Statistical Learning Methods for Integrative and Dynamic Modeling in Precision Medicine

by

Qingzhi Liu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2025

Doctoral Committee:

Professor Veera Baladandayuthapani, Co-Chair
Associate Professor Gen Li, Co-Chair
Professor Brian Athey
Associate Professor Irina Gaynanova

Qingzhi Liu
qingzliu@umich.edu
ORCID iD: 0000-0003-0682-5487

© Qingzhi Liu 2025

DEDICATION

In memory of my maternal grandparents.

During my years studying abroad, I lost you; I never lost your love.

ACKNOWLEDGEMENTS

As this five-year PhD journey draws to a close, it feels long–yet I can still clearly remember my days in the Department of Mathematics at the University of Maryland, preparing my PhD applications. This journey was driven by curiosity about how data and science come together, and by the joy of exploring cutting-edge research. It was sustained by advisors who offered patient guidance, friends who filled my days with laughter and support, and family whose unconditional love reached me from the other side of the world. To everyone who walked with me–thank you. I could not have reached this milestone without you.

I would first like to express my deepest gratitude to my co-advisor, Professor Veera Baladandayuthapani. He is an exceptional teacher and scholar, and a mentor in life. It has been an honor and a privilege to be his student. From him I learned far more than how to publish papers; more importantly, he taught me how to think, how to stay calm in the face of setbacks, and how to communicate–both in daily conversation and through storytelling in my writing. These lessons are a lifelong gift. I am equally grateful to my co-advisor, Professor Gen Li, a brilliant and patient mentor. When I felt uncertain about my research direction in my second year, his guidance helped me find a clear path. Throughout my work he offered detailed guidance and incisive suggestions. Even when parts of my dissertation fell outside his primary research areas, he always made time for me and provided prompt, thoughtful feedback.

I would also like to express my heartfelt thanks to the two professors on my dissertation committee. Professor Brian Athey has collaborated with me since my second year of the PhD, provided nearly three years of research funding, inspired the studies of bipolar disorder in Chapters 3 and 4, and made available invaluable clinical data. Throughout this collaboration, he has maintained a deep curiosity and enthusiasm for frontier science, exemplifying true academic leadership that has profoundly influenced me. I first came to know and admire Professor Irina Gaynanova through her papers for their clarity and rigor; I am honored to have received her patient and meticulous feedback on this dissertation, which substantially strengthened the work.

Beyond my advisors and committee, I am grateful to the many professors who supported and encouraged me at different stages: Professors Shuo Chen (my first mentor in biostatistics,

whose invaluable guidance took me from zero to one on the road to data-driven discovery), Jian Kang, Yongqun He, Danny Gulick, Limin Peng, Xiangming Fang, Thomas Braun, Zhenke Wu, and Lars Fritsche. I also thank Dr. Melvin McInnis and Dr. Anastasia Yocum for sharing the PrBP cohort data. I extend sincere thanks to mentors in industry–Dr. Erik Bloomquist, Dr. Yabing Mai, Dr. Annpey Pong, Dr. Qi Cao, Dr. Zikun Wang, and Dr. Boxiong Tang–for their guidance, perspective, and generous advice. I am grateful to all the teachers who taught me from primary school through university, and I would like to give special thanks to my high-school homeroom teacher, Ms. Xiaoqing Wang.

To my friends who walked alongside me–thank you. As we've grown, most of us now live far apart, in many corners of the world, yet our friendship has remained constant. During graduate school, Jiaheng Xie, Lingfeng Luo, Tiancheng Shan, Lang Zeng, Jiahan Chen, Molin Yue, Peijun Wu, Zheng Li, Leyi Wang, Tongbo Xu, Chao Gao, Tao Zhang, Chang Xu, Bu Zhao, Jiahui Ji, Ben Brennan, Rupam Bhattacharyya, and many others: the time we shared in Ann Arbor is full of joy and good memories. During college, Yang Xu, Zheyu Wu, Xiang Gao, Xuefang Cui, Jiahan Zhu, Zhaoqi Sun, and all my roommates–especially Haotian Cai–thank you. Since my undergraduate years, many senior classmates and mentors have offered generous help, and I am grateful to them as well. To my close friends from middle and high school–Chen Chen, Xinru Li, He Li, Qianyu Long, Tianhang Huang, and Shuyi Qu–thank you for your enduring friendship. Many thanks to my childhood friend Yichen Wang, and special thanks to my companions of twenty years–Yuchao Zhu and Lujin Zhao–who have been like brothers to me.

Above all, I owe special thanks to my family–my parents; my late maternal grandparents; my paternal grandparents; my uncles and aunts; and my cousins. Your unconditional love has filled my life for 30 years. To my fiancée, thank you for your steadfast companionship and encouragement. Together, you are my strongest foundation.

And finally, a word of thanks to myself–for never giving up, for staying grateful, for keeping my curiosity alive, and for embracing a future filled with love, happiness, hope, and challenges.

TABLE OF CONTENTS

DEDICATION ii

ACKNOWLEDGEMENTS iii

LIST OF FIGURES vii

LIST OF TABLES ix

LIST OF APPENDICES x

ABSTRACT xi

CHAPTER

1 Introduction 1

1.1 Precision medicine for complex diseases 1
1.2 Drug response prediction in precision oncology 3
1.2.1 Importance and translation gap of cell lines 3
1.2.2 Drug response prediction methods and limitations 4
1.3 Mood dynamics in bipolar disorder 5
1.3.1 Clinical and conceptual background 5
1.3.2 Heterogeneity of mood dynamics in bipolar disorder 5
1.3.3 Methods for measuring heterogeneity in correlations 6
1.4 Dissertation outline 7

2 Pan-Cancer Drug Response Prediction Using Integrative Principal Component Regression 10

2.1 Introduction 10
2.2 Methods 14
2.2.1 Data formats 14
2.2.2 iPCR model 15
2.2.3 Model estimation 18
2.2.4 Gene co-expression network estimation 21
2.3 Simulation study 23
2.3.1 Simulation design and comparative analyses 23
2.3.2 Sensitivity analysis of rank selection for iPCR 26
2.4 Application to pan-cancer drug response data 26

2.4.1 Dataset description . . . 27
2.4.2 Patient drug response prediction . . . 28
2.4.3 Shared variation and biological interpretations . . . 30
2.4.4 Gene co-expression network analyses . . . 32
2.5 Discussion . . . 33

3 Time-Varying and Covariate-Dependent (TiVAC) Correlation . . . 35

3.1 Introduction . . . 35
3.2 TiVAC model . . . 38
3.3 Estimation . . . 39
3.3.1 Spline representation of functional covariate effects . . . 39
3.3.2 Spline coefficient estimation . . . 40
3.3.3 Simultaneous confidence band estimation . . . 42
3.4 Simulation study . . . 43
3.4.1 Simulations with deterministic correlation functions . . . 45
3.4.2 Impact of noise in correlation functions . . . 50
3.5 Application to mood trajectories in bipolar patients . . . 50
3.6 Discussion . . . 55

4 Generalized TiVAC Correlation for Zero-Inflated Non-Gaussian Data with Application to Bipolar Mood Dynamics . . . 56

4.1 Introduction . . . 56
4.2 Copula-TiVAC model . . . 60
4.2.1 Data structure and notation . . . 60
4.2.2 Preliminaries: copula framework . . . 60
4.2.3 Copula–TiVAC model . . . 62
4.3 Estimation . . . 64
4.3.1 Bayesian TiVAC: spline representation and priors . . . 64
4.3.2 Copula transformation to the latent Gaussian scale . . . 64
4.3.3 Posterior inference . . . 65
4.4 Simulation study . . . 68
4.5 Application to mood dynamics in bipolar disorder . . . 73
4.6 Discussion . . . 77

5 Conclusion . . . 79

5.1 Summary of proposed statistical methods . . . 79
5.2 Key insights from real-world analyses . . . 80
5.3 Future directions . . . 81

APPENDICES . . . 84

BIBLIOGRAPHY . . . 120

LIST OF FIGURES

FIGURE

1.1 Overview of dissertation methods and applications 3

2.1 Conceptual framework of the iPCR model. 13
2.2 Simulation results for drug response prediction. 25
2.3 Prediction results of iPCR for five drugs. 30
2.4 Biological interpretations of iPCR's results. 31
2.5 Analysis of model-specific and shared gene co-expression networks between cell lines and patients. 33

3.1 PHQ-9 and GAD-7 trajectories and correlation heterogeneity 36
3.2 Representative curve estimations in Scenario 1. 47
3.3 Representative heatmap estimations under $\mathscr{T}_{Moderate}$ setting in Scenario 2. . . . 49
3.4 Results of TiVAC coefficient functions on PrBP data. 52
3.5 Results of TiVAC correlations on PrBP data. 54

4.1 Visualization of mood scores in BP-I 57
4.2 Copula-TiVAC workflow. 59
4.3 Conceptual diagram of the zero-inflated Gaussian copula model. 62
4.4 RMSE of estimated correlations across zero proportions. 70
4.5 Representative correlation curve estimations under $\mathscr{T}_{\text{Real}}$ 71
4.6 Representative heatmap estimations under the multiple–covariate scenario ($\mathscr{T}_{\text{Real}}$) 72
4.7 Copula-TiVAC coefficient functions for PHQ-ASRM correlation 74
4.8 Copula-TiVAC correlations between PHQ-9 and ASRM 75

A.1 Sensitivity analysis of rank selection for iPCR. 91
A.2 Cell line-specific gene co-expression network. 93
A.3 Patient-specific gene co-expression network. 93
A.4 Number of hub genes. 94

B.1 Representative heatmap estimations under $\mathscr{T}_{Low}$ setting in Scenario 2. 100
B.2 Representative heatmap estimations under $\mathscr{T}_{High}$ setting in Scenario 2. 101
B.3 Representative heatmap estimations under $\mathscr{T}_{PrBP}$ setting in Scenario 2. 102
B.4 Empirical analysis on PrBP data. 107
B.5 TiVAC results of intercept-only model on PrBP data. 108

C.1 Additional simulation RMSE across zero proportions 110
C.2 Representative correlation curve estimations under $\mathscr{T}_{\text{Sparse}}$ 111

C.3 Representative heatmap estimations under the continuous–covariate scenario . . 112
C.4 Representative heatmap estimations under the multiple–covariate scenario ($\mathscr{T}_{\text{Sparse}}$)113
C.5 Copula-TiVAC coefficient functions for GAD-ASRM correlation 115
C.6 Copula-TiVAC coefficient functions for PHQ-GAD correlation 116
C.7 Copula-TiVAC correlations between GAD-7 and ASRM 117
C.8 Copula-TiVAC correlations between PHQ-9 and GAD-7 118
C.9 TiVAC correlations between PHQ-9 and ASRM 119

LIST OF TABLES

TABLE

2.1 Results for human drug response prediction. 29

3.1 Simulation results for Scenario 1 46
3.2 Simulation results for Scenario 2. 48

A.1 Summary of notation. 85
A.2 Rank selection results of iPCR models for drug response prediction in patients. 92
A.3 List of hub genes. 95

B.1 Additional simulation results of CoCoA for Scenario 1. 98
B.2 Additional simulation results of CoCoA for Scenario 2. 99
B.3 Simulation results for Scenario 3. 103
B.4 Additional simulation results of CoCoA for Scenario 3. 104
B.5 Simulation results for Scenario 4. 105
B.6 Additional simulation results of CoCoA for Scenario 4. 106
B.7 Descriptive statistics of the preprocessed PrBP dataset 107

C.1 Descriptive statistics across analysis cohorts 114

LIST OF APPENDICES

A Appendix Chapter 2 . . . . . **84**
Mathematical notations 84
Gene co-expression network estimation 86
Additional information of simulation studies 90
Additional results of pan-cancer application 92
B Appendix Chapter 3 . . . . . **96**
Estimation algorithm 96
Additional simulation studies 97
Additional application results 106
C Appendix Chapter 4 . . . . . **109**
Additional simulation results 109
Additional application results 114

ABSTRACT

One of the goals of precision medicine is to better understand heterogeneity in disease mechanisms and drug response across patients. These endeavors have been accelerated by the collection of large clinical and molecular datasets but also raises new analytic challenges. This dissertation develops AI-driven and advanced statistical methods for two problems in precision oncology and psychiatry. The first problem is drug response prediction in oncology, where models trained on cell lines face challenges when generalizing to patient tumors. The second problem is the characterization of mood dynamics in bipolar disorder, where correlations between depression, anxiety, and mania vary across patients and over time.

