Supporting information for Bayesian Hierarchical Modeling on Covariance Valued Data

Satwik Acharyya
Department of Biostatistics, University of Michigan,
Ann Arbor, MI
Zhengwu Zhang
Department of Statistics and Operations Research, University of North Carolina,
Chapel Hill, NC
Anirban Bhattacharya
Department of Statistics, Texas A&M University,
College Station, TX
Debdeep Pati
Department of Statistics, Texas A&M University,
College Station, TX

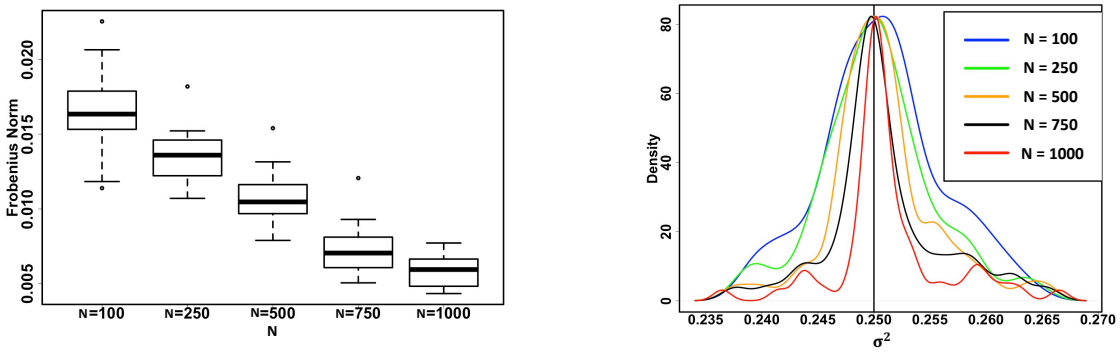
In the Supporting Information, we first explore the detailed simulation study of the independence model (Section 3.1 of the Paper) in Section S1. In section S2, we validate our methodology. We explore a dynamic extension of our hierarchical change point model (8) in Section S3. Section S4 contains sensitivity analysis for the hierarchical change point model (8) and some additional simulations to study robustness. Section S5 contains results for the gambling task for the HCP data. Section S6 & S7 contain details of posterior computation under the independence & hierarchical models respectively.

S1 Simulation Study for Independence Model

We conduct a detailed simulation study to illustrate the performance of the independence model in terms of recovering the true parameters. We fixed $p = 50$ and varied $N \in \{100, 250, 500, 750, 1000\}$. The true intrinsic rank of the data generating mechanism was fixed at 3 to mimic the observation in Figure 2 of the Paper. We set $\phi = p + 1$, the true $\sigma_0^2 = 0.25$, the true $\tilde{D}_0 = \{1.25, 2, 1.55\}$ and considered V_0 from a uniform distribution on Stiefel manifold $\mathcal{V}_{p,r^*}$. 100 independent datasets have been generated from the idependence model (1). We denote the true covariance matrix $\sigma_0^2(V_0\tilde{D}_0V_0^{\mathrm{T}} + I_p)$ by Ω_0.

For model fitting, we set $r^* = 10$; see Section S4 for a sensitivity analysis. The inverse-gamma hyperparameters α_σ and β_σ were elicited in an empirical Bayes approach. Specifically, we used

a method-of-moments type estimator for these hyperparameters. We ran our MCMC algorithm for 10,000 many iterations, discarding the first 5000 many iterates as burn-in. Letting $\widehat{\Omega}_B$ denote an estimate of the posterior mean based on the retained MCMC samples, we provide boxplots of the scaled Frobenius norm difference $\|\widehat{\Omega}_B - \Omega_0\|/p$ across the 100 replicates for the different values of N in Figure S1a. Here, and elsewhere, $\|A\| = \sqrt{\mathrm{tr}\,(A^{\mathrm{T}}A)}$ denotes the Frobenius norm of a matrix. As expected, both the center and spread of the boxplots tend to decrease with increasing N, implying the consistency of the posterior mean in recovering the population mean. Figure S1b shows density plots of the posterior samples of σ^2 which increasingly concentrate around the true value, $\sigma_0^2 = 0.25$, with increasing N.



(a) (b)

Figure S1: *Results for the independence model with* $p = 50$, $\sigma_0^2 = 0.25$, $\tilde{D}_0 = \{1.25, 2, 1.55\}$, *and* $N \in \{100, 250, 500, 750, 1000\}$. *(S1a) Boxplots of the scaled Frobenius norm difference* $\|\widehat{\Omega}_B - \Omega_0\|/p$ *across the 100 replicates for the different values of* N, *where* $\widehat{\Omega}_B$ *is the posterior mean and* Ω_0 *is the population mean. (S1b) Posterior density of* σ^2 *for different values of* N *increasingly concentrate around the true value* 0.25.

Next, we compare the performance of the posterior mean $\widehat{\Omega}_B$ with the sample mean $\widehat{\Omega}_s = N^{-1}\sum_{j=1}^{N} S_i$, which is an unbiased estimator of Ω_0. We consider three different norms between covariance matrices (Ian L. Dryden and Zhou, 2009) listed in Table S1.

Name	**Notation**	**Form**
Euclidean	$d_E(S_1, S_2)$	$\|S_1 - S_2\|$
Riemannian	$d_R(S_1, S_2)$	$\|\log(S_1^{-1/2} S_2 S_1^{-1/2})\|$
Cholesky	$d_C(S_1, S_2)$	$\|\mathrm{chol}(S_1) - \mathrm{chol}(S_2)\|$

Table S1: *Notation and definition of distances between two covariance matrices.*

For each distance d, we compare $d(\widehat{\Omega}_B, \Omega_0)$ and $d(\widehat{\Omega}_s, \Omega_0)$ for $p \in \{50, 100\}$; summary measures are tabulated in Table S2 and Table S3 for $p = 50$ and $p = 100$ respectively. We scaled the distances except the Riemannian norm by p for tabulation; the Riemannian norm is scale-invariant. Boxplots of the distances (in their original scale) across the 100 replicates are provided in Figure S2. It is evident that the posterior mean overall provides a substantial improvement over the sample mean, especially in higher dimensional situations.

$p = 50$	Euclidean		Riemannian		Cholesky	
	PM	SM	PM	SM	PM	SM
$N = 100$	$1.65_{0.20}$	$2.10_{0.22}$	$44.4_{1.8}$	$71.9_{2.0}$	$0.89_{0.4}$	$1.34_{0.7}$
$N = 250$	$1.35_{0.17}$	$1.48_{0.20}$	$42.2_{1.5}$	$62.8_{1.7}$	$0.83_{0.4}$	$1.24_{0.6}$
$N = 500$	$1.01_{0.15}$	$1.20_{0.17}$	$36.1_{1.4}$	$53.6_{1.7}$	$0.76_{0.3}$	$0.99_{0.6}$
$N = 750$	$0.71_{0.09}$	$1.08_{0.11}$	$33.3_{1.0}$	$48.5_{1.3}$	$0.71_{0.1}$	$0.90_{0.3}$
$N = 1000$	$0.51_{0.06}$	$0.93_{0.10}$	$32.2_{0.9}$	$44.6_{1.1}$	$0.67_{0.1}$	$0.82_{0.2}$

Table S2: *The average of* $100 \times d(\widehat{\Omega}_B, \Omega_0)$ *and* $100 \times d(\widehat{\Omega}_s, \Omega_0)$ *over 100 replicates for* $p = 50$*, where* d *generically refers to one of the three distances in Table S1. Subscripts denote* $100\times$*standard deviation across the 100 replicates. PM & SM correspond the distances for the posterior mean and sample mean respectively. All the displayed values are scaled by* $p = 50$*.*

