, 31(4):1688–1714, 2010.
- [157] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.
- [158] Yûki Kubo. Why do some probabilistic forecasts lack reliability? *Journal of Space Weather and Space Climate*, 9:A17, 2019.
- [159] Volodymyr Kuleshov and Percy S Liang. Calibrated structured prediction. *Advances in Neural Information Processing Systems*, 28, 2015.
- [160] Meelis Kull, Telmo M Silva Filho, and Peter Flach. Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. *Electronic Journal of Statistics*, 11:5052–5080, 2017. doi: 10.1214/17-EJS1338SI.
- [161] Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration. *Advances in Neural Information Processing Systems*, 32, 2019.
- [162] Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. *Advances in Neural Information Processing Systems*, 32, 2019.
- [163] JM Landsberg. *Tensors: Geometry and Applications: Geometry and Applications*, volume 128. American Mathematical Soc., 2011.
- [164] Michael LeBlanc and Robert Tibshirani. Combining estimates in regression and classification. *Journal of the American Statistical Association*, 91(436):1641–1650, 1996.
- [165] Olivier Ledoit and Michael Wolf. The power of (non-) linear shrinking: A review and guide to covariance matrix estimation. *Journal of Financial Econometrics*, 2020.

- [166] Donghwan Lee, Xinmeng Huang, Hamed Hassani, and Edgar Dobriban. T-Cal: An optimal test for the calibration of predictive models. *arXiv preprint arXiv:2203.01850*, 2022.
- [167] K. D. Leka, G. Barnes, and E. L. Wagner. The nwra classification infrastructure: description and extension to the discriminant analysis flare forecasting system (daffs). *J. Space Weather Space Clim.*, 8:A25, 2018. doi: 10.1051/swsc/2018004.
- [168] K. D. Leka, Graham Barnes, and Eric Wagner. The NWRA Classification Infrastructure: description and extension to the Discriminant Analysis Flare Forecasting System (DAFFS). *Journal of Space Weather and Space Climate*, 8:A25, April 2018. doi: 10.1051/swsc/2018004.
- [169] KD Leka and G Barnes. Photospheric magnetic field properties of flaring versus flare-quiet active regions. ii. discriminant analysis. *The Astrophysical Journal*, 595(2):1296, 2003.
- [170] KD Leka and Graham Barnes. Modeling and interpreting the effects of spatial resolution on solar magnetic field maps. *Solar Physics*, 277(1):89–118, 2012.
- [171] KD Leka and Graham Barnes. Solar flare forecasting: present methods and challenges. In *Extreme events in geospace*, pages 65–98. Elsevier, 2018.
- [172] KD Leka, G Barnes, and EL Wagner. Evaluating (and improving) estimates of the solar radial magnetic field component from line-of-sight magnetograms. *Solar Physics*, 292(2):36, 2017.
- [173] KD Leka, Sung-Hong Park, Kanya Kusano, Jesse Andries, Graham Barnes, Suzy Bingham, D Shaun Bloomfield, Aoife E McCloskey, Veronique Delouille, David Falconer, et al. A comparison of flare forecasting methods. ii. benchmarks, metrics, and performance results for operational solar flare forecasting systems. *The Astrophysical Journal Supplement Series*, 243(2):36, 2019.
- [174] KD Leka, Sung-Hong Park, Kanya Kusano, Jesse Andries, Graham Barnes, Suzy Bingham, D Shaun Bloomfield, Aoife E McCloskey, Veronique Delouille, David Falconer, et al. A comparison of flare forecasting methods. ii. benchmarks, metrics, and performance results for operational solar flare forecasting systems. *The Astrophysical Journal Supplement Series*, 243(2):36, 2019.
- [175] KD Leka, Sung-Hong Park, Kanya Kusano, Jesse Andries, Graham Barnes, Suzy Bingham, D Shaun Bloomfield, Aoife E McCloskey, Veronique Delouille, David Falconer, et al. A comparison of flare forecasting methods. iii. systematic behaviors of operational solar flare forecasting systems. *The Astrophysical Journal*, 881(2):101, 2019.
- [176] James R Lemen, David J Akin, Paul F Boerner, Catherine Chou, Jerry F Drake, Dexter W Duncan, Christopher G Edwards, Frank M Friedlaender, Gary F Heyman, Neal E Hurlburt, et al. The atmospheric imaging assembly (aia) on the solar dynamics observatory (sdo). In *The solar dynamics observatory*, pages 17–40. Springer, 2011.