In Chapter 2, I present integrative principal component regression (iPCR), an AI-driven transfer learning model for pan-cancer drug response prediction. iPCR bridges the gap between patient tumors and preclinical cancer models such as cell lines by jointly modeling their genomic profiles. It uncovers both shared and model-specific structured variation in the genomic data of cell lines and patient tumors through matrix decompositions. The shared features are then used to predict patient drug response based on pharmacologic data measured in preclinical models. The model supports interpretation by identifying driver genes and pathways and by estimating shared and model-specific gene co-expression networks. Simulations and real data analyses show improved drug response prediction and biological interpretability.

In Chapter 3, I introduce the Time-Varying and Covariate-Dependent (TiVAC) correlation model under a bivariate Gaussian framework. The model represents covariate effects on pairwise correlations as smooth functions of time using B-splines. Parameters are estimated by penalized maximum likelihood with a Newton-Raphson algorithm. I also construct simultaneous confidence bands to support inference on time-varying covariate effects. Simulations show accurate estimation under a range of scenarios. An application to longitudinal data from bipolar I patients illustrates how clinical interventions and patient characteristics affect the correlation between depression and anxiety across age.

In Chapter 4, I generalize the TiVAC framework to handle zero-inflated and non-Gaussian outcomes. I develop a zero-inflated Gaussian-copula-based TiVAC (Copula-TiVAC) model that assumes an underlying latent Gaussian layer linked to the observed scores, and models

time-varying covariate effects on latent correlation with Bayesian P-splines. Estimation uses a blocked Gibbs sampler that cycles between updating latent Gaussian variables and updating spline coefficients with Hamiltonian Monte Carlo. Inference uses simultaneous confidence bands from joint posterior draws. Simulations demonstrate advantages over competing approaches on zero-inflated data. An application to about 200 bipolar I participants shows how medications and patient characteristics relate to the concurrent correlation between mania, depression, and anxiety over age. One promising finding is that greater medication exposure is associated with a decoupling of depression and mania and of anxiety and mania.

Overall, the methods in this dissertation provide novel statistical approaches for integrative modeling in precision oncology and dynamic correlation modeling in precision psychiatry.

CHAPTER 1

Introduction

1.1 Precision medicine for complex diseases

Clinical care is undergoing a profound transition as precision medicine moves from concept to routine practice, supported by rapid advances in data-acquisition technologies and analytic methods (Akhoon, 2021). Historically, clinical decisions were grounded in experience and pathophysiology and often followed a "one-size-fits-all" approach. In contrast, precision medicine tailors prevention, diagnosis, and treatment to each person's health status, accounting for differences in genes, environment, and lifestyle (Marques et al., 2024). This goal is commonly summarized by the "five rights": the right drug for the right patient, at the right time, in the right dose, and via the right route (Grissinger, 2010). Launched in 1990, the Human Genome Project accelerated the shift to precision medicine by enabling affordable DNA sequencing and targeted therapies that revolutionized health care(Gibbs, 2020). Today, large-scale biomedical datasets—such as electronic health records and diverse omics profiles—are being generated rapidly (Marques et al., 2024). To turn these complex data into a clearer understanding of patient heterogeneity in disease mechanisms and into actionable guidance for individual care, there is a critical need for new computational approaches grounded in statistics and machine learning (Denny and Collins, 2021; Subbiah, 2023).

Precision Oncology. Cancer is a complex disease that differs widely across and within tumor types. Tumors that appear similar on pathology can carry distinct molecular profiles, so patients with the same diagnosis often respond differently to the same therapy (Ely, 2009; Arjmand et al., 2017). Large genomic and pharmacologic profiling efforts across human tumors and biological model systems now make it possible to study these differences at scale. Patient tumor resources, including The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013), The International Cancer Genome Consortium (ICGC) (Zhang et al., 2011), and The Cancer Proteome Atlas (TCPA) (Li et al., 2013), provide multi-platform ("multi-omics")

measurements for thousands of samples spanning >30 cancer types. Also, the resources of biological model systems, including the Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012) and the MD Anderson Cell Lines Project (MCLP) (Li et al., 2017), curate multi-omics profiles for $>1{,}000$ cell lines. In addition, drug-perturbation programs such as GDSC (Yang et al., 2013), DepMap (Tsherniak et al., 2017), LINCS (Subramanian et al., 2017), and the Connectivity Map (Subramanian et al., 2017) systematically characterize responses to a wide range of anti-cancer agents. Taken together, these datasets enable discovery of prognostic and diagnostic biomarkers and have accelerated targeted and precision therapies informed by genome-level heterogeneity (Jiang and Wang, 2010; McCarthy et al., 2013; Kumar-Sinha and Chinnaiyan, 2018; Liu et al., 2019; Bhattacharyya et al., 2020).

Precision Psychiatry. Unlike oncology, where diagnosis, prognosis, and treatment already benefit from precision medicine, psychiatry is newer yet increasingly important in this shift (Fernandes et al., 2017). Progress has been slower because the brain is more complex and dynamic than other organs. In addition, environment and experience interact with biology and create wide heterogeneity in symptoms across patients. Because of this complexity, the current focus of precision psychiatry is less on developing entirely new molecular compounds and more on improving classification, prognosis, and treatment selection (Fernandes et al., 2017). To address these challenges, large cohorts – such as the Heinz C. Prechter Bipolar Research Project (PrBP) (Yocum et al., 2023; McInnis et al., 2017), the Genomic Psychiatry Cohort (GPC) (Pato et al., 2013), GNC (German National Cohort (GNC) Consortium, 2014), the Systematic Treatment Enhancement Program for Bipolar Disorder (STEP-BD) (Sachs et al., 2003), and the Course and Outcome of Bipolar Youth (COBY) (Birmaher et al., 2009) – provide rich clinical and patient-characteristic data. With these data, researchers can potentially identify biomarkers that guide treatment choice and predict response to common agents, including antidepressants and antipsychotics (Fernandes et al., 2017). For example, the Clinical Pharmacogenetics Implementation Consortium (CPIC) recommends genetic screening before prescribing certain drugs to support appropriate metabolism (Alchakee et al., 2022). Another key direction is to identify mechanisms that underlie psychiatric symptoms, which can improve patient classification and guide existing treatments by disease class (Fernandes et al., 2017).

In this dissertation, as illustrated in Figure 1.1, I focus on two separate application problems in precision medicine: (1) predicting patient drug response in oncology and (2) characterizing the heterogeneity of mood dynamics in bipolar disorder. The next two sections provide the application and methodological background.

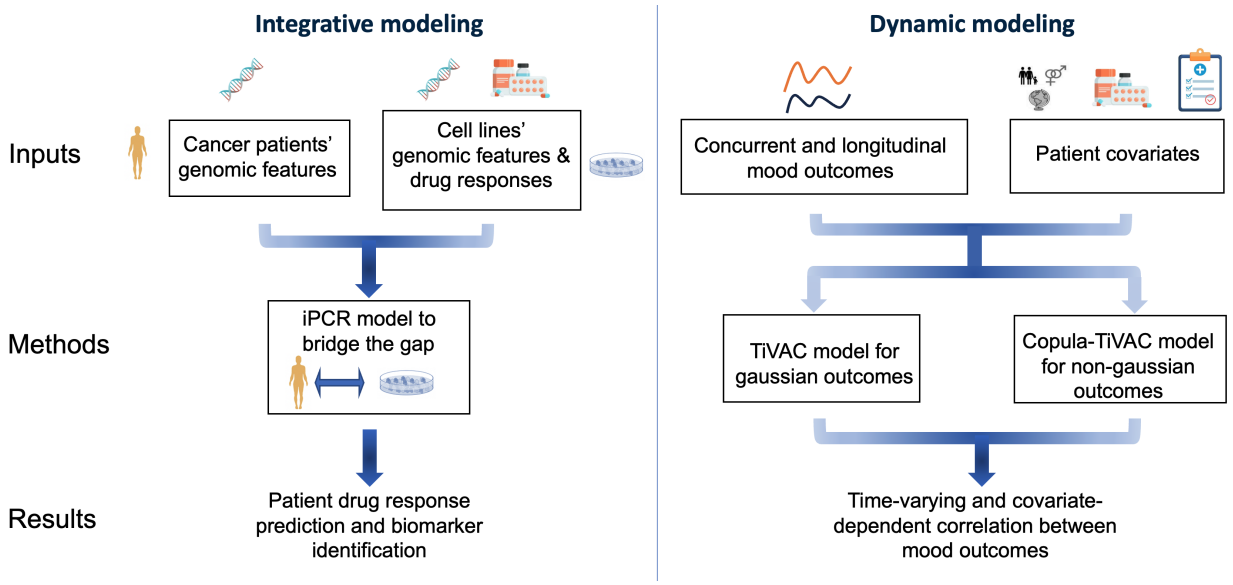


Figure 1.1: **Overview of dissertation methods and applications.** Left panel: Integrative modeling for precision oncology. Inputs are patients' genomic features and cell line genomic features with drug responses. The iPCR method jointly models these data to bridge the gap between cell lines and patients and supports patient drug response prediction and biomarker identification. Right panel: Dynamic modeling for precision psychiatry. Inputs include concurrent and longitudinal mood outcomes and patient covariates. The TiVAC model addresses Gaussian outcomes, and the Copula-TiVAC model generalizes to zero-inflated and non-Gaussian outcomes, yielding time-varying and covariate-dependent correlations between mood outcomes.

1.2 Drug response prediction in precision oncology

1.2.1 Importance and translation gap of cell lines

Cell lines are derived from human tumors and grown in culture for oncology studies (Sharma et al., 2010). They have long served as foundational models in cancer research and drug development, and continue to play a central role in the era of precision oncology. These immortalized *in vitro* systems provide a simplified and controllable platform for mechanistic studies and high-throughput drug screening. For example, large and well-characterized public cell line resources (e.g., CCLE, GDSC) provide rich genomic and transcriptomic profiles together with pharmacologic response measurements, enabling systematic discovery of genotype–phenotype associations and candidate biomarkers of drug response (Sharma et al., 2010; Barretina et al., 2012; Yang et al., 2013). Cell line assays are also faster and less expensive than animal studies, human clinical trials, or large cohort studies, supporting rapid and scalable experimentation (Kaur and Dufour, 2012a).