$p = 100$	Euclidean		Riemannian		Cholesky	
	PM	SM	PM	SM	PM	SM
$N = 100$	$1.04_{0.46}$	$1.31_{0.41}$	$39.3_{2.1}$	$98.7_{2.2}$	$0.58_{0.4}$	$0.91_{0.5}$
$N = 250$	$0.92_{0.40}$	$1.18_{0.39}$	$35.5_{1.8}$	$86.1_{2.0}$	$0.51_{0.3}$	$0.79_{0.4}$
$N = 500$	$0.68_{0.28}$	$0.74_{0.26}$	$31.4_{1.6}$	$76.1_{1.8}$	$0.46_{0.3}$	$0.68_{0.3}$
$N = 750$	$0.39_{0.20}$	$0.48_{0.23}$	$30.3_{1.4}$	$69.0_{1.5}$	$0.39_{0.2}$	$0.61_{0.2}$
$N = 1000$	$0.36_{0.17}$	$0.45_{0.18}$	$28.8_{1.4}$	$63.4_{1.4}$	$0.35_{0.1}$	$0.54_{0.2}$

Table S3: *Same setting as in Table S2 with* $p = 100$*.*

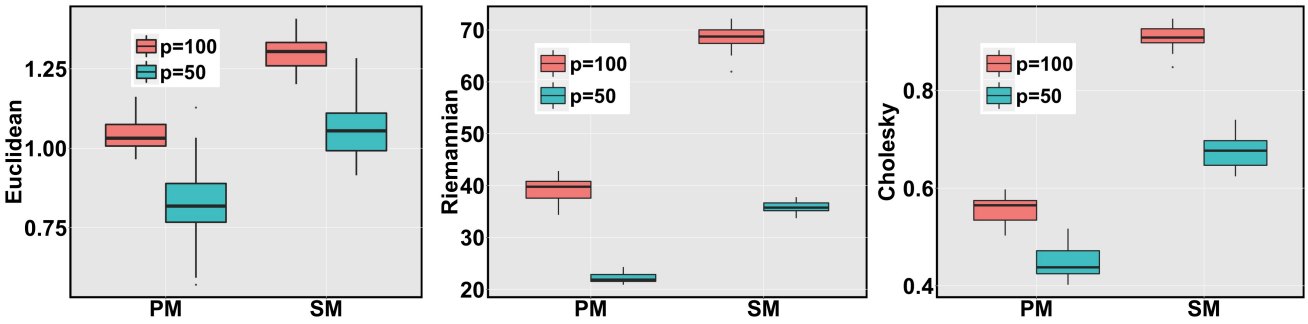
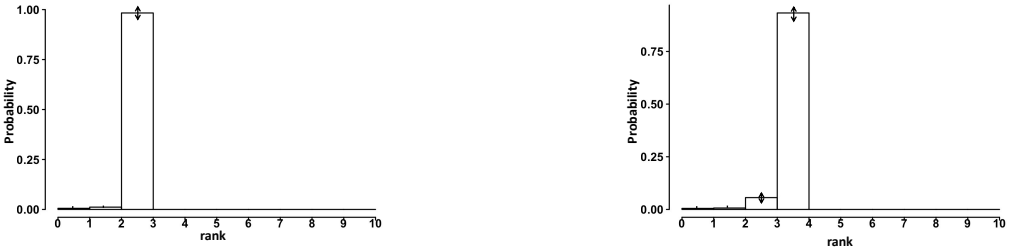


Figure S2: *Boxplots of* $100\times$ $d(\widehat{\Omega}_B, \Omega_0)$ *and* $d(\widehat{\Omega}_s, \Omega_0)$ *over 100 replicates for* $N = 100$*, where* d *generically refers to one of the three distances in Table S1. PM & SM correspond the distances for the posterior mean and sample mean respectively.*

Finally, we illustrate the performance of the post-processing step outlined in the previous subsection to estimate the effective rank. We only consider $N = 100$ and 2 different settings of $\tilde{D}$, (i) $\{1.25, 2, 1.55\}$ and (ii) $\{0.75, 1.25, 2, 1.55\}$. Setting (ii) has a weaker signal strength compared to (i). Following the discussed methodology and setup, we provide the rank estimates under different scenarios in Figure S3 which shows high probability mass at 3 and 4 in Figure S3a and S3b respectively. In different simulation settings, our proposed method is able to recover the true ranks. We highlight that the simulation is performed under mild model misspecification - the true covariance matrix in our simulation study is chosen to be exactly low-rank whereas our model only assumes a near low-rank structure. Despite this, our model is able to detect the true rank.



(a) (b)

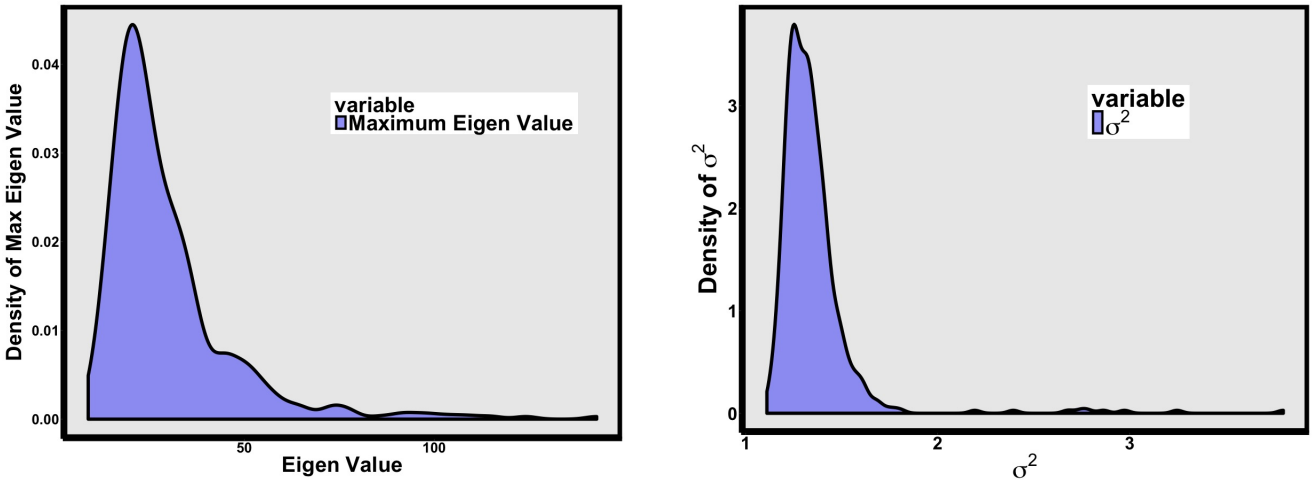
Figure S3: *Rank estimate for Independence model with* $N = 100$, $p = 50$ *and* $\sigma_0^2 = 0.25$. *Lines on the bars shows the standard errors of probability for each point across replicates. The probability is defined as average posterior probability across replicates for each point. (S3a) Left panel plot provides the rank estimate for setting* (i) $\tilde{D} = \{1.25, 2, 1.55\}$ *which indicates high probability mass at* 3. *(S3b) Right panel plot provides the rank estimate for setting* (ii) $\tilde{D} = \{0.75, 1.25, 2, 1.55\}$. *There is a moderately high probability mass at* 4 *and a significant amount of mass at* 3 *because of existence of a weak signal.*

S2 Model Validation

In the following, we first consider an adhoc graphical summary measure of the posterior to justify the extension to hierarchical covariance model from the independence model for the HCP dataset. The exploratory analyses in Figures S4 and S5 were conducted to attest to two aspects of our hierarchical covariance model, namely i) using subject specific D_i and variance components, and ii) using a common semi-orthogonal matrix V across different subjects. A formal model comparison is done later using the WAIC (Gelman et al., 2014) to provide more support to our visual illustration in Figures S4 and S5.

Figure S4 shows the variation of the posterior estimate of the maximum eigenvalue and σ^2 across the individuals, where data for each individual is fitted using the independence model on the 26 time points. The density plot is obtained by smoothing posterior estimates across the individuals. As clearly seen, there is a substantial variability in both the variance component and the maximum eigenvalue which prompted us to consider individual specific σ_i^2 and eigenvalues D_i for the hierarchical covariance model.