- [177] Chenlei Leng and Guangming Pan. Covariance estimation via sparse kronecker structures. *Bernoulli*, 24(4B):3833–3863, 2018.
- [178] Lexin Li and Xin Zhang. Parsimonious tensor response regression. *Journal of the American Statistical Association*, 112(519):1131–1146, 2017.
- [179] Sijia Li, Martín López-García, Neil D Lawrence, and Luisa Cuttillo. Scalable bi-graphical lasso: Two-way sparse network inference for count data. *arXiv preprint arXiv:2203.07912*, 2022.
- [180] Xuebao Li, Yanfang Zheng, Xinshuo Wang, and Lulu Wang. Predicting solar flares using a novel deep convolutional neural network. *The Astrophysical Journal*, 891(1):10, 2020.
- [181] Zhijin Li, Yi Chao, James C McWilliams, and Kayo Ide. A three-dimensional variational data assimilation scheme for the regional ocean modeling system. *Journal of Atmospheric and Oceanic Technology*, 25(11):2074–2090, 2008.
- [182] Daye Lim, Yong-Jae Moon, Eunsu Park, Jongyeob Park, Kangjin Lee, Jin-Yi Lee, and Soojeong Jang. Ensemble forecasting of major solar flares with short-, mid-, and long-term active region properties. *The Astrophysical Journal*, 885(1):35, 2019.
- [183] Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.
- [184] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*, pages 3122–3130. PMLR, 2018.
- [185] Chang Liu, Na Deng, Jason TL Wang, and Haimin Wang. Predicting solar flares using sdo/hmi vector magnetic data products and the random forest algorithm. *The Astrophysical Journal*, 843(2):104, 2017.
- [186] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.
- [187] Gang Liu, Xiaoqing Tan, Chao Dang, Shuangquan Tan, Shihui Xing, Nianwei Huang, Kangqiang Peng, Chuanmiao Xie, Xiaoying Tang, and Jinsheng Zeng. Regional shape abnormalities in thalamus and verbal memory impairment after subcortical infarction. *Neurorehabilitation and Neural Repair*, 33(6):476–485, 2019.
- [188] Hao Liu, Chang Liu, Jason TL Wang, and Haimin Wang. Predicting solar flares using a long short-term memory network. *The Astrophysical Journal*, 877(2):121, 2019.
- [189] Y Liu, JT Hoeksema, PH Scherrer, J Schou, S Couvidat, RI Bush, TL Duvall, K Hayashi, X Sun, and X Zhao. Comparison of line-of-sight magnetograms taken by the solar dynamics observatory/helioseismic and magnetic imager and solar and

- p heliospheric observatory/michelson doppler imager.
- Solar Physics*
- , 279(1):295–316, 2012.
- [190] Carlos Llosa-Vite and Ranjan Maitra. Reduced-rank tensor-on-tensor regression and tensor-variate analysis of variance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2022. doi: 10.1109/TPAMI.2022.3164836. URL <https://doi.org/10.1109/TPAMI.2022.3164836>.
 - [191] Zichao Long, Yiping Lu, Xianzhong Ma, and Bin Dong. PDE-net: Learning PDEs from data. In *International Conference on Machine Learning*, pages 3208–3216. PMLR, 2018.
 - [192] Xiang Lyu, Will Wei Sun, Zhaoran Wang, Han Liu, Jian Yang, and Guang Cheng. Tensor graphical model: Non-convex optimization and statistical inference. *IEEE transactions on pattern analysis and machine intelligence*, 42(8):2024–2037, 2019.
 - [193] Yingbo Ma, Shashi Gowda, Ranjan Anantharaman, Chris Laughman, Viral Shah, and Chris Rackauckas. Modelingtoolkit: A composable graph transformation system for equation-based modeling, 2021.
 - [194] Subha Maity, Debarghya Mukherjee, Moulinath Banerjee, and Yuekai Sun. Predictor-corrector algorithms for stochastic optimization under gradual distribution shift. *arXiv preprint arXiv:2205.13575*, 2022.
 - [195] Ameer M Manceur and Pierre Dutilleul. Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion. *Journal of Computational and Applied Mathematics*, 239:37–49, 2013.
 - [196] Agostino Manzato. A note on the maximum peirce skill score. *Weather and Forecasting*, 22(5):1148–1154, 2007.
 - [197] Kanti V Mardia and Colin R Goodall. Spatial-temporal analysis of multivariate environmental monitoring data. *Multivariate environmental statistics*, 6(76):347–385, 1993.
 - [198] Ian Mason. On reducing probability forecasts to yes/no forecasts. *Monthly Weather Review*, 107(2):207–211, 1979.
 - [199] SK Mathew, V Martínez Pillet, SK Solanki, and NA Krivova. Properties of sunspots in cycle 23-i. dependence of brightness on sunspot size and cycle phase. *Astronomy & Astrophysics*, 465(1):291–304, 2007.
 - [200] Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.
 - [201] Aoife E McCloskey, Peter T Gallagher, and D Shaun Bloomfield. Flare forecasting using the evolution of mcintosh sunspot classifications. *Journal of Space Weather and Space Climate*, 8:A34, 2018.