However, cancer cell lines often poorly predict human responses to therapy, since intrinsic differences between these biological models and patient tumors limit translation from bench to bedside (Ben-David et al., 2018; Gillet et al., 2011, 2013; Brubaker and Lauffenburger, 2020). From a genomics perspective, prolonged culture can lead to genetic drift, with new mutations and traits that reduce similarity to the original tumors (Gillet et al., 2011). This challenge highlights the need to systematically compare preclinical models with patient tumors to identify their shared and distinct genomic features. Integrative statistical models that jointly analyze genomic profiles from cell lines and patient tumors can help bridge this translational gap, enabling more effective use of cell lines to improve target discovery and drug response prediction in human cancer.

1.2.2 Drug response prediction methods and limitations

Past work has advanced biomarker discovery and drug response prediction using cell lines as the main training resource. Predictive models were trained on cell line genomic and drug-perturbation data and tested on out-of-sample cell lines or patient tumors (Falgreen et al., 2015; Zhang et al., 2015; Iorio et al., 2016). Public resources such as CCLE and GDSC provide high-throughput drug response data across hundreds of cell lines for many drugs, whereas patient datasets are much smaller and costly to generate (Costello et al., 2014). To reduce the gap between cell lines and patients, some methods incorporated patient tumor genomics during training and treated cross-platform differences as batch effects (Geeleher et al., 2014, 2017; Sakellaropoulos et al., 2019). More recent studies applied domain adaptation (a subcategory of transfer learning) and factor analysis to learn a latent space shared by cell lines and patients (Mourragui et al., 2019; Sharifi-Noghabi et al., 2020; Mourragui et al., 2021; Saha et al., 2022). Despite these advances, *in vitro* responses do not fully capture *in vivo* behavior in primary tumors (Kaur and Dufour, 2012b; van Staveren et al., 2009). Models trained only on cell lines face a domain shift when applied to patients because training and testing come from different biological systems (Saha et al., 2022). Methods that integrate patient and cell line data by aligning domains can help. However, these methods do not explicitly assess system-specific structured variation in gene expression (from either cell lines or patients) during model construction. If a model assumes only shared latent structure and residual noise, system-specific information may be absorbed into these components. Including this system-specific (individual) structure can potentially improve interpretation of genomic differences between cell lines and patients, and may increase prediction accuracy.

1.3 Mood dynamics in bipolar disorder

1.3.1 Clinical and conceptual background

In psychiatry, three main mood dimensions are depression, anxiety, and mania. These guide symptom evaluation, diagnosis, and treatment. Depression is defined as a persistent low mood and loss of interest or pleasure in usual activities, and it is often measured with PHQ-9 score (Kroenke et al., 2001). Anxiety refers to excessive worry and fear, often accompanied by physical tension, and it is commonly evaluated with GAD-7 score (Spitzer et al., 2006). Mania is characterized by elevated or irritable mood together with increased energy and activity, and it is frequently rated using ASRM score (Altman et al., 1997). Although these states are clinically distinct, they often overlap or interact in practice, which results in complex emotional patterns (Nunes et al., 2022). In this dissertation, I focus on dynamic statistical modeling to examine mood dynamics in bipolar disorder and to assess how patient characteristics and clinical factors influence these dynamics.

Bipolar disorder is a chronic and complex psychiatric illness that affects about 2.4% of the global population (Bonnín et al., 2019). It is defined by recurrent episodes of mania or hypomania together with major depression. In bipolar I disorder (BP-I), patients experience at least one manic episode, and depressive episodes are common. In bipolar II disorder, patients experience at least one hypomanic episode and one major depressive episode, but they do not experience a full manic episode (Black and Grant, 2014). As a lifelong and recurrent condition, bipolar disorder is strongly associated with reduced daily functioning, cognitive impairment, and a substantial decrease in quality of life (Bonnín et al., 2019). Its causes are still not fully understood. Current evidence suggests that many genetic variants with small effects, together with environmental risks such as childhood abuse and long-term stress, contribute to the development of the disorder (Anderson et al., 2012). Beyond the causes, the course of mood symptoms shows complex and heterogeneous dynamics (Cochran et al., 2018). Relative to unipolar depression or mania, bipolar disorder, particularly BP-I, presents greater challenges for clinicians in selecting proper treatment strategies to manage shifting and unpredictable mood states.

1.3.2 Heterogeneity of mood dynamics in bipolar disorder

Traditionally, bipolar mood has been viewed as either (i) one-dimensional, where increases in mania naturally align with decreases in depression and vice versa, or (ii) multistable, where the illness tends to settle into distinct, persistent states (e.g., mania and depression) (Cochran et al., 2018). Statistically, both views imply a negative concurrent correlation between mania

and depression. However, a recent analysis of the BP-I cohort (Cochran et al., 2018) found that the concurrent correlation between mania and depression can be negative or positive across individuals, with values ranging from about -0.60 to 0.75, revealing a more complex pattern. A positive concurrent correlation aligns with mixed states or episodes, as formally defined in DSM-5 (Muneer, 2017). Clinically, mixed states correspond to dysphoric mania or agitated depression, where patients feel depressed yet activated or irritable; such states often carry higher risks of suicidality (Muneer, 2017). Anxiety adds another layer of complexity because comorbid anxiety disorders are common in bipolar disorder and can interact with depressive and manic symptoms over time (Spoorthy et al., 2019). Moreover, heterogeneity is present not only between patients but also within individuals, as mood symptom profiles are dynamic and can change with age in BP-I (Coryell et al., 2009). However, few studies examine which patient characteristics and clinical factors are associated with this heterogeneity in pairwise correlation between mania, depression and anxiety. This gap motivates the statistical question of treating the correlation (ρ) itself as a function of both time (t) and covariates ($\mathbf{x}$)–that is, $\rho(t, \mathbf{x})$– rather than assuming it is fixed at the population level.

1.3.3 Methods for measuring heterogeneity in correlations

We first review methods that model heterogeneity in correlations as a function of covariates. Simpson's paradox shows that subgroup patterns can reverse after aggregation, which motivates covariate-dependent correlations (Blyth, 1972). Early work treated correlation as a function of a discrete third variable but required large subgroup sizes (Bartlett, 1993). Later approaches allowed both continuous and discrete covariates in a bivariate Gaussian framework, with estimation via restricted maximum likelihood and via second order generalized estimating equations (Wilding et al., 2011; Ho et al., 2010). More recently, Tu et al. (2022) introduced the CoCoA framework for covariate-dependent pairwise correlation. Other methods model covariate-dependent covariance or correlation for more than two outcomes through parameterizations or transformations that ensure the covariance or correlation matrix remains positive definite (Pourahmadi, 1999; Chiu et al., 1996; Hoff and Niu, 2012; Archakov and Hansen, 2021). These approaches capture covariate effects but usually treat correlation as fixed in time. Note that simply using time as one covariate imposes a linear trend and does not allow time-varying covariate effects.

In addition, many biomedical outcomes deviate from Gaussian assumptions and often show both zero inflation and heavy skewness. For example. in the Heinz C. Prechter Bipolar Research Project cohort (McInnis et al., 2017; Yocum et al., 2023), PHQ-9 depression scores and GAD-7 anxiety scores contain about 15% zeros, whereas ASRM mania scores exhibit

about 40% zeros and show a long right tail. Similar zero inflation appears in single-cell RNA-sequencing due to dropout (Van den Berge et al., 2018). Simple monotone transformations, such as logarithmic transforms, do not remove the discrete mass at zero. Applying Gaussian-based correlation models, such as CoCoA, to such data risks biased estimates and misleading inference.

To handle zero-inflated and non-Gaussian data, several alternative approaches have emerged. Conditional copula methods use kernel-based weights to model covariate dependence with flexibility, but they offer limited direct inference on how specific covariates affect correlation (Gijbels et al., 2011, 2012). The zero-inflated negative binomial dynamic correlation (ZENCO) model links zero-inflated counts to latent Gaussian variables to study the dependence structure, but it relies on strict distributional assumptions, such as a negative binomial component, that may not fit observed data (Yang and Ho, 2021). Multivariate hurdle models are related but address a different question. They study conditional correlations in zero-inflated graphical models, where one outcome is modeled given the others, while our target is the pairwise correlation between two outcomes given covariates (McDavid et al., 2019). Importantly, all of these models are designed for static settings and cannot flexibly capture time-varying correlation, which is essential for understanding mood dynamics in bipolar I.

There are time-varying dependence models such as DCC-GARCH (Engle, 2002), DSEM (Asparouhov et al., 2017), and idPAC (Kundu et al., 2021). Though these approaches track change over time, none directly interpret how covariates affect correlations over time and provide statistical inference on these time-varying effects.

1.4 Dissertation outline

To advance the two problems in oncology and psychiatry described above, I develop new AI-driven and advanced statistical methods that address the key analytic challenges and gaps. Chapter 2 focuses on predicting patient drug response in oncology. Chapter 3 develops a time-varying and covariate-dependent correlation model. Chapter 4 generalizes the Chapter 3 model to handle zero-inflated mood scores and uses it to study mood dynamics in BP-I patients.

In Chapter 2, as illustrated in the left panel of Figure 1.1, I develop integrative principal component regression (iPCR), an AI-driven transfer learning model, to bridge the gap between preclinical cancer models and patient tumors for drug-response prediction. iPCR uncovers both joint and model-specific structured variations in the genomic data of cell lines and patient tumors through matrix decompositions. The extracted shared features from

joint variation is then used to predict patient drug responses based on the pharmacological data of preclinical models. Moreover, the interpretability of our model allows for the identification of key driver genes and pathways associated with the treatment-specific response in patients across multiple cancers. I demonstrate that the outputs of the iPCR model can assist in inferring both model-specific and shared co-expression networks between cell lines and patients. I show that iPCR performs favorably compared to competing approaches in predicting patient drug responses, in both simulation studies and real-world applications, in addition to identifying key genomic drivers of cancer drug responses.

In Chapter 3, as illustrated in the right panel of Figure 1.1, I propose the Time-Varying and Covariate-Dependent (TiVAC) correlation model under a bivariate Gaussian framework. The model captures covariate effects on pairwise correlation as smooth functions of time using B-splines, accommodating both binary and continuous covariates. The model parameters are estimated via penalized maximum likelihood using a Newton–Raphson algorithm. I further construct simultaneous confidence bands to facilitate rigorous inference on the time-varying covariate effects. Simulation studies show that TiVAC outperforms existing methods in accurately estimating correlations across diverse scenarios, including both correctly specified and misspecified correlation functions. Finally, I apply TiVAC to longitudinal mood data from BP-I patients, illustrating how clinical interventions and patient characteristics influence the correlation between depression and anxiety across ages.

In Chapter 4, I generalize TiVAC to handle zero-inflated, non-Gaussian biomedical outcomes, enabling a deeper and precise examination of mood dynamics in BP-I where mood scores often contain excess zeros and heavy right tails. I introduce a zero-inflated Gaussian-copula–based TiVAC (Copula-TiVAC) model to estimate the concurrent correlation between two longitudinal outcomes and how it varies with covariates and over time. Copula-TiVAC employs a truncated Gaussian copula model (Yoon et al., 2020), assuming latent Gaussian variables that induce the observed non-Gaussian, zero-inflated outcomes and fully capture their dependence, providing flexibility without restrictive distributional assumptions. On the latent bivariate normal scale, I mirror the TiVAC framework and model the time-varying covaraite effects with Bayesian P-splines (Lang and Brezger, 2004). Estimation proceeds via a blocked Gibbs sampler that iterates between (i) a copula stage updating latent Gaussian variables and zero thresholds given the current TiVAC correlation estimation, and (ii) a TiVAC stage updating spline coefficients using Hamiltonian Monte Carlo. Inference on time-varying covariate effects is based on simultaneous confidence bands constructed from joint MCMC draws. Through simulations, I compare Copula-TiVAC against competing approaches and show advantages across diverse settings. Finally, I apply Copula-TiVAC to longitudinal mood trajectories in 200 BP-I participants, focusing on

pairwise correlations between mania, depression, and anxiety. A key finding is that greater exposure to psychiatric medications is associated with a decoupling of depression–mania and anxiety–mania, consistent with reduced risk of mixed episodes in BP-I.