Another important modeling assumption in (5) that requires empirical justification is the use of common semi-orthogonal matrix V across all the individuals as opposed to having individual specific semi-orthogonal matrices in the independence model. In the following, we develop a simple diagnostic to this effect. First we fit the independence model separately for each individual and for each i we calculate the Karcher mean (Marks, 2012) of posterior samples of $\{V_i\}_{i=1}^n$ to obtain individual specific posterior estimate of the semi-orthogonal matrices, denoted $\{\hat{V}_i^1\}_{i=1}^n$. For each i, variability of the estimate of the semi-orthogonal matrix is measured as $d(\hat{V}_i^1, \bar{V})$ where $\bar{V} =$ Karcher mean of $\{\hat{V}_i^1\}_{i=1}^n$ and $d(U, W) =|| P_U - P_W ||$ where $P_U = UU^{\mathrm{T}}$. In Figure S5, "Different V" shows the histogram of $\{d(\hat{V}_i^1, \bar{V})\}_{i=1}^n$ describing the variability for the individual specific semi-orthogonal matrices. We generated data from (1) with individual specific Ω_i set as $\bar{V}\hat{D}_i\bar{V}^T + \hat{\sigma}_i^2 I$



(a) (b)

Figure S4: *Density plots of the estimated maximum eigenvalue (S4a) and* σ^2 *(S4b) across individuals obtained from fitting the independence model separately for each individual in the HPC dataset.*

for all individuals $i = 1, \ldots, n$ where $\hat{D}_i$ and $\hat{\sigma}_i$ are the individual specific posterior estimates from the independence model. We then refit the independence model to this new dataset and acquired individual specific posterior estimates of the semi-orthogonal matrices $= \{\hat{V}_i^2\}_{i=1}^n$. Variability of these semi-orthogonal matrices is measured as $d(\hat{V}_i^2, \bar{V})$ for each $i(= 1, \ldots, n)$. "Same V" in Figure S5 denotes the histogram of $\{d(\hat{V}_i^2, \bar{V})\}_{i=1}^n$. Figure S5 clearly indicates a reduction in the variability of semi-orthogonal matrices if only a single semi-orthogonal matrix is considered across all individuals.

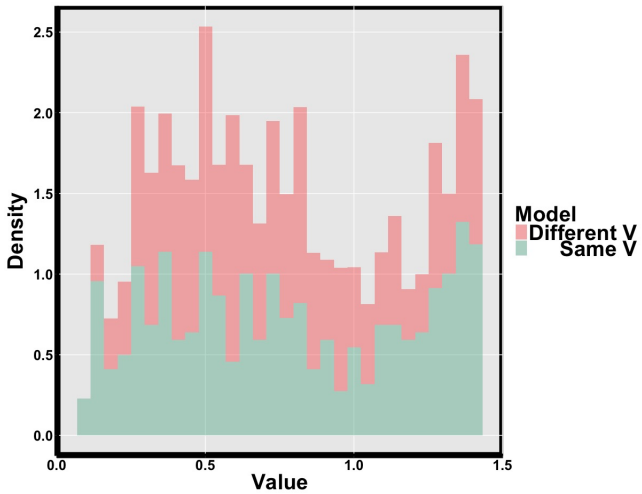


Figure S5: *"Different V" indicates the variation in* V_i*s which are obtained from fitting independence model separately for each individual and measured the deviation of* V_i*s from* $\bar{V}$ *which is karcher mean of individual specific* V_i*s. Later we have simulated a datum with* $\bar{V}$*, estimated* $\tilde{D}$ *and estimated* σ^2 *and measured the variation in a similar fashion which is represented with "Same V". Here deviations are measured as* $d(U, W) = || P_U - P_W ||$ *where* $P_U = UU^{\mathrm{T}}$*.*

After fitting a complex Bayesian model, it is important to compare its predictive accuracy with other models, both simple and complex (Geisser and Eddy, 1979; Hoeting et al., 1999; Vehtari and Ojanen, 2012). Cross-validation and information criteria are two approaches to estimate out-of-sample predictive accuracy using within-sample fits. The deviance information criterion (DIC) (Spiegelhalter et al., 2002) is a generalization of the Akaike information criterion (AIC) (Akaike, 1998) for hierarchical settings. DIC has gained in popularity in recent years, in part through its implementation in the graphical modeling package BUGS, but it is known to suffer from issues from not being fully Bayesian. The Watanabe-Akaike information criterion (WAIC) (Watanabe, 2010) can be viewed as an improvement on the deviance information criterion (DIC) for Bayesian models. In this article, we use WAIC to compare the Independence model, the hierarchical covariance model and the its extension to detect changepoints. Computation time for the WAIC is negligible compared to the cost of fitting the model and obtaining posterior draws. The WAIC values suggest

Model	WAIC
Independence Model	37.7500
Hierarchical Covariance Model	20.1703
Hierarchical Change-point Model	19.6739

Table S4: *WAIC values for three models which are defined in (1), (5) and (8) respectively for HCP dataset under motor task with fitted rank value* 5. *Reported WAIC values are in scale of* 10^4.

progressively better fits as we move from the independence model to the most complex change point model.

S2.1 Validation of the value of ϕ

In this section, we validate the choice of ϕ based on WAIC. Under the same simulation setting as in Section S1, the data is generated from the independence model (1) by setting the true values of ϕ in $\{11, 15, 20\}$. The fitted values of ϕ are varied over a range $\{11, \ldots, 20\}$ and corresponding WAIC values are reported in Table (S5). Irrespective of the true value of ϕ, the lowest WAIC value is obtained when $\phi = 11 \in \{11, \ldots, 20\}$. For the HCP data analysis, we compare the WAIC values from the hierarchical change point model (8) for several values of ϕ as shown in Table (S6). The lowest WAIC is attained at $\phi = 11$.

S2.2 Model validation with ADNI Dataset

We applied the hierarchical change point model on the previously described ADNI dataset, where the subjects are believed to be at a resting state. As anticipated, our model did not detect any change points in the dataset for both supernormals and health control groups. To test for model adequacy, we used the Watanabe-Akaike information criterion (WAIC) (Gelman et al., 2014). WAIC is a fully Bayesian approach to measure model accuracy computed with log pointwise posterior predictive density and then adding a correction for the effective number of parameters to adjust for

True $\phi = 11$				True $\phi = 15$				True $\phi = 20$			
ϕ	WAIC	ϕ	WAIC	ϕ	WAIC	ϕ	WAIC	ϕ	WAIC	ϕ	WAIC
11	223980.5	16	982594.9	11	224046.5	16	891307.5	11	227826.6	16	832607.7
12	350710.9	17	1161531.7	12	331893.7	17	1052557.1	12	326247.8	17	980285.6
13	492837.7	18	1346223.5	13	454632.8	18	1219633.2	13	437647.2	18	1133538.3
14	646175.0	19	1535622.7	14	590192.5	19	1391076.2	14	559376.4	19	1291091.1
15	810530.4	20	1729035.9	15	736743.1	20	1566453.2	15	692127.0	20	1452505.0

Table S5: *WAIC values obtained from the independence model (1) with true values of* $\phi = \{11, 15, 20\}$ *and fitted values are varied over a range* $\{11, \ldots, 20\}$.

Real Data Analysis			
ϕ	WAIC	ϕ	WAIC
11	196739	16	257329
12	209801	17	271796
13	218313	18	289761
14	233022	19	297537
15	242873	20	312590

Table S6: *WAIC values obtained by implementing hierarchical change point model (8) on the HCP dataset with different values of* $\phi \in \{11, \ldots, 20\}$.

overfitting. Table S7 provides the WAIC values for ADNI dataset with respect to the independence model (1), hierarchical covariance model (5) and the change point model (8) respectively. The WAIC values are lowest in the case of hierarchical covariance model which reaffirms the lacks of change points in resting state fMRI dataset. Higher WAIC values for the hierarchical change-point model suggests overfitting from a more complex model where the data do not have a changepoint.