- [202] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 34(3):1436–1462, 2006.
- [203] Keqian Min, Qing Mai, and Xin Zhang. Fast and separable estimation in high-dimensional tensor gaussian graphical models. *Journal of Computational and Graphical Statistics*, 31(1):294–300, 2022.
- [204] Shentong Mo, Jingfei Xia, Xiaoqing Tan, and Bhiksha Raj. Point3d: tracking actions as moving points with 3d cnns. *arXiv preprint arXiv:2203.10584*, 2022.
- [205] Aaron J Molstad and Adam J Rothman. A penalized likelihood method for classification with matrix-valued predictors. *Journal of Computational and Graphical Statistics*, 28(1):11–22, 2019.
- [206] Huw Morgan and Youra Taroyan. Global conditions in the solar corona from 2010 to 2017. *Science advances*, 3(7):e1602056, 2017.
- [207] Allan H Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12(4):595–600, 1973.
- [208] Allan H Murphy. Climatology, persistence, and their linear combination as standards of reference in skill scores. *Weather and forecasting*, 7(4):692–698, 1992.
- [209] Allan H Murphy and Edward S Epstein. Verification of probabilistic predictions: A brief review. *Journal of Applied Meteorology and Climatology*, 6(5):748–755, 1967.
- [210] Allan H Murphy and Robert L Winkler. Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 26(1):41–47, 1977.
- [211] Allan H Murphy and Robert L Winkler. A general framework for forecast verification. *Monthly weather review*, 115(7):1330–1338, 1987.
- [212] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.
- [213] Sophie A Murray. The importance of ensemble techniques for operational space weather forecasting. *Space Weather*, 16(7):777–783, 2018.
- [214] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.
- [215] Khanh Nguyen and Brendan O’Connor. Posterior calibration and exploratory analysis for natural language processing models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1587–1598, 2015.
- [216] Weili Nie, Yang Zhang, and Ankit Patel. A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. In *International Conference on Machine Learning*, pages 3809–3818. PMLR, 2018.

- [217] N Nishizuka, K Sugiura, Y Kubo, M Den, S Watari, and M Ishii. Solar flare prediction model with three machine-learning algorithms using ultraviolet brightening and vector magnetograms. *The Astrophysical Journal*, 835(2):156, 2017.
- [218] Naoto Nishizuka, Komei Sugiura, Yuki Kubo, Mitsue Den, and Mamoru Ishii. Deep flare net (defn) model for solar flare prediction. *The Astrophysical Journal*, 858(2):113, 2018.
- [219] Naoto Nishizuka, Yuki Kubo, Komei Sugiura, Mitsue Den, and Mamoru Ishii. Reliable probability forecast of solar flares: Deep flare net-reliable (defn-r). *The Astrophysical Journal*, 899(2):150, 2020.
- [220] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.
- [221] Martin Ohlson, M Rauf Ahmad, and Dietrich Von Rosen. The multilinear normal distribution: Introduction and some basic properties. *Journal of Multivariate Analysis*, 113:37–47, 2013.
- [222] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [223] Margaret Sullivan Pepe. *The statistical evaluation of medical tests for classification and prediction*. Oxford university press, 2003.
- [224] Alexandre Perez-Lebel, Marine Le Morvan, and Gaël Varoquaux. Beyond calibration: Estimating the grouping loss of modern neural networks. In *The Eleventh International Conference on Learning Representations*, 2023.
- [225] Jeffrey M Perkel. Julia: come for the syntax, stay for the speed. *Nature*, 572(7768):141–143, 2019.
- [226] W. D. Pesnell, B. J. Thompson, and P. C. Chamberlin. The solar dynamics observatory (sdo). *Solar Physics*, 275:3–15, 2012. doi: 10.1007/s11207-011-9841-3.
- [227] W. Dean Pesnell. Solar Cycle Predictions (Invited Review). *Solar Physics*, 281(1):507–532, November 2012. doi: 10.1007/s11207-012-9997-5.
- [228] W. Dean Pesnell. Lessons learned from predictions of Solar Cycle 24. *Journal of Space Weather and Space Climate*, 10:60, October 2020. doi: 10.1051/swsc/2020060.
- [229] John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74, 1999.
- [230] Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in Artificial Intelligence*, pages 844–853. PMLR, 2021.