In Chapter 5, I conclude by bringing together the methodological contributions, highlighting insights from the real-world applications, and discussing potential directions for future research.

Although these methods were developed for two specific applications in precision oncology and psychiatry, they are broadly applicable to other domains, provided the data structure and inferential goals align. Code and software are available in the GitHub repositories iPCR, TiVAC, and Copula–TiVAC. For the public data used in Chapter 2, sources are specified, whereas the cohort data used for Chapters 3 and 4 are used under license and are available with permission from the Heinz C. Prechter Bipolar Research Program.

CHAPTER 2

Pan-Cancer Drug Response Prediction Using Integrative Principal Component Regression

2.1 Introduction

Cancer is a complex disease at multiple levels. Due to inherent tumor complexity and molecular heterogeneity, it is well evidenced that patients with similar pathological and diagnostic cancer phenotypes often respond differently to the same treatment (Ely, 2009; Arjmand et al., 2017). Thus, this has spurned efforts to transition from traditional one-size-fits-all approach to precision or personalized medicine – allowing clinicians to match the right drug to the right patient (Chae et al., 2017). Precision medicine and targeted therapies are progressing rapidly as scientists develop a deeper understanding of patients' genome-level heterogeneity as well as prognostic and diagnostic biomarkers (Jiang and Wang, 2010; McCarthy et al., 2013; Kumar-Sinha and Chinnaiyan, 2018; Liu et al., 2019; Bhattacharyya et al., 2020). There are two related axes of investigation that have contributed to these efforts: pre-clinical model systems and high-throughput molecular profiling techniques; as we expound below.

Cell lines are samples taken from human tumors and cultivated in labs, serve as fundamental preclinical models for studying the pathological mechanisms and treatment efficacy of cancer (Sharma et al., 2010). They are pivotal because they offer a controlled approach to discover underlying biological mechanisms and hence, effective cancer treatments (Sharma et al., 2010). However, accumulating evidence suggests that inherent biological differences between preclinical models and patient tumors systematically hinder translational research efforts – from bench-side knowledge to its use at the clinical bedside (Ben-David et al., 2018; Gillet et al., 2011, 2013; Brubaker and Lauffenburger, 2020). Specifically, from a genomics perspective, cell lines can genetically drift over prolonged culture periods, potentially adopting mutations and traits that render them less representative of the original patient tumors (Gillet et al., 2011). This challenge underscores the need for a systematic and comprehensive

examination of the similarities and differences between preclinical and clinical (i.e. patient tumors) model systems to uncover novel drug targets and prediction models in cancer.

The advances in large-scale genomic and pharmacological profiling across patient tumors and preclinical models make this investigation possible. The sources of patient tumor datasets such as The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013), the International Cancer Genome Consortium (ICGC) (Zhang et al., 2011) and The Cancer Proteome Atlas (TCPA) (Li et al., 2013) have provided extensive multi-platform genomics (multi-omics) profiles of thousands of cancer and normal samples spanning > 30 cancer types. The data resources of preclinical model systems, including Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012), and MD Anderson Cell Lines Project (MCLP) (Li et al., 2017) contain multi-omics data of $> 1,000$ cell lines. The large-scale drug perturbation studies of biological model systems such as Genomics of Drug Sensitivity in Cancer (GDSC) (Yang et al., 2013), Dependency Map (DEPMAP) (Tsherniak et al., 2017), Library of Integrated Cellular Signatures (LINCS) (Subramanian et al., 2017), and Connectivity Map (CMAP) (Subramanian et al., 2017) have characterized their responses to a wide range of anti-cancer therapeutics. These datasets enable the assessment of the translational potential of cell lines to patient tumors, by identifying their differential and shared genomic characterization and its effect in drug response prediction.

To this end, computational and clinical researchers have made significant progress in therapeutic biomarker detection and drug response prediction by using cell lines alone. In the context of biomarker detection, high-throughput genomic studies have examined the correlation between drug response with mutations and copy number alterations based on cell lines (Williams and McDermott, 2017). With respect to drug response prediction, statistical models have been trained solely on cell lines' genomic and drug perturbation data, and tested on out-of-sample data from cell lines or patient tumors (Falgreen et al., 2015; Zhang et al., 2015; Iorio et al., 2016). The data from cell lines have been widely used for model training because high-throughput drug response data across hundreds of cell lines for many drugs are publicly available in large databases, such as the aforementioned CCLE and GDSC databases. On the other hand, since clinical data of patients are often very expensive and time-consuming to generate, these data are still relatively small in sample size and present challenges to be used as reliable training data for drug response prediction (Costello et al., 2014).

Even though cell lines have been typically used to train drug response prediction models, the *in vitro* response from cell lines is not expected to adequately represent *in vivo* response from primary patient tumors (Kaur and Dufour, 2012b; van Staveren et al., 2009). The main reason is that cell lines evolve quite differently from patient tumors in the absence of the

complex *in vivo* microenvironment, and genotypic and phenotypic variation in cell lines can happen over a large number of passages (Goodspeed et al., 2016; Brodaczewska et al., 2016). Thus, given *in vivo* drug response prediction as the primary goal, the variation between cell lines and patient tumors makes the models, which are trained solely on cell lines, suffer from the fundamental problem that training and test sets come from two different biological model systems (Saha et al., 2022).

To mitigate the effect of the differences between biological model systems on drug response prediction, genomic data from patients have been used in model training to integrate genomic features between cell lines and patients. Previous drug response prediction methods (Geeleher et al., 2014, 2017; Sakellaropoulos et al., 2019) assumed that batch effects were the main source of differences between cell lines and patients' gene expression data. More recently, some methods (Mourragui et al., 2019; Sharifi-Noghabi et al., 2020; Mourragui et al., 2021; Saha et al., 2022) have employed domain adaptation technique, a subcategory of transfer learning, and factor analysis to bridge the gap between preclinical models and patient tumors, by identifying a shared latent space common to both the cell lines' and patients' domains. However, these methods do not explicitly assess the model-specific structured variation in gene expression data (pertaining to either cell lines or patients) during model construction. In such instances, if the modeling of gene expression data for cell lines and patients solely comprises shared latent structure and residual noise, the model-specific information could be incorporated into the estimated shared latent structure and/or residual matrix. Including the model-specific information, also called individual structures, in the model could enhance the interpretation of genomic differences between cell lines and patients, as well as potentially increase prediction accuracy.

To this end, we propose an integrative principal component regression (iPCR) model that serves four key purposes: 1) to quantify the similarities and differences in gene expression data between cell lines and patient tumors, 2) to use the patient-cell line similarities to predict the patient drug response, 3) to find key genomic drivers (e.g., pathways, gene sets) for patient drug-specific response, and 4) to infer model-specific and shared co-expression networks between cell lines and patients and detect hub genes in them. Viewed as a domain adaptation model (a subset of transfer learning), our innovative approach not only adapts multi-view statistical techniques to address the unique challenges of multi-system prediction problem in precision oncology, but also comprehensively investigates the common biology and differences between cell lines and patient tumors. This approach provides enhanced interpretability of both gene effects on drug response and gene-gene interactions, offering a fresh perspective in this critical domain. Figure 2.1 shows the overall conceptual flow of our pipeline. In stage I of iPCR model (Figure 2.1A), we integrate gene expression data between

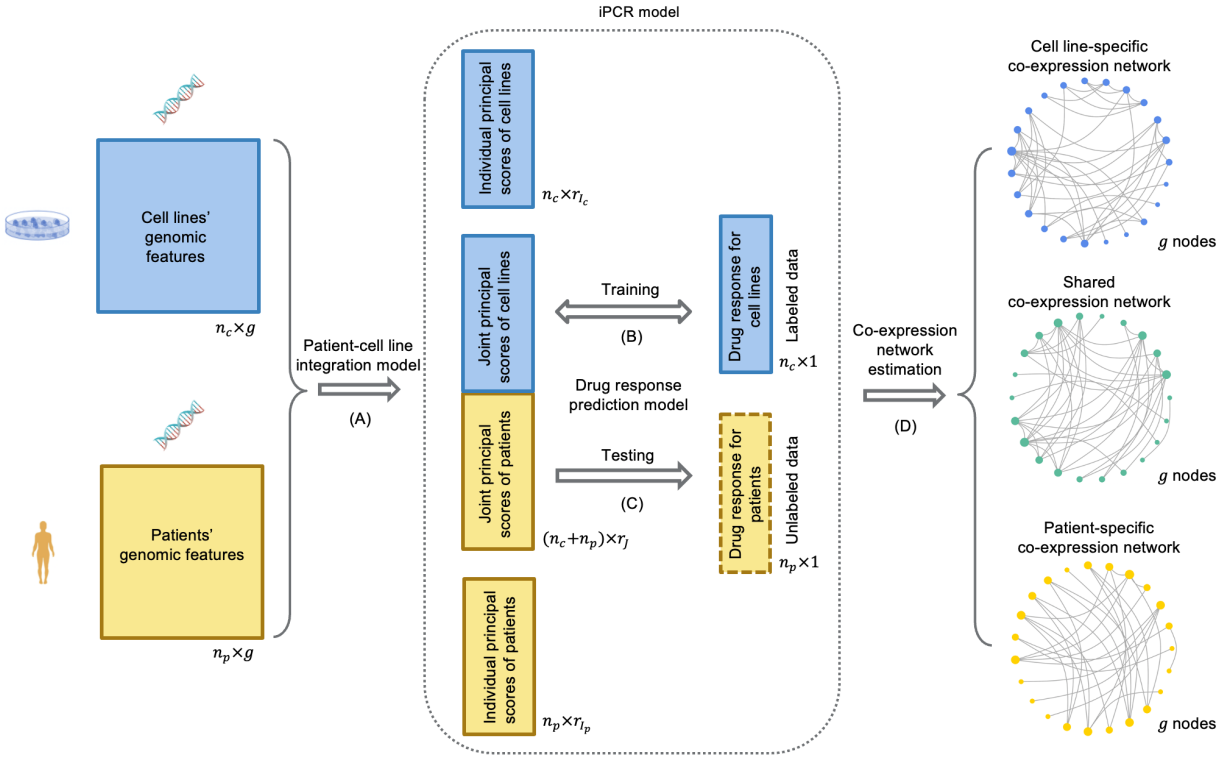


Figure 2.1: **Conceptual framework of the iPCR model.** The joint model is specified by two sub-models, namely the patient-cell line integration model (A) specifying the biological similarity (joint principal scores) and difference (individual principal scores) between cell lines and patients, and the drug response prediction model (B and C) specifying the relationship between drug response and the joint principal scores of cell lines and patients, which is trained on the labeled cell lines' data and tested on the unlabeled patients' data. The results of iPCR model are then used to infer model-specific and shared co-expression networks through co-expression network estimation (D).

cell lines and patients, and identify latent shared features (i.e., joint principal scores) and model-specific features (i.e., individual principal scores). These latent features essentially capture the underlying shared and model-specific biological process across the two model systems. In stage II of iPCR model (Figure 2.1B-C), the shared features from stage I and drug responses of cell lines are used to train a penalized regression model for the prediction of patient drug responses. Finally, the joint and individual variations across the two biological model systems detected by the iPCR model are used to capture the model-specific and shared gene co-expression networks between the biological model systems.