S3 Dynamic Change Point Model

In our current change-point estimation framework, the observational units before and after the change-point are assumed to be independent and identically distributed. A dynamic extension of our model is certainty interesting to incorporate the dependence across the time-points. To that end, we explore an auto-regressive Wishart process on the subject specific loadings. This is studied

Model	SN	HC
Independence Model	80.32	76.71
Hierarchical Covariance Model	49.07	44.63
Hierarchical Change-point Model	68.57	61.96

Table S7: WAIC values for ADNI dataset with respect to three models which are defined in (1), (5) and (8) respectively. Reported WAIC values are in scale of 10^2.

under the hierarchical change point model for one subject with a single change point:

$$\begin{aligned} S_{it} &\sim W(\phi_1, \phi_1^{-1}\Omega_{1_i}^{(t)}), \quad \Omega_{1_i}^{(t)} = V_1 D_{1_i}^{(t)} V_1^{\mathrm{T}} + \sigma_{1_i}^2 I_p, \quad t = 1, \ldots, c_i\ , \\ S_{it} &\sim W(\phi_2, \phi_2^{-1}\Omega_{2_i}^{(t)}), \quad \Omega_{2_i}^{(t)} = V_2 D_{2_i}^{(t)} V_2^{\mathrm{T}} + \sigma_{2_i}^2 I_p, \quad t = c_i + 1, \ldots, T. \end{aligned} \tag{S1}$$

We use the same orthogonal factor model type decomposition on Ω at each time point of an individual. As a parsimonious model for the time dependent covariance matrices, the orthogonal matrix V is held constant with subject specific time varying process for the loadings $D_i^{(t)}$ and subject specific variance components. For computational convenience, we consider $\tilde{D}_i^{(t)} = D_i^{(t)}/\sigma^2$ which are assumed to evolve following an auto-regressive process:

$$\left.\begin{aligned} \log \tilde{d}_{1_{ih}}^{(t)} &= \rho_1 \log \tilde{d}_{1_{ih}}^{(t-1)} + \epsilon_{1_{ih}}^{(t)}, \quad t = 2, \ldots, c_i \\ \log \tilde{d}_{2_{ih}}^{(t)} &= \rho_2 \log \tilde{d}_{2_{ih}}^{(t-1)} + \epsilon_{2_{ih}}^{(t)}, \quad t = c_i + 2, \ldots, T \end{aligned}\right\} \quad h = 1, \ldots, r^*. \tag{S2}$$

The diagonal elements at the first time point are shrunk using global and local shrinkage:

$$\begin{aligned} \log \tilde{d}_{1_{ih}}^{(1)} &= \log \lambda_{1_{ih}}^{(1)} + \log \tau_{1_i}^{(1)} \\ \log \tilde{d}_{2_{ih}}^{(c_i+1)} &= \log \lambda_{2_{ih}}^{(c_i+1)} + \log \tau_{2_i}^{(c_i+1)}. \end{aligned} \tag{S3}$$

Here $\tau_{1_i}^{(1)}$, $\tau_{2_i}^{(c_i+1)}$ are the global shrinkage parameters while the $\lambda_{ih}^{(1)}$, $\lambda_{2_{ih}}^{(c_i+1)}$ parameters allow for coordinate specific deviations. We place independent half-Cauchy priors on the local parameters and half-Cauchy prior truncated to $(0, 1)$ on the global parameters. Also, $\epsilon_{1_{ih}}^{(t)} \sim N(0, v_1)$ and $\epsilon_{2_{ih}}^{(t)} \sim N(0, v_2)$ with $v_1, v_2 \overset{iid.}{\sim}$ Gamma(1/2,2).

We detected change points of 62 individuals from the dynamic change point model which is provided in Figure S6. There are indeed certain change points which are detected by both (8) and (S1). For example, the change points for individual 66, 133 and 173 are detected at 24th, 23rd and 10th time point using both the models. Although the specific dynamic framework can be extended to a full-blown dynamic model, initial indications of the WAIC values suggest that it may not be a good fit for the current dataset. We leave this as a topic for future research.

S4 Sensitivity and robustness analysis of the hierarchical change point model

We considered different values of the fitted rank ($r^* = 5, 6, 7, 8$) and ran the hierarchical change point model (8) on the HCP dataset under the Motor task to check the sensitivity of the results with respect to the fitted rank. The findings are provided in Table S8. We obtained similar results for 26 individuals from different values of fitted ranks which are provided in the last column of Table S8. Next we considered five different initializations for the model parameters in (8) and fixed $r^* = 5$. Detected change points with the corresponding individuals are shown in Table S9 and the frequently appeared 34 individuals for all those 5 settings are provided in last column of the table.

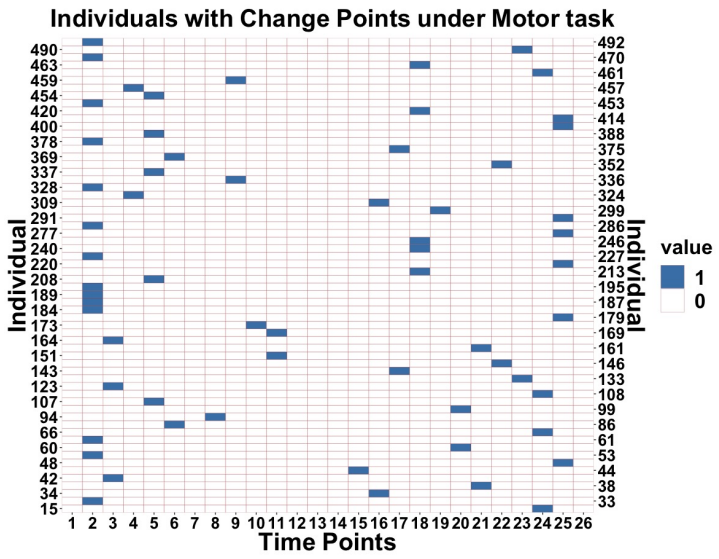


Figure S6: *Detected change points for different individuals under Motor task from dynamic change point model* (S1). *On both sides of the y-axis of the plot we have labeled the individuals. Blue lines denote the individual specific change points.*

The description of these initial values are provided next. For setting 1, we set $\{\sigma^2_{1_i}, \sigma^2_{2_i}\}_{i=1}^n$ to be 2.5, $\{\tau_{1_i}, \tau_{2_i}\}_{i=1}^n = 0.1$ and $\{\lambda_{1_i}, \lambda_{2_i}\}_{i=1}^n = 2.5$. In case of setting 2, we consider $\{\sigma^2_{1_i}, \sigma^2_{2_i}\}_{i=1}^n$ to be 1.75, $\{\lambda_{1_i}, \lambda_{2_i}\}_{i=1}^n = 2.25$ and values for all values of τ were same with setting 1. Values of $\{\sigma^2_{1_i}, \sigma^2_{2_i}\}_{i=1}^n$ was set to be 1.5, $\{\lambda_{1_i}, \lambda_{2_i}\}_{i=1}^n = 3$ and same value for all τ parameters for setting 3. In setting 4, we have $\{\sigma^2_{1_i}, \sigma^2_{2_i}\}_{i=1}^n$ to be 2.25, $\{\lambda_{1_i}, \lambda_{2_i}\}_{i=1}^n = 3.5$ and $\{\tau_{1_i}, \tau_{2_i}\}_{i=1}^n = 0.2$. In case of setting 5, we consider $\{\sigma^2_{1_i}, \sigma^2_{2_i}\}_{i=1}^n$ to be 2.5, $\{\lambda_{1_i}, \lambda_{2_i}\}_{i=1}^n = 3.25$ and $\{\tau_{1_i}, \tau_{2_i}\}_{i=1}^n = 0.25$.