- [231] Teodora Popordanoska, Raphael Sayer, and Matthew Blaschko. A consistent and differentiable l_p canonical calibration error estimator. *Advances in Neural Information Processing Systems*, 35:7933–7946, 2022.
- [232] Mohsen Pourahmadi. Covariance estimation: The glm and regularization perspectives. *Statistical Science*, 26(3):369–387, 2011.
- [233] Saeed Pouryazdian, Soosan Beheshti, and Sridhar Krishnan. Candecomp/parafac model order selection based on reconstruction error in the presence of kronecker structured colored noise. *Digital Signal Processing*, 48:12–26, 2016.
- [234] Guillaume Rabusseau and Hachem Kadri. Low-rank regression with tensor responses. *Advances in Neural Information Processing Systems*, 29:1867–1875, 2016.
- [235] Christopher Rackauckas and Qing Nie. DifferentialEquations.jl—a performant and feature-rich ecosystem for solving differential equations in julia. *Journal of Open Research Software*, 5(1):15, 2017.
- [236] Amir Rahimi, Amirreza Shaban, Ching-An Cheng, Richard Hartley, and Byron Boots. Intra order-preserving functions for calibration of multi-class neural networks. *Advances in Neural Information Processing Systems*, 33:13456–13467, 2020.
- [237] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Numerical Gaussian processes for time-dependent and nonlinear partial differential equations. *SIAM Journal on Scientific Computing*, 40(1):A172–A198, 2018.
- [238] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, et al. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, 2019.
- [239] F Ribeiro and ALS Gradvohl. Machine learning techniques applied to solar flares forecasting. *Astronomy and Computing*, 35:100468, 2021.
- [240] Emile Richard and Andrea Montanari. A statistical model for tensor pca. *Advances in Neural Information Processing Systems*, 27:2897–2905, 2014.
- [241] P Riley, M Ben-Nun, JA Linker, Z Mikic, L Svalgaard, J Harvey, L Bertello, T Hoeksema, Y Liu, and R Ulrich. A multi-observatory inter-comparison of line-of-sight synoptic solar magnetograms. *Solar Physics*, 289(3):769–792, 2014.
- [242] Thomas R Rimmele, Mark Warner, Stephen L Keil, Philip R Goode, Michael Knölker, Jeffrey R Kuhn, Robert R Rosner, Joseph P McMullin, Roberto Casini, Haosheng Lin, et al. The daniel k. inouye solar telescope—observatory overview. *Solar Physics*, 295(12):1–49, 2020.
- [243] B Roberts. Slow magnetohydrodynamic waves in the solar atmosphere. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 364(1839):447–460, 2006.

- [244] B Roberts and AR Webb. Vertical motions in an intense magnetic flux tube. *Solar Physics*, 56:5–35, 1978.
- [245] Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael C Mozer. Mitigating bias in calibration error estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 4036–4054. PMLR, 2022.
- [246] P Romano and F Zuccarello. Photospheric magnetic evolution of super active regions. *Astronomy & Astrophysics*, 474(2):633–637, 2007.
- [247] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.
- [248] Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, 2002.
- [249] P. H. Scherrer, R. S. Bogart, R. I. Bush, J. T. Hoeksema, A. G. Kosovichev, J. Schou, W. Rosenberg, L. Springer, T. D. Tarbell, A. Title, C. J. Wolfson, I. Zayer, and MDI Engineering Team. The Solar Oscillations Investigation - Michelson Doppler Imager. *Solar Physics*, 162(1-2):129–188, December 1995. doi: 10.1007/BF00733429.
- [250] P. H. Scherrer, J. Schou, R. I. Bush, A. G. Kosovichev, R. S. Bogart, J. T. Hoeksema, Y. Liu, T. L. Duvall, J. Zhao, A. M. Title, C. J. Schrijver, T. D. Tarbell, and S. Tomczyk. The Helioseismic and Magnetic Imager (HMI) investigation for the Solar Dynamics Observatory (SDO). *Solar Physics*, 275:207–227, 2012. doi: 10.1007/s11207-011-9834-2.
- [251] J. Schou, P. H. Scherrer, R. I. Bush, R. Wachter, S. Couvidat, M. C. Rabello-Soares, R. S. Bogart, J. T. Hoeksema, Y. Liu, T. L. Duvall, D. J. Akin, B. A. Allard, J. W. Miles, R. Rairden, R. A. Shine, T. D. Tarbell, A. M. Title, C. J. Wolfson, D. F. Elmore, A. A. Norton, and S. Tomczyk. Design and Ground Calibration of the Helioseismic and Magnetic Imager (HMI) Instrument on the Solar Dynamics Observatory (SDO). *Solar Physics*, 275:229–259, January 2012. doi: 10.1007/s11207-011-9842-2.
- [252] Carolus J Schrijver. A characteristic magnetic field pattern associated with all major solar flares and its use in flare forecasting. *The Astrophysical Journal*, 655(2):L117, 2007.
- [253] Carolus J. Schrijver. A characteristic magnetic field pattern associated with all major solar flares and its use in flare forecasting. *The Astrophysical Journal*, 655(2):L117–L120, jan 2007. doi: 10.1086/511857. URL <https://doi.org/10.1086/511857>.
- [254] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