We demonstrate that iPCR outperforms alternative approaches in patient drug response prediction, under multiple simulated scenarios and multiple cancer drugs in a pan-cancer application. By implementing our iPCR and gene co-expression network decomposition methods, we identify key genomic drivers (e.g., pathways, gene sets) of patient drug-specific responses, and hub genes in the shared co-expression network between cell lines and patients.

The rest of this manuscript is organized as follows: Section 2.2.1 introduces the formats of genomic and drug response data used in our study. Section 2.2.2.1 formulates the stage I of iPCR, which employs a dimension reduction technique to decompose multi-system genomic data of cell lines and patients into shared and individual components. Section 2.2.2.2 describes the stage II of iPCR, which utilizes the outputs from stage I to predict the drug response in patients based on cell lines' drug response data. Section 2.2.3 delineates the details of iPCR model estimation. Section 2.2.4 introduces the gene co-expression network decomposition method based on the outputs from the iPCR model. Section 2.3 evaluates the prediction performance of iPCR and its sensitivity to rank selection through simulation studies. In Section 2.4, we apply iPCR and gene co-expression network decomposition methods on large-scale pan-cancer data from cell lines and patient tumors. We conclude with a discussion in Section 2.5

2.2 Methods

2.2.1 Data formats

Following the conceptual framework in Figure 2.1, the specific data formats are defined as follows. We consider two gene expression matrices $\mathbf{C} \in \mathbb{R}^{n_c \times g}$ from n_c cell lines with g gene features, and $\mathbf{P} \in \mathbb{R}^{n_p \times g}$ from n_p patients with the same set of g gene features. Then, these two matrices can be combined into a single data matrix $\mathbf{G} = \left[\mathbf{C}^\top, \mathbf{P}^\top\right]^\top_{n \times g}$, where $n = n_c + n_p$ is the total number of samples. Following standard preprocessing in Lock et al. (2013); Mourragui et al. (2021), we performed a centralization and standardization for each

column of $\mathbf{C}$ and $\mathbf{P}$.

For a given drug D assayed on n_c cell lines, we denote vector $\mathbf{D}_c = (D_c^1, D_c^2, \ldots, D_c^{n_c})^T \in \mathbb{R}^{n_c}$ as the observed continuous drug responses for those cell lines. For the same drug taken by N_p patients, the vector $\mathbf{D}_p^* \in \mathbb{R}^{n_p}$ contains the drug responses for those patients, where $(*)$ represents the unobserved labels. Our primary objective is to predict unobserved patient drug responses $\mathbf{D}_p^*$ based on the available datasets $\{\mathbf{C}, \mathbf{P}, \mathbf{D}_c\}$. In this paper, $\mathbf{D}_c$ is measured by the area under the drug response curve (drug-AUC), although other omic platforms (e.g., proteomics, metabolomics) and drug response metrics (e.g., IC50) can easily be used in our framework. To enhance clarity of heavy notations throughout the entire paper, we primarily used the first letter of each concept for our notations. For instance, $\mathbf{D}_c$ and $\mathbf{D}_p^*$ represent the drug data of cell lines and patients, respectively. A comprehensive table of these notations is provided in Table A.1.

2.2.2 iPCR model

In this section, we elaborate on the two stages of our iPCR model, which is designed to enable the transfer of genomic information from cell lines to patients, thereby enhancing the accuracy of drug response predictions. First, we develop an integrative model for patient-cell line data based on variation decomposition. This approach is crucial for identifying shared patterns of variation (termed joint structure) and model-specific variation (referred to as individual structure) within the gene expression data of cell lines and patients. Subsequently, we introduce the drug response prediction model, which leverages the latent shared features between patients and cell lines from the joint structure to predict drug responses in patients and to identify key therapeutic biomarkers.

2.2.2.1 Stage I: patient-cell line integration model

Our proposed patient-cell line data integration model assumes the existence of both (genomically driven) biological similarity and difference between the two model systems. Based on this assumption, the model decomposes $\mathbf{C}$ and $\mathbf{P}$ into a low-rank joint structure that captures the common features between cell lines and patients; two low-rank individual structures that preserve individual features that are not shared, and the residual noise. The model is formalized as

$$\begin{bmatrix} \mathbf{C} \\ \mathbf{P} \end{bmatrix} = \begin{bmatrix} \boldsymbol{J}_c \\ \boldsymbol{J}_p \end{bmatrix} + \begin{bmatrix} \boldsymbol{I}_c \\ \boldsymbol{I}_p \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_c \\ \boldsymbol{\varepsilon}_p \end{bmatrix}, \tag{2.1}$$

where for all $i \in \{c, p\}$, $\boldsymbol{J}_i$ is the sub-matrix of the joint structure matrix $\boldsymbol{J} = [\boldsymbol{J}_c^\top, \boldsymbol{J}_p^\top]^\top$ with rank r_J, $\boldsymbol{I}_i$ is the individual structure matrix with rank r_{I_i}, and $\boldsymbol{\varepsilon_i}$ is the error matrix containing the independent entries with $\mathbb{E}(\boldsymbol{\varepsilon}_i) = \mathbf{0}_{n_i \times g}$. We denote r_J as joint rank, r_{I_c} as individual rank of cell lines, and r_{I_p} as individual rank of patients.

Model 2.1 necessitates the conditions detailed in Proposition 1 for unique identifiability, as follows:

Proposition 1. *(Lock et al., 2013; Gaynanova and Li, 2019) Let row(·) and col(·) be the column and row space of a matrix, respectively. Given any set of matrices* $\{\boldsymbol{G}_i,\ i{=}c,\ p\}$*, there are unique sets of matrices* $\{\boldsymbol{J}_i,\ i = c,\ p\}$ *and* $\{\boldsymbol{I}_i,\ i = c,\ p\}$ *such that*

(1) $\boldsymbol{G}_i = \boldsymbol{J}_i + \boldsymbol{I}_i$ *for* $i = c,\ p$

(2) $col(\boldsymbol{J}_i) \cap col(\boldsymbol{I}_i) = 0$ *for* $i = c,\ p$

(3) $row(\boldsymbol{J}_i) = row(\boldsymbol{J})$ *for* $i = c,\ p$*, where* $\boldsymbol{J} = \begin{bmatrix} \boldsymbol{J}_c \\ \boldsymbol{J}_p \end{bmatrix}$

(4) $row(\boldsymbol{J}_i) \perp row(\boldsymbol{I}_i)$ *for* $i = c,\ p$

(5) $row(\boldsymbol{I}_c) \cap row(\boldsymbol{I}_p) = 0$

Interpretation of Model 2.1 based on Proposition 1. Intuitively, these conditions suggest that gene expression data from either cell lines or patients can be decomposed into a joint structure and an individual structure, excluding noise (Condition 1). Moreover, the data from cell lines and patients share a latent common feature space within the joint structure (Condition 3), and this shared feature space is linearly independent of the individual feature space of either patients or cell lines (Condition 4). Additionally, the individual structure of patients doesn't share a feature space with the individual structure of cell lines (Condition 5). This aligns well with our assumption about the existing similarities and differences in gene expression data between cell lines and patients. As mentioned in Gaynanova and Li (2019), Condition 2 on the column spaces is equivalent to the rank condition in Lock et al. (2013), which can be expressed as $rank(\boldsymbol{J}_i + \boldsymbol{I}_i) = rank(\boldsymbol{J}_i) + rank(\boldsymbol{I}_i)$ for $i = c, p$. This condition is necessary for the uniqueness of the decomposition at Stage I, in conjunction with Condition 4.

To fit the various components of Model 2.1, we employ PCA or singular value decomposition (SVD) on the joint structure $\boldsymbol{J}$, as well as the two individual structures $\boldsymbol{I}_c$ and $\boldsymbol{I}_p$. To this end, Model 2.1 can be further factorized as follows:

$$\begin{bmatrix} \mathbf{C} \\ \mathbf{P} \end{bmatrix} = \begin{bmatrix} \boldsymbol{S}_{J_c} \\ \boldsymbol{S}_{J_p} \end{bmatrix} \boldsymbol{V}_J^\top + \begin{bmatrix} \boldsymbol{S}_{I_c} \boldsymbol{V}_{I_c}^\top \\ \boldsymbol{S}_{I_p} \boldsymbol{V}_{I_p}^\top \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_c \\ \boldsymbol{\varepsilon}_p \end{bmatrix}. \tag{2.2}$$

Interpretation of Model 1 based on its factorization The joint score matrix $\boldsymbol{S}_J = [\boldsymbol{S}_{J_c}^\top, \boldsymbol{S}_{J_p}^\top]^\top \in \mathbb{R}^{n \times r_J}$ contains r_J joint principal scores from $\boldsymbol{J}$, which can be viewed as the latent shared features of cell lines and patients, capturing their biological similarity. The orthonormal joint loading matrix $\boldsymbol{V}_J \in \mathbb{R}^{g \times r_J}$ quantifies the extent to which each original feature in $\boldsymbol{J}$ is related to the latent shared features, such that $\boldsymbol{V}_J^\top \boldsymbol{V}_J = \mathbb{I}_{r_J \times r_J}$. From another angle, the columns of $\boldsymbol{V}_J$ (i.e., shared principal axes) define an r_J-dimensional shared feature space between cell lines and patients, while each row of $\boldsymbol{S}_J$ provides the location of the corresponding sample in this shared feature space. The score matrices $\boldsymbol{S}_{I_c} \in \mathbb{R}^{n_c \times r_{I_c}}$ and $\boldsymbol{S}_{I_p} \in \mathbb{R}^{n_p \times r_{I_p}}$ summarize the features individual to cell lines and patients, with the corresponding loadings $\boldsymbol{V}_{I_c} \in \mathbb{R}^{g \times r_{I_c}}$ and $\boldsymbol{V}_{I_p} \in \mathbb{R}^{g \times r_{I_p}}$, respectively.

Relation to JIVE. The model formalization above is similar in some respect to the JIVE model proposed by Lock et al. (2013), since both decompose the original unaligned data into joint and individual structures as well as an error matrix. However, the difference is that JIVE and its related methods (Gaynanova and Li, 2019; Feng et al., 2018; Kaplan and Lock, 2017) are designed to integrate multi-source data (i.e., the same set of samples with data from different sources), while iPCR aims to integrate multi-system data (i.e., different biological model systems with the same set of features). Consequently, in the JIVE model, the two input matrices share the same sample count, but in our model, $\mathbf{C}$ and $\mathbf{P}$ share the same gene count. This difference implies that the joint loading matrix $\boldsymbol{V}_J \in \mathbb{R}^{g \times r_J}$ in our Model 2.2 corresponds to a joint score matrix $\boldsymbol{S}_J \in \mathbb{R}^{n \times r_J}$ in JIVE, leading to different interpretations. Here, n represents the common sample size in JIVE's multi-source context.