We conducted another experiment to test the robustness of the change point model when the data generating distribution indeed has correlation across time points. We simulated data from the time varying change point model (S1) consisting of 10 subjects with 26 time points. All the subjects have individual specific change points which are generated uniformly in the interval $\{2, \ldots, 25\}$. For clarity, any parameter with subscript "1" correspond to the pre-change-point regime (Group 1) and the ones with subscript "2" corresponds to the post-change-point (Group 2). Individual specific variances $\{\sigma^2_{01_i}\}_{i=1}^n$ and $\{\sigma^2_{02_i}\}_{i=1}^n$ are generated from uniform$(0.25, 0.50)$. Local parameters $\{\lambda^{(1)}_{1_i}, \lambda^{(c_i+1)}_{2_i}\}_{i=1}^n$ and global parameters $\{\tau^{(1)}_{1_i}, \tau^{(c_i+1)}_{2_i}\}_{i=1}^n$ are generated from half-Cauchy distribution. $\{D^{(1)}_{01_i}\}_{i=1}^n$ and $\{D^{(1)}_{02_i}\}_{i=1}^n$ are generated by following equation (S3). We use AR(1) process (S2) to generate rest of the columns of D_{01} and D_{02}. Figure S7 shows that our non-time varying model(8) model is able to detect the individual specific change points for 60% cases even though the data is generated from the time varying hierarchical change point model.

S5 Case study for gambling task of HCP dataset

A detailed description of the gambling task corresponding to HCP is in Delgado et al. (2000). We applied our hierarchical change point model on data which was scaled with subject specific

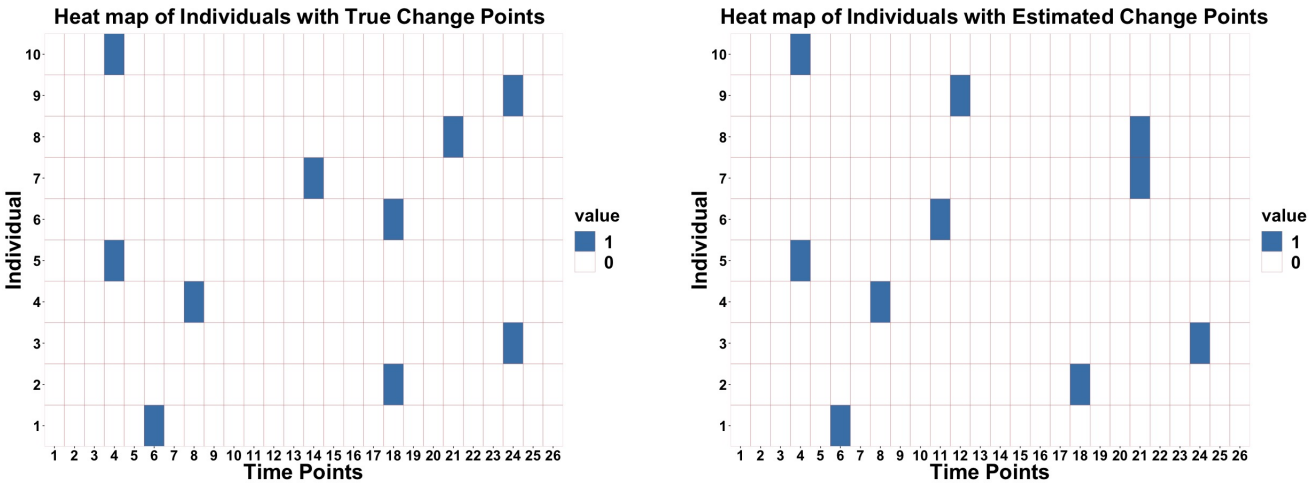


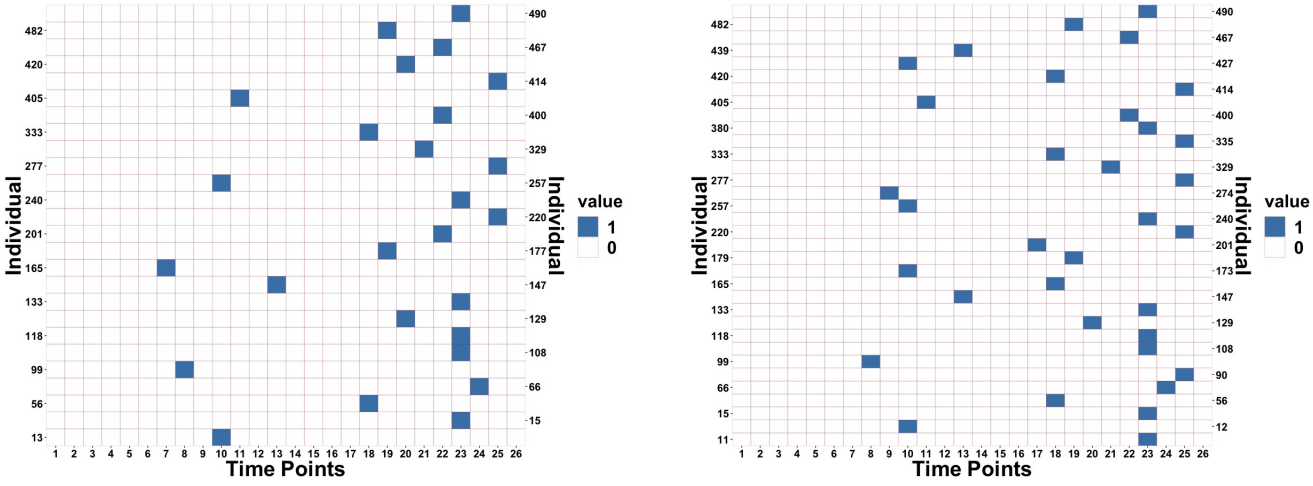
Figure S7: Individual specific change point estimates obtained from hierarchical change point model (8) whereas the data is generated from time-varying hierarchical change point model (S1).

$r^* = 5$		$r^* = 6$		$r^* = 7$		$r^* = 8$		Common Ind	
Change Point	Ind	Change Points	Ind	Change Points	Ind	Change Points	Ind	Change Points	Ind
23	11	10	13	10	13	23	11	10	13
10	13	24	15	24	15	10	13	23	15
23	15	18	56	18	56	23	15	18	56
18	56	24	66	24	66	18	56	24	66
24	66	21	69	21	69	24	66	8	99
25	90	25	90	25	90	21	69	23	108
8	99	8	99	3	93	23	82	23	118
23	108	22	108	25	101	25	90	20	129
23	118	24	118	25	108	8	99	23	133
20	129	20	129	20	129	25	101	13	147
23	133	23	133	23	133	22	108	7	165
13	147	13	147	13	147	23	118	19	177
18	165	18	165	7	165	21	122	22	201
10	173	19	179	10	173	20	129	25	220
19	179	22	201	19	179	23	133	23	240
17	201	25	211	22	201	13	147	10	257
25	220	25	220	25	220	7	165	25	277
23	240	23	240	23	240	10	173	21	329
10	257	10	257	10	257	18	177	18	333
25	270	25	277	25	270	19	179	22	400
9	274	18	333	9	274	22	201	11	405
25	277	25	335	25	277	25	220	25	414
21	329	23	380	21	329	20	224	20	420
18	333	25	383	18	333	17	225	22	467
25	335	22	400	25	383	2	240	19	482
23	380	11	405	22	400	10	257	23	490
23	381	25	414	11	405	25	264		
22	400	18	420	25	414	24	267		
11	405	22	467	28	420	25	270		
25	414	19	482	10	427	9	274		
18	420	23	490	22	467	25	277		
10	427			19	482	6	291		
13	439			23	490	21	329		
22	467					18	333		
19	482					23	380		
23	490					23	381		
						14	399		
						22	400		
						11	405		
						25	414		
						18	420		
						18	427		
						22	467		
						19	482		
						23	490		