- [255] Mehmet Umut ŞEn and Hakan Erdogan. Linear classifier combination and selection using group sparse regularization and hinge loss. *Pattern Recognition Letters*, 34(3): 265–274, 2013.
- [256] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR, 2017.
- [257] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.
- [258] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [259] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.
- [260] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012.
- [261] J Springenberg, Alexey Dosovitskiy, Thomas Brox, and M Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, 2015.
- [262] Graham Steward, Vasili Lobzin, Iver H Cairns, Bo Li, and David Neudegg. Automatic recognition of complex magnetic regions on the sun in sdo magnetogram images and prediction of flares: Techniques and results for the revised flare prediction program flarecast. *Space Weather*, 15(9):1151–1164, 2017.
- [263] Thomas Stocker. *Introduction to climate modelling*. Springer Science & Business Media, 2011.
- [264] Amos Storkey. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, 30:3–28, 2009.
- [265] Peter Strobach. Low-rank detection of multichannel gaussian signals using block matrix approximation. *IEEE Transactions on Signal Processing*, 43(1):233–242, 1995.
- [266] Guohao Sun, Zishu He, Jun Tong, and Xuejing Zhang. Knowledge-aided covariance matrix estimation via Kronecker product expansions for airborne STAP. *IEEE Geoscience and Remote Sensing Letters*, 15(4):527–531, 2018.
- [267] Will Wei Sun, Botao Hao, and Lexin Li. Tensors in modern statistical learning.

- [268] Xudong Sun, J. Todd Hoeksema, Yang Liu, Thomas Wiegmann, Keiji Hayashi, Qingrong Chen, and Julia Thalmann. Evolution of Magnetic Field and Energy in a Major Eruptive Active Region Based on SDO/HMI Observation. *The Astrophysical Journal*, 748(2):77, April 2012. doi: 10.1088/0004-637X/748/2/77.
- [269] Zeyu Sun, Monica Bobra, Xiantong Wang, Yu Wang, Hu Sun, Tamas Gombosi, Yang Chen, and Alfred Hero. Predicting solar flares using cnn and lstm on two solar cycles of active region data. *Earth and Space Science Open Archive*, page 32, 2021. doi: 10.1002/essoar.10508256.1. URL <https://doi.org/10.1002/essoar.10508256.1>.
- [270] Zeyu Sun, Monica G Bobra, Xiantong Wang, Yu Wang, Hu Sun, Tamas Gombosi, Yang Chen, and Alfred Hero. Predicting solar flares using cnn and lstm on two solar cycles of active region data. *The Astrophysical Journal*, 931(2):163, 2022. doi: 10.3847/1538-4357/ac64a6.
- [271] Zeyu Sun, Dogyoon Song, and Alfred Hero. Minimum-risk recalibration of classifiers. *arXiv preprint arXiv:2305.10886*, 2023.
- [272] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3319–3328, 2017.
- [273] Leif Svalgaard, Thomas L Duvall, and Philip H Scherrer. The strength of the sun’s polar fields. *Solar Physics*, 58(2):225–239, 1978.
- [274] Leif Svalgaard, Edward W Cliver, and Yohsuke Kamide. Sunspot cycle 24: Smallest cycle in 100 years? *Geophysical Research Letters*, 32(1), 2005.
- [275] Naoya Takeishi, Yoshinobu Kawahara, and Takehisa Yairi. Learning Koopman invariant subspaces for dynamic mode decomposition. *arXiv preprint arXiv:1710.04340*, 2017.
- [276] Xiaoqing Tan, Christopher A Ross, Michaeli Miller, and Xiaoying Tang. Change-point analysis of putamen and thalamus subregions in premanifest huntington’s disease. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 531–535. IEEE, 2018.
- [277] Xiaoqing Tan, Judah Abberbock, Priya Rastogi, and Gong Tang. Identifying principal stratum causal effects conditional on a post-treatment intermediate response. In *Conference on Causal Learning and Reasoning*, pages 734–753. PMLR, 2022.
- [278] Xiaoqing Tan, Chung-Chou H Chang, Ling Zhou, and Lu Tang. A tree-based model averaging approach for personalized treatment effect estimation from heterogeneous data sources. In *International Conference on Machine Learning*, pages 21013–21036. PMLR, 2022.
- [279] Xiaoqing Tan, Zhengling Qi, Christopher W Seymour, and Lu Tang. Rise: Robust individualized decision learning with sensitive variables. *arXiv preprint arXiv:2211.06569*, 2022.