2.2.2.2 Stage II: Drug response prediction model

At this stage, we use the estimated $\boldsymbol{S}_J = [(\boldsymbol{S}_{J_c})^\top, (\boldsymbol{S}_{J_p})^\top]^\top$ from Stage I, which captures the shared biology between cell lines and patients, along with the observed $\mathbf{D}_c$, to predict the unobserved $\mathbf{D}_p^*$ using a penalized linear regression model. We choose a penalized model ranther than a non-penalized one because our pan-cancer application in Section 2.4 is in a high-dimensional setting, and the value of r_J may not be very small regarding the sample size. For example, in our real application, we choose $r_J = 55$ for most drugs through rank selection. To avoid overfitting, in this paper, we use the elastic net regularization for the estimation of Model 2.3. We note here that the penalized Gaussian linear regression model suffices for our case study and is easy for interpretation, although any other linear or nonlinear prediction models could be potentially used in this stage.

Specifically, the drug responses are regressed on the r_J shared features in $\boldsymbol{S}_J$ from Stage I, rather than the original g gene features. The drug response model for cell lines can be

formularized as follows:

$$\mathbf{D}_c = \mathbf{1}_{n_c}\beta_0 + \boldsymbol{S}_{J_c}\boldsymbol{\beta}_J + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\sigma}^2\mathbf{I}), \tag{2.3}$$

where $\mathbf{1}_{n_c}$ is a $n_c \times 1$ vector of ones, β_0 is a constant intercept, $\boldsymbol{\beta}_J$ is a $r_J \times 1$ vector of regression coefficients that measure the effect of the corresponding shared features on drug responses, and $\boldsymbol{\epsilon}$ is a $n_c \times 1$ error vector. Since $\boldsymbol{S}_{J_c}$ and $\boldsymbol{S}_{J_p}$ capture the shared biology between the two biological model systems, we assume that the same latent shared feature in both $\boldsymbol{S}_{J_c}$ and $\boldsymbol{S}_{J_p}$ has an identical effect on their drug responses. Based on this assumption, we can then use $\boldsymbol{S}_{J_p}$, the estimated β_0 and the estimated $\boldsymbol{\beta}_J$ to predict the unobserved patient drug responses $\mathbf{D}_p^*$, as detailed in Section 2.2.3.1.

To enhance the interpretability of regression coefficients, it is essential to transition from the latent feature space to the gene space, which offers clearer biological significance. Specifically, $\boldsymbol{S}_{J_c}$ can be approximated by projecting the original cell lines' gene expression data onto the latent shared feature space such that $\hat{\boldsymbol{S}}_{J_c} = \mathbf{C}\hat{\boldsymbol{V}}_J$. By substituting this into Equation 2.3, we can get

$$\hat{\mathbf{D}}_c = \mathbf{1}_{n_c}\hat{\beta}_0 + \mathbf{C}\hat{\boldsymbol{V}}_J\hat{\boldsymbol{\beta}}_J, \tag{2.4}$$

where $\hat{\boldsymbol{V}}_J\hat{\boldsymbol{\beta}}_J = \hat{\boldsymbol{\beta}}_{genes} = (\hat{\beta}_{gene_1}, \hat{\beta}_{gene_2}, \ldots, \hat{\beta}_{gene_g})^\top$ represents $g \times 1$ vector of estimated regression coefficients for shared gene effects between cell lines and patients on drug responses. In addition, the ranked list of these gene effects can be used as input in the PreRanked gene set enrichment analysis (Subramanian et al., 2005).

Transferability of Drug Responses from Cell Lines to Patients. It is crucial to note that our model does not assume complete transferability of drug responses from cell lines to patients. Our hypothesis is anchored in the principle of joint variation observed in Stage I of the iPCR process. We demonstrate that a higher degree of joint variation signifies an increased likelihood of successful transferability, as detailed in the later simulation section. This rationale is central to our approach for isolating shared variation from the gene expression data of both cell lines and patients, with the objective of assessing and leveraging this joint variation to improve the precision of our predictions.

2.2.3 Model estimation

In this section, we outline the estimation process for our two-stage iPCR model. First, we discuss the estimation of latent joint and individual structures at Stage I, given their predetermined ranks. Next, we explain how the model training and drug response prediction

are conducted at Stage II, using shared features between cell lines and patients derived from the joint structure. Then, we describe the method for rank selection, optimized through cross-validation. Lastly, when the sample sizes of cell lines and patients differ significantly, we introduce a scaling technique at Stage I to balance the data integration process, followed by a corresponding re-scaling at Stage II.

2.2.3.1 Stages I and II given joint and individual ranks

For the estimation of patient-cell line integration model at Stage I (Model 2.1), suppose the joint rank r_J and individual ranks $\{r_{I_c}, r_{I_p}\}$ are given. As mentioned in Section 2.2.2.1, although our patient-cell line integration model and JIVE have different problem settings and interpretations, their model formulations are similar. Hence, we use the algorithm introduced by Lock et al. (2013) and "r.jive" package (O'Connell and Lock, 2016) to estimate the joint and individual structures with the given ranks $\{r_J, r_{I_c}, r_{I_p}\}$. The final estimated joint and individual structures are denoted as $\hat{\boldsymbol{J}}$, $\hat{\boldsymbol{I}}_c$, and $\hat{\boldsymbol{I}}_p$. Then, SVD is performed on $\hat{\boldsymbol{J}}$ such that $\hat{\boldsymbol{J}} = \hat{\boldsymbol{U}}_J \hat{\boldsymbol{\Sigma}}_J \hat{\boldsymbol{V}}_J^\top$. Those r_J estimated shared features between cell lines and patients can be acquired from the estimated joint score matrix $\hat{\boldsymbol{S}}_J = [\hat{\boldsymbol{S}}_{J_c}^\top, \hat{\boldsymbol{S}}_{J_p}^\top]^\top = \hat{\boldsymbol{U}}_J \hat{\boldsymbol{\Sigma}}_J$, which allows us to use Model 2.3 to predict patient drug response given the responses from cell lines to a certain drug D.

Given $\hat{\boldsymbol{S}}_J$ from Stage I, we use $\mathbf{D}_c$ and $\hat{\boldsymbol{S}}_{J_c}$ as the training data to fit Model 2.3 with an elastic net regularization (Zou and Hastie, 2005). The regression coefficients $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}_J$ are shrunk based on both L1-norm and the L2-norm. The elastic net model training is done in R based on the caret package (Kuhn, 2008). Finally, the predicted responses of n_p patients to the given drug D are given by

$$\hat{\mathbf{D}}_p^* = \mathbf{1}_{n_p} \hat{\beta}_0 + \hat{\boldsymbol{S}}_{J_p} \hat{\boldsymbol{\beta}}_J. \tag{2.5}$$

Note that the intercept term is assumed to be identical in Model 2.3 (for cell lines) and Model 2.5 (for patients) for two reasons. Firstly, since patient drug responses are completely unobserved, we assume the estimated intercept for cell lines to be the optimized intercept for patients. Secondly, the interpretation of drug response metrics, such as drug-AUC or IC50, is nuanced; a single value cannot conclusively determine whether it represents a response or a non-response. Instead, comparative analysis across a cohort of samples is required to distinguish responsive from non-responsive subjects. Given that the intercept uniformly affects all predictions, it does not alter the relative rankings of patients, thus preserving the integrity of our comparative analysis.

2.2.3.2 Rank selection of joint and individual structures

The unsupervised rank selection methods (Lock et al., 2013) tend to neglect those latent signals that explain a very small proportion of variance in the data, although they could be strong predictors of outcomes. This could result in underestimation of the true rank and subsequently impact prediction accuracy. Hence, analysts sometimes select the optimal number of principal components based on the prediction performance through cross-validation in the labelled source data (Dadousis et al., 2014; Lee et al., 2015). In addition, a domain adaptation study (Pan et al., 2011) showed the advantage of using label information in the source domain data to improve prediction performance. Inspired by these studies, we select an optimal rank set $\{r_J, r_{I_c}, r_{I_p}\}^{opt}$ by performing a 10-fold cross-validation based on $\mathbf{D}_c$ and $\hat{\boldsymbol{S}}_{J_c}$ for a certain drug.

Algorithm 1 iPCR model

Require: cell lines' genomic data $\mathbf{C}$ and drug data $\mathbf{D_c}$, patients' genomic data $\mathbf{P}$, and a set of rank set candidates $\mathbf{R}_{cand} = \{\mathbf{r}_1, \ldots, \mathbf{r}_m\}$

1: **for** k:=1 **to** m **do**
2: Calculate $(\hat{\boldsymbol{J}})_k$, $(\hat{\boldsymbol{I}}_c)_k$, and $(\hat{\boldsymbol{I}}_p)_k$ in Model 2.1, given $\mathbf{r}_k$.
3: Calculate $(\hat{\boldsymbol{S}}_J)_k = (\hat{\boldsymbol{U}}_J)_k(\hat{\boldsymbol{\Sigma}}_J)_k$, where $(\hat{\boldsymbol{J}})_k = (\hat{\boldsymbol{U}}_J)_k(\hat{\boldsymbol{\Sigma}}_J)_k(\hat{\boldsymbol{V}}_J^\top)_k$.
4: Randomly split labeled cell line cohort $\{\mathbf{D}_c, (\hat{\boldsymbol{S}}_{J_c})_k\}$ into 10 folds, and perform 10-fold cross validation based on Model 2.3.
5: Calcualte the mean squared prediction error (MSPE) across all testing folds, through the cross validation.
6: **end for**
7: Determine the optimal rank set $\{r_J, r_{I_c}, r_{I_p}\}^{opt}$, corresponding to the smallest MSPE from the 10-fold cross validation.
8: Calculate $\hat{\boldsymbol{J}}$, $\hat{\boldsymbol{I}}_c$, and $\hat{\boldsymbol{I}}_p$ in Model 2.1 based on the optimal rank set.
9: Calculate $\hat{\boldsymbol{S}}_J = \hat{\boldsymbol{U}}_J\hat{\boldsymbol{\Sigma}}_J$, where $\hat{\boldsymbol{J}} = \hat{\boldsymbol{U}}_J\hat{\boldsymbol{\Sigma}}_J\hat{\boldsymbol{V}}_J^\top$.
10: Regress $\mathbf{D_c}$ on $\hat{\boldsymbol{S}}_{J_c}$, and get $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}_J$ in Model 2.3.
11: The predicted patient drug response $\hat{\mathbf{D}}_p^* = \mathbf{1}_{n_p}\hat{\beta}_0 + \hat{\boldsymbol{S}}_{J_p}\hat{\boldsymbol{\beta}}_J$.

To provide more details for Algorithm 1, we first choose m candidates of rank set $\{\mathbf{r}_1, \ldots, \mathbf{r}_m\}$ for grid search, such that $\mathbf{r}_k = \{r_J, r_{I_c}, r_{I_p}\}_k$, where $k = 1, \ldots, m$. Since the number of combinations of r_J, r_{I_c} and r_{I_p} grows cubically with the number of candidates for each type of rank, the exhaustive grid research is only feasible for low-dimensional training data. Due to the high dimension of pan-cancer genomic data, we limit the number of combinations by only doing a grid search on r_J. Given a certain r_J, the corresponding r_{I_c} and r_{I_p} are calculated by subtracting r_J from the estimated ranks of denoised $\mathbf{C}$ and $\mathbf{P}$, respectively; the rank of denoised $\mathbf{C}$ or $\mathbf{P}$ is estimated based on the inflection point of the cumulative eigenvalues through PCA, or, if there is no obvious inflection point, the parallel

analysis (Franklin et al., 1995) through "paran" package (Dinno, 2009).