Table S8: *Detected change points and corresponding individuals from the hierarchical change point model (8) for different values of fitted rank.*

Setting 1		Setting 2		Setting 3		Setting 4		Setting 5		Common Ind	
Change Point	Ind	Change Points	Ind	Change Points	Ind	Change Points	Ind	Change Points	Ind	Change Points	Ind
23	11	23	11	23	11	23	11	23	11	23	11
23	15	10	13	10	13	10	13	10	13	10	13
18	56	23	15	23	15	23	15	23	15	23	15
24	66	18	56	18	56	19	23	18	56	18	56
21	69	24	66	24	66	19	32	24	66	24	66
25	90	25	90	25	90	18	56	25	90	25	90
8	99	8	99	8	99	24	66	8	99	8	99
23	108	23	108	23	108	21	69	24	108	23	108
23	118	23	118	23	118	25	90	23	118	23	118
20	129	20	129	20	129	8	99	20	129	20	129
23	133	23	133	23	133	25	101	23	133	23	133
13	147	13	147	13	147	23	108	13	147	13	147
18	165	7	165	7	165	23	118	18	165	18	165
10	173	10	173	10	173	20	129	10	173	10	173
19	179	19	179	19	179	23	133	19	179	19	179
17	201	17	201	17	201	25	139	17	201	17	201
25	220	25	220	25	220	13	147	25	220	25	220
23	240	23	240	23	240	7	165	23	240	23	240
10	257	10	257	10	257	10	173	10	257	10	257
9	274	25	270	9	274	19	179	25	270	9	274
25	277	9	274	25	277	22	201	9	274	25	277
21	329	25	277	21	329	25	220	25	277	21	329
18	333	21	329	18	333	23	240	21	329	18	333
25	335	18	333	25	335	10	257	18	333	25	335
23	380	25	335	22	400	25	270	25	335	23	380
23	381	23	380	25	414	9	274	23	380	22	400
22	400	25	381	18	420	25	277	23	381	11	405
11	405	22	400	13	439	21	329	22	400	25	414
25	414	11	405	22	467	18	333	11	405	18	420
18	420	25	414	19	482	25	335	25	414	10	427
24	427	18	420	23	490	23	380	18	420	13	439
13	439	10	427			23	381	10	427	22	467
22	467	13	439			14	399	13	439	19	482
19	482	22	467			22	400	22	467	23	490
23	490	19	482			11	405	19	482		
		23	490			25	414	23	490		
						18	420				
						10	427				
						13	439				
						22	467				
						19	482				
						23	490				

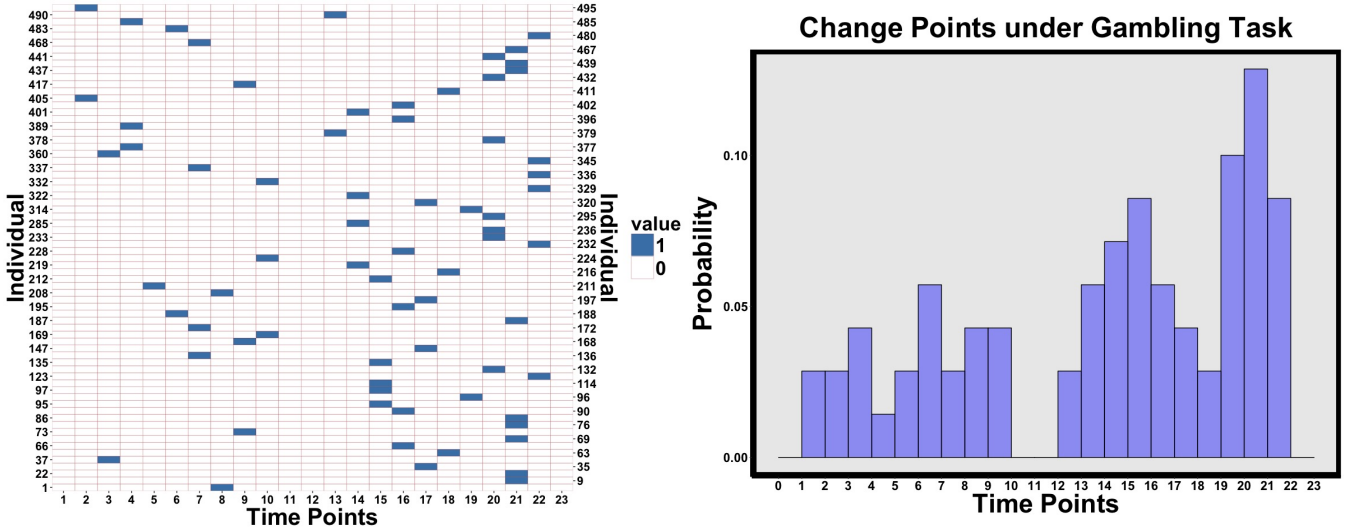
Table S9: *Detected change points and corresponding individuals from the hierarchical change point model (8) for different initial setting.*



(a) (b)

Figure S8: *Common change points and corresponding individuals detected from the hierarchical change point model (8) for different values of fitted rank (S8a) and initial setting (S8b).*

minimum eigen value. We are able to detect 70 individuals with change points under Gambling task. In Figure S9 we labeled the 70 individuals with their corresponding change points. Figure S9a consists of 70 individuals with their most prominent change point detected through the hierarchical model. Figure S9b shows the overall pattern of the most prominent change points across all the individuals. The histogram in Figure S9b shows that more than 10% individuals have change points at 21. Next we extended our study to detect multiple change points under gambling task.



(a) (b)

Figure S9: *Right panel shows heat map of binary matrix consisting of* 1 *to* $(i, j)^{th}$ *position which corresponds to* i^{th} *individual and* j^{th} $(j = 1, \ldots, 23)$ *time point which is a change point for the corresponding individual and* 0 *otherwise. Individuals with one change point under gambling task are labeled on y-axis. In left panel we have histogram of change points under gambling task which shows most of the individuals have change point at* 21*.*

We applied the methodology discussed in Section 5.2 on individuals under gambling task and detected multiple change points for different individuals. In Figure S10, we listed the individuals with at least two change points and their corresponding change points.

S6 Posterior Computation under Independence Model

We first record two useful identities. We have

$$|\Omega| = |\sigma^2 I_p + VDV^{\mathrm{T}}| = (\sigma^2)^p \left| I_p + V\tfrac{D}{\sigma^2}V^{\mathrm{T}} \right| = (\sigma^2)^p |I_p + V\tilde{D}V^{\mathrm{T}}|$$

$$= (\sigma^2)^p |I_{r^*} + (V^{\mathrm{T}}V)\tilde{D}| = (\sigma^2)^p \prod_{j=1}^{r^*} \left(1 + \frac{d_j}{\sigma^2}\right) = (\sigma^2)^p \prod_{j=1}^{r^*} (1 + \tilde{d}_j) \ ,$$

where $\tilde{D} = D/\sigma^2$. Moreover,

$$\Omega^{-1} = (\sigma^2 I_p + VDV^{\mathrm{T}})^{-1} = \sigma^{-2}\left(I_p + V\tfrac{D}{\sigma^2}V^{\mathrm{T}}\right)^{-1} = \sigma^{-2}(I_p + V\tilde{D}V^{\mathrm{T}})^{-1}$$

$$= \sigma^{-2}\{I_p - V(\tilde{D}^{-1} + V^{\mathrm{T}}V)^{-1}V^{\mathrm{T}}\} = \sigma^{-2}\{I_p - V(\tilde{D}^{-1} + I_{r^*})^{-1}V^{\mathrm{T}}\} = \sigma^{-2}\{I_p - VE^{-1}V^{\mathrm{T}}\} \ ,$$