- [280] Xiaoqing Tan, Shu Yang, Wenyu Ye, Douglas E Faries, Ilya Lipkovich, and Zbigniew Kadziola. When doubly robust methods meet machine learning for estimating treatment effects from real-world data: A comparative study. *arXiv preprint arXiv:2204.10969*, 2022.
- [281] Li Tang and Earl Lawrence. In-situ data analysis with Julia for E3SM at large scale. JuliaCon: <https://pretalx.com/juliacon2021/talk/GJRKY3/>, 2021.
- [282] Dirk Tasche. Calibrating sufficiently. *Statistics*, 55(6):1356–1386, 2021.
- [283] Siew Leng Teng and Haiyan Huang. A statistical framework to infer functional gene relationships from biologically interrelated microarray experiments. *Journal of the American Statistical Association*, 104(486):465–473, 2009.
- [284] The SunPy Community, Will T. Barnes, Monica G. Bobra, Steven D. Christe, Nabil Freij, Laura A. Hayes, Jack Ireland, Stuart Mumford, David Perez-Suarez, Daniel F. Ryan, Albert Y. Shih, Prateek Chanda, Kolja Glogowski, Russell Hewett, V. Keith Hughitt, Andrew Hill, Kaustubh Hiware, Andrew Inglis, Michael S. F. Kirk, Sudarshan Konge, James Paul Mason, Shane Anthony Maloney, Sophie A. Murray, Asish Panda, Jongyeob Park, Tiago M. D. Pereira, Kevin Reardon, Sabrina Savage, Brigitta M. Sipőcz, David Stansby, Yash Jain, Garrison Taylor, Tannmay Yadav, Rajul, and Trung Kien Dang. The sunpy project: Open source development and status of the version 1.0 core package. *The Astrophysical Journal*, 890:68–, 2020. doi: 10.3847/1538-4357/ab4f7a. URL <https://iopscience.iop.org/article/10.3847/1538-4357/ab4f7a>.
- [285] Junjiao Tian, Yen-Cheng Liu, Nathaniel Glaser, Yen-Chang Hsu, and Zsolt Kira. Posterior re-calibration for imbalanced datasets. *Advances in Neural Information Processing Systems*, 33:8101–8113, 2020.
- [286] Kai Ming Ting and Ian H Witten. Issues in stacked generalization. *Journal of artificial intelligence research*, 10:271–289, 1999.
- [287] Ljupčo Todorovski and Sašo Džeroski. Combining classifiers with meta decision trees. *Machine learning*, 50(3):223–249, 2003.
- [288] Anestis Touloumis, John C. Marioni, and Simon Tavaré. Hypothesis testing for the covariance matrix in high-dimensional transposable data with Kronecker product dependence structure. *Statistica Sinica*, 31:1309–1329, 2021.
- [289] Theodoros Tsiligkaridis and Alfred O Hero. Covariance estimation in high dimensions via kronecker product expansions. *IEEE Transactions on Signal Processing*, 61(21):5347–5360, 2013.
- [290] Theodoros Tsiligkaridis, Alfred O Hero, and Shuheng Zhou. On convergence of kronecker graphical lasso algorithms. *IEEE transactions on signal processing*, 61(7):1743–1755, 2013.

- [291] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York ; London, 2009. ISBN 978-0-387-79051-0 978-0-387-79052-7.
- [292] Lisa Upton and David H Hathaway. Predicting the sun’s polar magnetic fields with a surface flux transport model. *The Astrophysical Journal*, 780(1):5, 2013.
- [293] Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3459–3467. PMLR, 2019.
- [294] Lidia van Driel-Gesztelyi and Lucie May Green. Evolution of active regions. *Living Reviews in Solar Physics*, 12(1):1–98, 2015.
- [295] Charles F Van Loan and Nikos Pitsianis. Approximation with kronecker products. In *Linear algebra for large scale and real-time applications*, pages 293–314. Springer, 1993.
- [296] Cristiano Varin, Nancy Reid, and David Firth. An overview of composite likelihood methods. *Statistica Sinica*, pages 5–42, 2011.
- [297] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. NIPS’17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.
- [298] Pantelis R Vlachas, Wonmin Byeon, Zhong Y Wan, Themistoklis P Sapsis, and Petros Koumoutsakos. Data-driven forecasting of high-dimensional chaotic systems with long short-term memory networks. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474(2213):20170844, 2018.
- [299] Peng Wang, Huijie Zhang, Zekai Zhang, Siyi Chen, Yi Ma, and Qing Qu. Diffusion models learn low-dimensional distributions via subspace clustering. *arXiv preprint arXiv:2409.02426*, 2024.
- [300] Rui Wang, Karthik Kashinath, Mustafa Mustafa, Adrian Albert, and Rose Yu. Towards physics-informed deep learning for turbulent flow prediction. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1457–1466, 2020.
- [301] Xiantong Wang, Yang Chen, Gabor Toth, Ward B Manchester, Tamas I Gombosi, Alfred O Hero, Zhenbang Jiao, Hu Sun, Meng Jin, and Yang Liu. Predicting solar flares with machine learning: investigating solar cycle dependence. *The Astrophysical Journal*, 895(1):3, 2020.
- [302] Y-M Wang and NR Sheeley Jr. Magnetic flux transport and the sun’s dipole moment—new twists to the Babcock-Leighton model. *The Astrophysical Journal*, 375:761–770, 1991.