As Algorithm 1 shows, for each candidate rank set $\mathbf{r}_i$, we then get $(\hat{\boldsymbol{S}}_{J_c})_k$ from the patient-cell line integration model, and randomly split labeled cell line cohorts $\{\mathbf{D}_c, (\hat{\boldsymbol{S}}_{J_c})_k\}$ into 10 folds. Next, we perform 10-fold cross validation based on Model 2.3, and calculate the mean squared prediction error (MSPE) across all testing folds corresponding to $\mathbf{r}_k$. Last, the optimal rank set $\{r_J, r_{I_c}, r_{I_p}\}^{opt}$ is chosen with the smallest MSPE from the cross validation. The optimal rank set is used for the training of the iPCR model to predict $\mathbf{D}_p^*$.

2.2.3.3 Data scaling at Stages I and II

In scenarios where the sample sizes in $\mathbf{C}$ and $\mathbf{P}$ exhibit substantial discrepancies, "the bigger dataset always wins" during data integration at Stage I. This bias can adversely affect the accuracy of predictions at Stage II. To avoid this issue, it is helpful to scale $\mathbf{C}$ and $\mathbf{P}$ by their total variation such that $\mathbf{C}^{scaled} = \frac{\mathbf{C}}{\|\mathbf{C}\|_F}$, and $\mathbf{P}^{scaled} = \frac{\mathbf{P}}{\|\mathbf{P}\|_F}$, so that $\|\mathbf{C}^{scaled}\|_F = \|\mathbf{P}^{scaled}\|_F = 1$, where $\|\cdot\|_F$ defines the Frobenius norm (Lock et al., 2013; Gaynanova and Li, 2019). At Stage I, using the scaled matrices $\mathbf{C}^{scaled}$ and $\mathbf{P}^{scaled}$ as inputs, we obtain the estimated scaled joint score matrices $\hat{\boldsymbol{S}}_{J_c}^{scaled}$ and $\hat{\boldsymbol{S}}_{J_p}^{scaled}$ following the methodology in Section 2.2.3.1. At Stage II, these matrices are scaled back such that $\hat{\boldsymbol{S}}_{J_c} = \|\mathbf{C}\|_F \hat{\boldsymbol{S}}_{J_c}^{scaled}$ and $\hat{\boldsymbol{S}}_{J_p} = \|\mathbf{P}\|_F \hat{\boldsymbol{S}}_{J_p}^{scaled}$, which are subsequently utilized for model training and patient drug response prediction.

2.2.4 Gene co-expression network estimation

Genes are naturally interconnected through biological networks (Goh et al., 2007). Various network-based models have been developed to identify gene co-expression networks, aiming to yield insights into underlying biological and regulatory pathways (Goh et al., 2007; Barabási et al., 2010; Marbach et al., 2016; Santolini and Barabási, 2018; Wu et al., 2021). However, it is not feasible to assume a universal network structure across different types of biological samples (Acharyya et al., 2022). Acharyya et al. Acharyya et al. (2022) proposed a model, known as SpaceX, to discern both shared and cluster-specific gene co-expression networks by decomposing the covariance matrix of genes into its shared and cluster-specific components. Inspired by this approach, in this section we provide a new angle to study the gene co-expression network in cell lines and patients, based on Model 2.1, which is also a byproduct of iPCR. Unlike SpaceX, which relies on a factor model and decomposes the true underlying covariance matrix, iPCR is related to PCA and focuses on decomposing the sample covariance matrix.

The estimated denoised gene expression matrices for cell lines and patients $\tilde{\boldsymbol{G}}_i = \hat{\boldsymbol{J}}_i + \hat{\boldsymbol{I}}_i$

$(i = c, p)$ from step 8 in Algorithm 1 are used as input herein this section. Let $|\cdot|$ be the cardinal number of a set. We define the gene co-expression network as a graph $\{V, E\}$, where the vertex set V denotes g genes with $|V| = g$ and E represents the set of undirected pairwise relationships (edges) between the g genes such that $|E| = g(g-1)/2$. The edge between two gene vertices denotes their co-expression level, which is defined using a similarity measure, Pearson correlation coefficient. Our gene co-expression network estimation method decomposes the gene network into two hierarchical components: 1) a shared component representing the shared gene co-expression network across the two biological model systems (cell line and patient); and 2) a model system specific component representing the gene co-expression network individual to the cell lines or patients.

For all $i \in \{c, p\}$, we propose a sample covariance decomposition method for $\tilde{\boldsymbol{G}}_i$, such that

$$\tilde{\boldsymbol{G}}_i \quad = \quad \tilde{\boldsymbol{J}}_i + \tilde{\boldsymbol{I}}_i, \text{with } \hat{cov}(\tilde{\boldsymbol{G}}_i) = \hat{cov}(\tilde{\boldsymbol{J}}_i) + \hat{cov}(\tilde{\boldsymbol{I}}_i), \tag{2.6}$$

where $\hat{cov}(\cdot)$ represents sample covariance matrix between features (i.e., genes). The matrices $\tilde{\boldsymbol{J}}_i$ and $\tilde{\boldsymbol{I}}_i$ can be uniquely determined under the conditions specified in Proposition 2. The unique solution and the proof of identifiability are provided in Appendix Section A.2.1 and Section A.2.2, respectively. Note that Propositions 1 and 2 have the same Condition 3, that there is a shared feature space between cell lines and patients. The only difference between them is that in Proposition 1 the row spaces between $\boldsymbol{J}_i$ and $\boldsymbol{I}_i$ are orthogonal, while Proposition 2 has the constraint that column spaces between $\tilde{\boldsymbol{J}}_i$ and $\tilde{\boldsymbol{I}}_i$ are orthogonal, for all $i \in \{c, p\}$. Condition 2 of Proposition 2, which focuses the orthogonality of column spaces, serves as a key constraint that ensures $\tilde{\boldsymbol{J}}_i$ and $\tilde{\boldsymbol{I}}_i$ satisfy the property of sample covariance decomposition in Model 2.6. Further, Condition 4 in Proposition 2 is essential for the uniqueness of the decomposition. When paired with Condition 2, it also guarantees the property such that $rank(\tilde{\boldsymbol{J}}_i + \tilde{\boldsymbol{I}}_i) = rank(\tilde{\boldsymbol{J}}_i) + rank(\tilde{\boldsymbol{I}}_i)$.

Proposition 2. *Given any set of matrices $\{\tilde{\boldsymbol{G}}_i, i = c, p\}$, there are unique sets of matrices $\{\tilde{\boldsymbol{J}}_i, i = c, p\}$ and $\{\tilde{\boldsymbol{I}}_i, i = c, p\}$ such that,*

(1) $\tilde{\boldsymbol{G}}_i = \tilde{\boldsymbol{J}}_i + \tilde{\boldsymbol{I}}_i$ *for* $i = c,\ p$

(2) $col(\tilde{\boldsymbol{J}}_i) \perp col(\tilde{\boldsymbol{I}}_i) = 0$ *for* $i = c,\ p$

(3) $row(\tilde{\boldsymbol{J}}_i) = row(\tilde{\boldsymbol{J}})$ *for* $i = c,\ p$, *where* $\tilde{\boldsymbol{J}} = \begin{bmatrix} \tilde{\boldsymbol{J}}_c \\ \tilde{\boldsymbol{J}}_p \end{bmatrix}$

(4) $row(\tilde{\boldsymbol{J}}_i) \cap row(\tilde{\boldsymbol{I}}_i) = 0$ *for* $i = c,\ p$

(5) $row(\tilde{\boldsymbol{I}}_c) \cap row(\tilde{\boldsymbol{I}}_p) = 0$

For easier interpretation, for all $i \in \{c, p\}$, the sample covariance matrices $\hat{cov}(\tilde{\boldsymbol{J}}_i)$ and $\hat{cov}(\tilde{\boldsymbol{I}}_i)$ are transformed to the sample correlation matrices $\hat{cor}(\tilde{\boldsymbol{J}}_i)$ and $\hat{cor}(\tilde{\boldsymbol{I}}_i)$, respectively. For example, the $(m,n)^{th}$ element of $\hat{cor}(\tilde{\boldsymbol{J}}_i)$, denoted as $\hat{cor}(\tilde{\boldsymbol{J}}_i)_{m,n}$, represents the shared sample correlation between m^{th} gene and n^{th} gene in the cell lines' gene expression data ($i = c$) or patients' gene expression data ($i = p$). An edge between two genes is considered to be significant in the co-expression network if the corresponding correlation's absolute value is larger than the threshold δ (e.g., $\delta = 0.7$). To be conservative, we define the edge between m^{th} gene and n^{th} gene as significant in the shared gene co-expression network between cell lines and patients, if both $\hat{cor}(\tilde{\boldsymbol{J}}_c)_{m,n}$ and $\hat{cor}(\tilde{\boldsymbol{J}}_p)_{m,n}$ are larger than δ. We then use the shared and model-specific gene co-expression networks to detect hub genes based on the number of edges linked to each gene.

2.3 Simulation study

We compare the drug response prediction accuracy of our iPCR model with other alternative methods in the simulated cell line and patient data mimicking our real data applications (Section 2.4). The simulated data are generated under different levels of similarity between cell lines and patients and signal-to-noise ratios. We also evaluate the sensitivity of rank selection relative to the prediction performance for iPCR.

2.3.1 Simulation design and comparative analyses

This simulation study examines the performance of iPCR and other methods under different levels of similarity between cell lines and patients and degrees of noise. We assume the dimensions of $\mathbf{C}$ and $\mathbf{P}$ are 1500×2000 and 10000×2000, respectively, where rows correspond to samples and columns correspond to genes. $\mathbf{C}$ and $\mathbf{P}$ are sampled based on the following models:

$$\mathbf{C} = \underbrace{\boldsymbol{U}_{J_c}\boldsymbol{\Sigma}_J\boldsymbol{V}_J^\top}_{\boldsymbol{J}_c} + \underbrace{\boldsymbol{U}_{I_c}\boldsymbol{\Sigma}_{I_c}\boldsymbol{V}_{I_c}^\top}_{\boldsymbol{I}_c} + \boldsymbol{\varepsilon}_c^{stage1}, \tag{2.7}$$

and

$$\mathbf{P} = \underbrace{\boldsymbol{U}_{J_p}\boldsymbol{\Sigma}_J\boldsymbol{V}_J^\top}_{\boldsymbol{J}_p} + \underbrace{\boldsymbol{U}_{I_p}\boldsymbol{\Sigma}_{I_p}\boldsymbol{V}_{I_p}^\top}_{\boldsymbol{I}_p} + \boldsymbol{\varepsilon}_p^{stage1}, \tag{2.8}$$