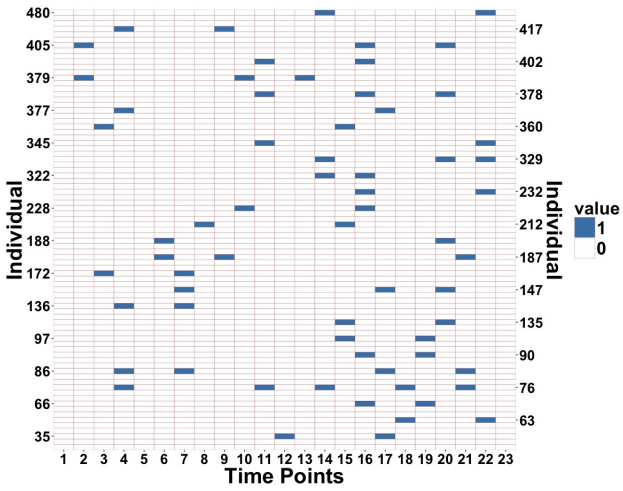


Figure S10: *Multiple change points for different individuals under Gambling task. On y-axis of the plot we have labeled the individuals with multiple change points and on x-axis we have time points. Blue lines denote the individual specific change points.*

where $E = \tilde{D}^{-1} + I_{r^*}$.

- **Likelihood under Independence Model**: The joint likelihood of (V, D, σ^2) is

  $L(V, D, \sigma^2) = |\Omega|^{-\frac{N\phi}{2}} \prod_{j=1}^{N} \exp\left\{ -\frac{1}{2}\text{tr}\,(\Omega^{-1} S_j)\right\} = |\Omega|^{-\frac{N\phi}{2}} \exp\left\{ -\frac{1}{2}\text{tr}\,(\Omega^{-1} \sum_{j=1}^{N} S_j)\right\}$

  $= |\Omega|^{-\frac{N\phi}{2}} \exp\left\{ -\frac{1}{2}\text{tr}\,(\Omega^{-1} S^N)\right\}$

  $= (\sigma^2)^{-\frac{N\phi p}{2}} \prod_{j=1}^{r^*} \left(1 + \frac{d_j}{\sigma^2}\right)^{-\frac{N\phi}{2}} \exp\left[-\frac{1}{2\sigma^2}\text{tr}\left\{(I_p - V(\sigma^2 D^{-1} + I_{r^*})^{-1} V^{\mathrm{T}}) S^N\right\}\right],$

  where $S^N = \sum_{j=1}^{N} S_j$.

  With the transformation $\tilde{D} = D/\sigma^2$, the joint likelihood of $(V, \tilde{D}, \sigma^2)$ is

  $L(V, \tilde{D}, \sigma^2) = (\sigma^2)^{-\frac{N\phi p}{2}+1} \prod_{j=1}^{r^*} (1 + \tilde{d_j})^{-\frac{N\phi}{2}} \exp\left[-\frac{1}{2\sigma^2}\text{tr}\left\{(I_p - V(\tilde{D}^{-1} + I_{r^*})^{-1} V^{\mathrm{T}}) S^N\right\}\right].$

  We now describe the full-conditional distributions that are used to implement a Metropolis within Gibbs sampler to sample from the joint posterior of $(V, \tilde{D}, \sigma^2)$. We use the notation $[\theta \mid -]$ to denote the full conditional distribution of a parameter.

- **Full conditional of V**:

  We have,

  $[V \mid -] \propto \exp\left[\frac{1}{2\sigma^2}\text{tr}\left\{V E^{-1} V^{\mathrm{T}} S^N\right\}\right] = \exp\left[\frac{1}{2\sigma^2}\text{tr}\left\{E^{-1} V^{\mathrm{T}} S^N V\right\}\right].$

  This is a Bingham$(S^N, E^{-1}/2\sigma^2)$ distribution, which can be sampled using the package Rstiefel in R.

- **Full conditional of β_h**:

We decomposed $\log \tilde{d}_h = \mu + \beta_h$ and considered $M = V^T S^N V$ and a transformation $w_h = (1 + \beta_h \mu)^{-1}$ for $h = 1, \ldots, r^*$. The full-conditional of w_h is give by

$$[w_h \mid -] \propto w_h^{\frac{N\phi}{2}-2} \exp\left(- \tfrac{\phi M_{hh}}{2\sigma^2} w_h \right) \frac{1}{\mu^2 + \left(\frac{1-w_h}{w_h}\right)^2},$$

To sample from the above, we consider a slice-sampling scheme. Specifically, augment a latent variable u_h such that the joint distribution of (w_h, u_h) is

$$[w_h, u_h \mid -] \propto w_h^{\frac{N\phi}{2}-2} \exp\left(- \tfrac{\phi M_{hh}}{2\sigma^2} w_h \right) I\left[0 < u_h < \frac{1}{\mu^2 + \left(\frac{1-w_h}{w_h}\right)^2} \right].$$

It is clearly seen that the marginal distribution of w_h is preserved under this joint distribution. We then successively sample from the conditionals $[u_h \mid w_h, -]$ and $[w_h \mid u_h, -]$. We have

$[u_h \mid w_h, -] \sim \text{Uniform}[0, \{\mu^2 + ((1-w_h)/w_h)^2\}^{-1}]$,

$[w_h \mid u_h, -] \sim \text{Gamma}(\text{shape}= (N\phi/2) - 1, \text{ rate} = \phi M_{hh}/2\sigma^2) I[w_h > \{1 + \sqrt{(1/u_h) - \mu^2}\}^{-1}]$.

- **Full conditional of μ**:

  The full conditional of μ is

$$[\mu \mid -] \propto \prod_{j=1}^{r^*} (1 + \beta_h \exp(\mu))^{-\frac{N\phi}{2}} \exp\left[\sum_{j=1}^{r^*} c_h (1 + \frac{1}{\beta_h \exp(\mu)})^{-1} \right] \frac{\exp(\mu)}{1 + \exp(2\mu)}.$$

  where $c_h = \phi M_{hh}/(2\sigma^2)$. We use a Metropolis–Hastings step to sample μ using a normal random-walk proposal as $\mu^* \sim N(\mu, s^2)$ with standard deviation $s = 0.1$. We accept μ^* with probability $\min\{\alpha(\mu, \mu^*), 1\}$ where

$$\alpha(\mu, \mu^*) = \frac{\Pi(\mu^* \mid -)}{\Pi(\mu \mid -)}.$$

- **Full conditional of σ^2:**

$$[\sigma^2 \mid -] \propto (\sigma^2)^{-\frac{N\phi p}{2}+1} \exp\left[-\tfrac{1}{2\sigma^2} \text{tr}\,\{QS^N\} \right] . (\sigma^2)^{\alpha_\sigma - 1} \exp\left(-\tfrac{\beta_\sigma}{\sigma^2} \right)$$

$$= (\sigma^2)^{-\frac{N\phi p}{2} - \alpha_\sigma} \exp\left[-\tfrac{1}{\sigma^2} \left\{ \beta_\sigma + \tfrac{tr(QS^N)}{2} \right\} \right],$$

  where $Q = (V\tilde{D}V^{\mathrm{T}} + I_p)^{-1} V^{\mathrm{T}}$. This implies

  $[\sigma^2 \mid -] \sim \text{Inverse-Gamma}(\alpha_\sigma - 1 + Np\phi/2, \beta_\sigma + \text{tr}\,(QS^N)/2)$.