- [303] Y-M Wang, JL Lean, and NR Sheeley Jr. Modeling the sun’s magnetic field and irradiance since 1713. *The Astrophysical Journal*, 625(1):522, 2005.
- [304] Yu Wang and Alfred Hero. Multiway ensemble kalman filter. *arXiv preprint arXiv:2112.04322*, 2021.
- [305] Yu Wang and Alfred Hero. Sg-palm: a fast physically interpretable tensor graphical model. *International Conference on Machine Learning (ICML)*, *arXiv preprint arXiv:2105.12271*, 2021.
- [306] Yu Wang, Byoungwook Jang, and Alfred Hero. The sylvester graphical lasso (syglasso). In *International Conference on Artificial Intelligence and Statistics*, pages 1943–1953. PMLR, 2020.
- [307] Yu Wang, Natalie Klein, Steven Morley, Vania Jordanova, Michael Henderson, Ayan Biswas, and Earl Lawrence. TributaryPCA: Distributed, streaming PCA for in situ dimension reduction with application to space weather simulations. In *2021 7th International Workshop on Data Analysis and Reduction for Big Scientific Data (DRBSD-7)*, pages 33–39. IEEE, 2021.
- [308] Yu Wang, Zeyu Sun, and Alfred Hero. Tensorgraphicalmodels: A julia toolbox for multiway covariance models and ensemble kalman filter. *Software Impacts*, 13:100308, 2022. ISSN 2665-9638. doi: <https://doi.org/10.1016/j.simpa.2022.100308>. URL <https://www.sciencedirect.com/science/article/pii/S266596382200046X>.
- [309] Yu Wang, Zeyu Sun, and Alfred Hero. Tensorgraphicalmodels: A julia toolbox for multiway covariance models and ensemble kalman filter. *Software Impacts*, page 100308, 2022. doi: 10.1016/j.simpa.2022.100308.
- [310] Yu Wang, Zeyu Sun, Dogyoon Song, and Alfred Hero. Kronecker-structured covariance models for multiway data. *Statistics Surveys*, 16:238–270, 2022. doi: 10.1214/22-SS139.
- [311] Xiaohan Wei and Stanislav Minsker. Estimation of the covariance structure of heavy-tailed distributions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL <https://proceedings.neurips.cc/paper/2017/file/10c272d06794d3e5785d5e7c5356e9ff-Paper.pdf>.
- [312] E Weinan, Jiequn Han, and Linfeng Zhang. Integrating machine learning with physics-based modeling. *arXiv*, 2020.
- [313] Brian T Welsch, Yan Li, Peter W Schuck, and George H Fisher. What is the relationship between photospheric flow fields and solar flares? *The Astrophysical Journal*, 705(1):821, 2009.
- [314] Jonathan Wenger, Hedvig Kjellström, and Rudolph Triebel. Non-parametric calibration for classification. In *International Conference on Artificial Intelligence and Statistics*, pages 178–190. PMLR, 2020.

- [315] Karl Werner, Magnus Jansson, and Petre Stoica. On estimation of covariance matrices with kronecker product structure. *IEEE Transactions on Signal Processing*, 56(2): 478–491, 2008.
- [316] MS Wheatland. Rates of flaring in individual active regions. *Solar Physics*, 203(1): 87–106, 2001. doi: 10.1023/A:1012749706764.
- [317] MS Wheatland. A bayesian approach to solar flare prediction. *The Astrophysical Journal*, 609(2):1134, 2004.
- [318] MS Wheatland. A statistical solar flare forecast method. *Space Weather*, 3(7), 2005.
- [319] Ami Wiesel, Teng Zhang, et al. Structured robust covariance estimation. *Foundations and Trends® in Signal Processing*, 8(3):127–216, 2015.
- [320] Daniel S Wilks. *Statistical methods in the atmospheric sciences*, volume 100. Academic press, 2011.
- [321] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2016.
- [322] David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.
- [323] Frank Woodcock. The evaluation of yes/no forecasts for scientific and administrative purposes. *Monthly Weather Review*, 104(10):1209–1214, 1976.
- [324] Jian Xu, Wen-Qin Wang, and Ronghua Gui. Computational efficient DOA, DOD, and Doppler estimation algorithm for MIMO radar. *IEEE Signal Processing Letters*, 26(1): 44–48, 2018.
- [325] Pan Xu, Tingting Zhang, and Quanquan Gu. Efficient algorithm for sparse tensor-variate gaussian graphical models via gradient descent. In *Artificial Intelligence and Statistics*, pages 923–932. PMLR, 2017.
- [326] Jing-Hao Xue and Peter Hall. Why does rebalancing class-unbalanced data improve auc for linear discriminant analysis? *IEEE transactions on pattern analysis and machine intelligence*, 37(5):1109–1112, 2014.
- [327] Anthony R Yeates. How good is the bipolar approximation of active regions for surface flux transport? *Solar physics*, 295(9):1–19, 2020.
- [328] KL Yeo, A Feller, SK Solanki, S Couvidat, S Danilovic, and NA Krivova. Point spread function of sdo/hmi and the effects of stray light correction on the apparent properties of solar surface phenomena. *Astronomy & Astrophysics*, 561:A22, 2014.
- [329] Kangwoo Yi, Yong-Jae Moon, Daye Lim, Eunsu Park, and Harim Lee. Visual explanation of a deep learning solar flare forecast model and its relationship to physical parameters. *The Astrophysical Journal*, 910(1):8, 2021.

- [330] Jianxin Yin and Hongzhe Li. Model selection and estimation in the matrix normal graphical model. *Journal of multivariate analysis*, 107:119–140, 2012.
- [331] Jun Ho Yoon and Seyoung Kim. Eiglasso: Scalable estimation of cartesian product of sparse inverse covariance matrices. In *Conference on Uncertainty in Artificial Intelligence*, pages 1248–1257. PMLR, 2020.
- [332] Daren Yu, Xin Huang, Huaning Wang, Yanmei Cui, Qinghua Hu, and Rui Zhou. Short-term solar flare level prediction using a bayesian network approach. *The Astrophysical Journal*, 710(1):869, 2010.
- [333] Kai Yu, Mats Bengtsson, Björn Ottersten, Darren McNamara, Peter Karlsson, and Mark Beach. Second order statistics of nlos indoor mimo channels based on 5.2 ghz measurements. In *GLOBECOM’01. IEEE Global Telecommunications Conference (Cat. No. 01CH37270)*, volume 1, pages 156–160. IEEE, 2001.
- [334] Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.
- [335] Yuan Yuan, Frank Y Shih, Ju Jing, and Hai-Min Wang. Automated flare forecasting using a statistical learning technique. *Research in Astronomy and Astrophysics*, 10(8):785, 2010.
- [336] Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012.
- [337] Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 609–616, 2001.
- [338] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 694–699, 2002.
- [339] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.
- [340] Anru Zhang and Dong Xia. Tensor svd: Statistical and computational limits. *IEEE Transactions on Information Theory*, 64(11):7311–7338, 2018.
- [341] Jie Zhang, Yuming Wang, and Yang Liu. Statistical properties of solar active regions obtained from an automatic detection system and the computational biases. *The Astrophysical Journal*, 723(2):1006, 2010.
- [342] Jize Zhang, Bhavya Kaikhura, and T Yong-Jin Han. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International Conference on Machine Learning*, pages 11117–11128. PMLR, 2020.

- [343] Linfeng Zhang, Jiequn Han, Han Wang, Wissam Saidi, Roberto Car, and Weinan E. End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
- [344] Yi Zhang and Jeff G Schneider. Learning multiple tasks with a sparse matrix-normal penalty. In *NIPS*, volume 6, page 2, 2010.
- [345] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR, 2021.
- [346] Yanfang Zheng, Xuebao Li, and Xinshuo Wang. Solar flare prediction with the hybrid deep convolutional neural network. *The Astrophysical Journal*, 885(1):73, 2019.
- [347] Hua Zhou, Lexin Li, and Hongtu Zhu. Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502): 540–552, 2013.
- [348] Jie Zhou, Will Wei Sun, Jingfei Zhang, and Lexin Li. Partially observed dynamic tensor response regression. *Journal of the American Statistical Association*, pages 1–16, 2021.
- [349] Shuheng Zhou. Gemini: Graph estimation with matrix variate normal instances. *The Annals of Statistics*, 42(2):532–562, 2014.