S7 Posterior Computation and Algorithm under Hierarchical Model

The joint likelihood of (V, D, σ^2) under the hierarchical covariance model (5) is

$$
\begin{aligned}
&L(V, D, \sigma^2) \\
&= \Big[\prod_{i=1}^{n}\prod_{t=1}^{T}|\Omega_i|^{-\frac{\phi}{2}}\exp\Big\{-\frac{\phi}{2}\mathrm{tr}\,(\Omega_i^{-1}S_{it})\Big\}\Big] = \prod_{i=1}^{n}|\Omega_i|^{-\frac{T\phi}{2}}\exp\Big\{-\frac{\phi}{2}\mathrm{tr}\,(\Omega_i^{-1}\sum_{t=1}^{T}S_{it})\Big\} \\
&= \Big[\prod_{i=1}^{n}\Big\{(\sigma_i^2)^{-\frac{pT\phi}{2}}\prod_{j=1}^{r^*}\Big(1+\frac{d_{ij}}{\sigma_i^2}\Big)^{-\frac{T\phi}{2}}\Big\}\Big]\exp\Big\{-\frac{\phi}{2}\mathrm{tr}\Big(\sum_{i=1}^{n}\sigma_i^{-2}\{I_p - V(\sigma_i^2 D_i^{-1} + I_{r^*})^{-1}V^{\mathrm{T}}\}\sum_{t=1}^{T}S_{it}\Big)\Big\}.
\end{aligned}
$$

With the transformation $\tilde{D} = D/\sigma^2$, the joint likelihood of $(V, \tilde{D}, \sigma^2)$ is

$$
\begin{aligned}
&L(V, \tilde{D}, \sigma^2) \\
&= \Big[\prod_{i=1}^{n}\Big\{(\sigma_i^2)^{-\frac{pT\phi}{2}+1}\prod_{j=1}^{r^*}(1+\tilde{d}_{ij})^{-\frac{T\phi}{2}}\Big\}\Big]\exp\Big\{-\frac{\phi}{2}\mathrm{tr}\Big(\sum_{i=1}^{n}\sigma_i^{-2}\{I_p - V(\tilde{D}_i^{-1} + I_{r^*})^{-1}V^{\mathrm{T}}\}\sum_{t=1}^{T}S_{it}\Big)\Big\}.
\end{aligned}
$$

Now we only describe detailed steps to derive the full-conditional of V. For the rest of the parameters, we cycled through the subject specific full-conditionals which are similarly derived following the posterior computation steps under the Independence model (1). We have

$$
[V \mid -] \propto \exp\Big[\frac{\phi}{2}\sum_{i=1}^{n}\mathrm{tr}\Big\{V\frac{E_i^{-1}}{\sigma_i^2}V^{\mathrm{T}}\sum_{t=1}^{T}S_{it}\Big\}\Big] = \exp\Big[\frac{\phi}{2}\sum_{i=1}^{n}\mathrm{tr}\Big\{V\frac{E_i^{-1}}{\sigma_i^2}V^{\mathrm{T}}S_i^*\Big\}\Big],
$$

where $S_i^* = \sum_{t=1}^{T}S_{it}$ and $E_i = (\tilde{D}_i^{-1} + I_{r^*})$. The detailed steps of the identity used in (6) are given as

$$
\begin{aligned}
&\Big(V\frac{\phi E_i^{-1}}{2\sigma_i^2}V^{\mathrm{T}}\Big)S_i^* = \sum_{j=1}^{r^*}(v_j v_j^{\mathrm{T}})\Big(\frac{\phi_1 S_i^*}{2e_{ij}\sigma_i^2}\Big), \\
&\sum_{i=1}^{n}\Big(V\frac{\phi E_i^{-1}}{2\sigma_i^2}V^{\mathrm{T}}\Big)S_i^* = \sum_{i=1}^{n}\sum_{j=1}^{r^*}(v_j v_j^{\mathrm{T}})\Big(\frac{\phi S_i^*}{2e_{ij}\sigma_i^2}\Big) \\
&= \sum_{j=1}^{r^*}(v_j v_j^{\mathrm{T}})\sum_{i=1}^{n}\Big(\frac{\phi S_i^*}{2e_{ij}\sigma_i^2}\Big) = \sum_{j=1}^{r^*}(v_j v_j^{\mathrm{T}})H_j = \sum_{j=1}^{r^*}(v_j^{\mathrm{T}}H_j v_j).
\end{aligned}
$$

The full-conditional of V under the hierarchical covariance model (5) can thus be expressed as

$$
[V \mid -] \propto \exp\Big[\sum_{j=1}^{r^*}v_j^{\mathrm{T}}H_j v_j\Big] = \prod_{j=1}^{r^*}\exp(v_j^{\mathrm{T}}H_j v_j) \;\text{ where }\; H_j = \sum_{i=1}^{n}\Big(\frac{\phi S_i^*}{2e_{ij}\sigma_i^2}\Big).
$$

To sample from the above, we follow the steps in §3.3 of Hoff (2009) and write $V = \{V_{[,1]}, V_{[,-1]}\} =$

$\{Nz, V_{[,-1]}\}$ with $\|z\| = 1$. Here, N is an $p \times (r^* - 1)$ orthonormal basis for the null space of $V_{[,-1]}$ and z is expressed as $z = N^{\mathrm{T}} V_{[,1]}$ because $N^{\mathrm{T}} N = I$. The conditional density of $z \mid V_{[,-1]}$ is derived as in Hoff (2009) as

$$p(z \mid V_{[,-1]}) \propto \exp(z^{\mathrm{T}} N^{\mathrm{T}} H_j N z) = \exp(z^{\mathrm{T}} \tilde{H}_j z).$$

We iterate through the steps 1–4 for each $j \in \{1, \ldots, r^*\}$ to obtain samples from the density (6):

1) N = null space of $V_{[,-j]}$ and $z_j = N^T V_{[,j]}$.
2) $\tilde{H}_j = N^{\mathrm{T}} H_j N$.
3) Sample z_j from a vector Bingham($\tilde{H}_j$) density using the package `rsteifel` (Hoff, 2013).
4) Set $v_j = N z_j$.

Next, we outline the full conditionals of the time points under hierarchical change point model (8). Full conditional updates of the rest of the parameters are similar to full conditional under hierarchical covariance model (5). We have

$$P(c_i = k|-) = \frac{A_{k_i}}{\sum_{k=1}^{T} A_{k_i}}, \quad k = 1, \ldots, T,$$

$$\text{where } A_{k_i} = (\sigma^2_{1_i})^{-\frac{p\phi_1 k}{2}+1} \prod_{j=1}^{r_1} (1 + \tilde{d}_{1_{ij}})^{-\frac{\phi_1 k}{2}} \exp\left[-\frac{\phi_1}{2\sigma^2_{1_i}} \mathrm{tr}\left\{ Q_{1_i} S'_{1_i} \right\} \right] \times$$

$$(\sigma^2_{2_i})^{-\frac{p\phi_2 (T-k)}{2}+1} \prod_{j=1}^{r_2} (1 + \tilde{d}_{2_{ij}})^{-\frac{\phi_2 (T-k)}{2}} \exp\left[-\frac{\phi_2}{2\sigma^2_{2_i}} \mathrm{tr}\left\{ Q_{2_i} S'_{2_i} \right\} \right],$$

with $S'_{1_i} = \sum_{t=1}^{k} S_{it}$, $S'_{2_i} = \sum_{t=k+1}^{T} S_{it}$.

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of Hirotugu Akaike*, pages 199–213. Springer.

Delgado, M., Nystrom, L., Fissell, C., Noll, D., and Fiez, J. (2000). Tracking the hemodynamic responses to reward and punishment in the striatum. *Journal of Neurophysiology*, 84:3072–3077.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and computing*, 24(6):997–1016.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401.

Hoff, P. (2013). Bayesian analysis of matrix data with rstiefel. *R-vignette*.

Hoff, P. D. (2009). Simulation of the matrix Bingham–von Mises–Fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, 18(2):438–456.

Ian L. Dryden, A. K. and Zhou, D. (2009). Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics*, 3(3):1102–1123.

Marks, J. (2012). Mean variants on matrix manifolds. *Doctor of Philosophy (Ph.D.) Thesis*, Colorado State University.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series B (statistical methodology)*, 64(4):583–639.

Vehtari, A. and Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec):3571–3594.