where $r_J^{True} = rank(\boldsymbol{J}_c) = rank(\boldsymbol{J}_p) = rank([\boldsymbol{J}_c^\top, \boldsymbol{J}_p^\top]^\top) = 40$, $r_{I_c}^{True} = rank(\boldsymbol{I}_c) = 30$, $r_{I_p}^{True} = rank(\boldsymbol{I}_p) = 50$, the elements of $\boldsymbol{\varepsilon}_c^{stage1}$ follow $N(0, \frac{\boldsymbol{\delta}^2_{J_c+I_c}}{SNR_{stage1}})$ and the elements of $\boldsymbol{\varepsilon}_c^{stage1}$ follow $N(0, \frac{\boldsymbol{\delta}^2_{J_p+I_p}}{SNR_{stage1}})$. For all $i \in \{c, p\}$, the signal-to-noise ratio SNR_{stage1} is defined as the ratio between the sample variance of elements in the denoised data $\boldsymbol{J}_i + \boldsymbol{I}_i$ (denoted as $\boldsymbol{\delta}^2_{J_i+I_i}$) and the variance of $\boldsymbol{\varepsilon}_i^{stage1}$. The shared feature space $row(\boldsymbol{V}_J)$ is orthogonal with the individual feature spaces $row(\boldsymbol{V}_{I_c})$ and $row(\boldsymbol{V}_{I_p})$, and $row(\boldsymbol{V}_{I_c})$ is also orthogonal with $row(\boldsymbol{V}_{I_p})$. We use the term "joint proportion" to refer to the ratio of $\|\boldsymbol{J}_i\|_F^2$ to $\|\boldsymbol{J}_i + \boldsymbol{I}_i\|_F^2$, which quantifies the proportion of variation of $\boldsymbol{J}_i$ relative to the denoised genomic data $\boldsymbol{J}_i + \boldsymbol{I}_i$ for all $i \in c, p$. In other words, a larger value of joint proportion corresponds to a higher level of similarity between cell lines and patients. In order to meet orthogonality constraints and control different levels of joint proportion from 20% to 80% in 10% increments, the SVD components of $\boldsymbol{J}_c$, $\boldsymbol{J}_p$, $\boldsymbol{I}_c$, and $\boldsymbol{I}_p$ in Model 2.7 and 2.8 are constructed separately (more details in Appendix Section A.3.1). We also set different levels of SNR_{stage1} at 0.5, 1, or 2.

We assume that $\mathbf{D}_c$ and $\mathbf{D}_p$ depend on the joint score matrices $\boldsymbol{S}_{J_c} = \boldsymbol{U}_{J_c}\boldsymbol{\Sigma}_J$ and $\boldsymbol{S}_{J_p} = \boldsymbol{U}_{J_p}\boldsymbol{\Sigma}_J$, respectively, such that

$$\mathbf{D}_c = \boldsymbol{S}_{J_c}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_c^{stage2}, \quad \mathbf{D}_p = \boldsymbol{S}_{J_p}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_p^{stage2}. \tag{2.9}$$

The regression coefficient $\boldsymbol{\beta}$ is a vector of length 40, where the elements are generated from $N(0, 4)$. The elements of $\boldsymbol{\varepsilon}_c^{stage2}$ and $\boldsymbol{\varepsilon}_p^{stage2}$ follow $N(0, \frac{\boldsymbol{\delta}^2_{S_{J_c}\beta}}{SNR_{stage2}})$ and $N(0, \frac{\boldsymbol{\delta}^2_{S_{J_p}\beta}}{SNR_{stage2}})$, where $\boldsymbol{\delta}^2_{S_{J_c}\beta}$ and $\boldsymbol{\delta}^2_{S_{J_p}\beta}$ represent the sample variances of $\boldsymbol{S}_{J_c}\boldsymbol{\beta}$ and $\boldsymbol{S}_{J_p}\boldsymbol{\beta}$, respectively. The signal-to-noise ratio in Model 2.9, SNR_{stage2}, is set to 5. The prediction performance is measured by the Spearman correlation between the true and predicted values of $\mathbf{D_p}$, focusing solely on the order of values. In each setting, we generate 50 replicates of $\{\mathbf{C}, \mathbf{P}, \mathbf{D_c}\}$ for training and $\mathbf{D}_p$ for testing.

For the iPCR model, six candidate rank sets are chosen for grid search. Specifically, the candidate rank sets $\{r_J, r_{I_c}, r_{I_p}\}_k$, k=1...6, include $\{10, 60, 80\}_1$, $\{20, 50, 70\}_2$, $\{30, 40, 60\}_3$, $\{40, 30, 50\}_4$, $\{50, 20, 40\}_5$, and $\{60, 10, 30\}_6$ based on the rank selection guideline in Section 2.2.3.3. We compare with 1) an elastic net regression trained solely on cell line data (Elastic Net); 2) an elastic net regression trained and tested on the batch-corrected data (Combat + Elastic Net), similar to Geeleher et al. (2017); 3) an elastic net regression trained and tested on the first K principal scores of $[\mathbf{C}^\top, \mathbf{P}^\top]^\top$ (K selected through 10-fold cross validation), which in this paper is called principal component regression (PCR); and 4) TRANSACT(Mourragui et al., 2021), a leading domain adaptation approach for drug response prediction. Both PCR and iPCR aim to detect the joint structure based on variation decomposition, but PCR does not consider individual structures during the modeling. In

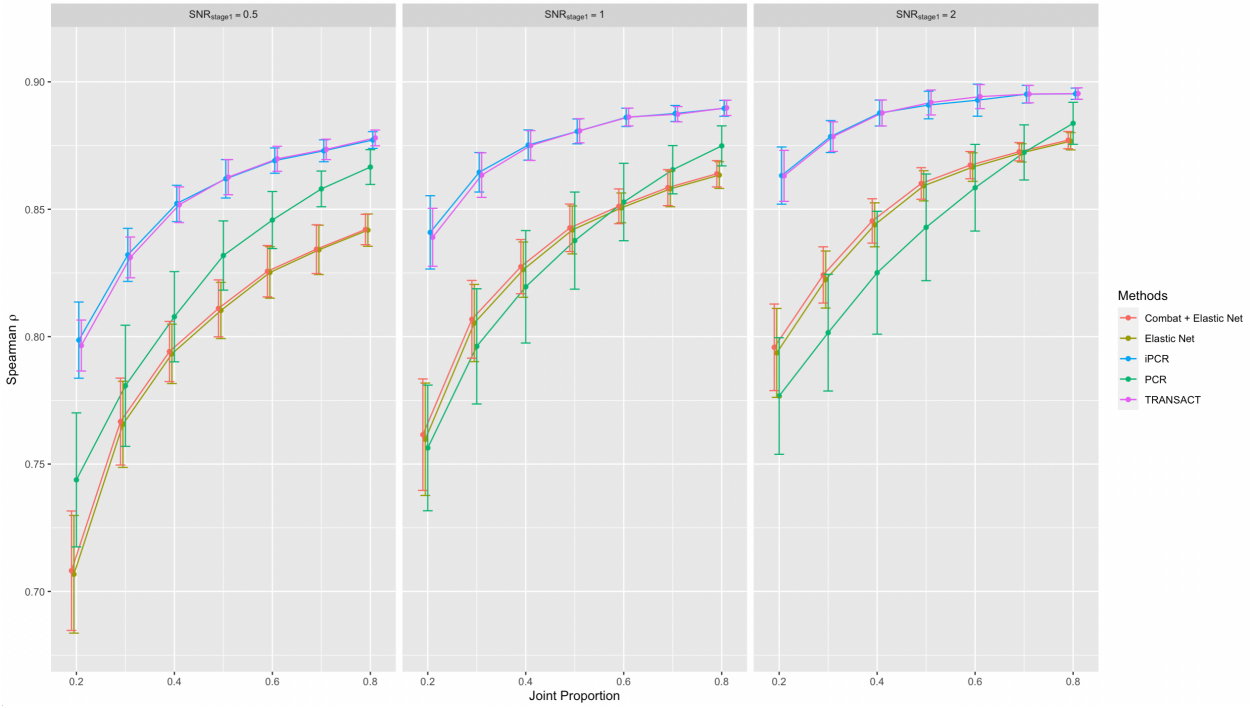


Figure 2.2: **Simulation results for drug response prediction.** Accuracy comparison of different methods in predicting $\mathbf{D_p}$ in simulation study. The prediction performance is measured by Spearman correlation between the true and predicted values of $\mathbf{D_p}$. The data are generated under different levels of SNR_{stage1} and joint proportion over 50 replications. The sample mean of Spearman correlation over replications correspond to the solid curves, and the error bars represent mean $\pm$ 1sd.

addition, TRANSACT focuses on aligning the kernel principal components of cell line and patient gene expression data. To make an apples-to-apples comparison, we consistently use Elastic Net as the regularization method across the above methods.

Figure 2.2 demonstrates the prediction accuracy (measured by the Spearman correlation between the predicted and true $\mathbf{D_p}$) of different methods over 50 replications for each combination of SNR_{stage1} and joint proportion. For better illustration, we plot the curves of Spearman correlation as a function of joint proportion in three subplots that correspond to three levels of SNR_{stage1}. As expected, when joint proportion increases, the prediction accuracy of all methods increases, since the expected drug responses are assumed to only depend on the similarities between cell lines and patients. In addition, a larger SNR_{stage1} leads to a higher prediction accuracy for all methods. iPCR outperforms the competing methods under different levels of SNR_{stage1} and joint proportion, with the exception of TRANSACT, which shows slightly better predictive performance than iPCR. Surprisingly, Combat + Elastic Net performs very similarly with Elastic Net, which means that the Combat batch

removal method does not filter model-specific differences in our setting. It is also fair to say that the batch effect is negligible in this instance. PCR performs worse than iPCR, in part because in PCR the first K principal scores used for prediction includes both shared and model-specific information. As joint proportion increases, those first K principal scores in PCR include more information from the joint structure, so it performs closer to iPCR in terms of the prediction accuracy. When $SNR_{stage1} = 0.5$, PCR always outperforms Elastic Net and Combat + Elastic Net at different levels of joint proportion. For $SNR_{stage1} = 1$ (or 2), PCR only outperforms Elastic Net and Combat + Elastic Net when joint proportion is larger than 0.6 (or 0.7); possibly because Elastic Net utilizes all the data including the information of similarity in prediction while PCR has difficulty detecting the information of similarity when joint proportion is low.

2.3.2 Sensitivity analysis of rank selection for iPCR

To measure the sensitivity of rank selection to the prediction accuracy for iPCR, we implement iPCR given a certain rank set without doing the rank selection, following the data generating process in Section 2.3.1. The rank sets of interest in the form of $\{r_J, r_{I_c}, r_{I_p}\}$ include 1) $\{40, 30, 50\}$ that corresponds to the true rank set $\{r_J^{True}, r_{I_c}^{True}, r_{I_p}^{True}\}$; 2) $\{30, 40, 60\}$ and $\{50, 20, 40\}$ that under- and over-estimate r_J^{True}, respectively, while the denoised ranks of $\mathbf{C}$ and $\mathbf{P}$ are correctly chosen; and 3) $\{40, 20, 40\}$ and $\{40, 40, 60\}$ that under and over estimate the true individual ranks, respectively, while the joint rank is correctly chosen. The accuracy comparison of iPCR given different rank sets in predicting $\mathbf{D_p}$ is shown in Figure A.1. We note that the accuracy curve of iPCR in Figure 2.2 is very similar to the accuracy curve of iPCR given the true rank set in Figure A.1. This observation illustrates that in Section 2.3.1 iPCR always correctly selects the true rank set for prediction; although it sometimes does not select the true rank set, it achieves comparable accuracy to iPCR given the true rank set. In addition, given the four misspecified rank sets defined in this section, iPCR still outperforms the competing methods in Section 2.3.1, except iPCR given $\{30, 40, 60\}$ (the case that under estimates r_J^{True}) when joint proportion is high, as shown by Figure 2.2 and A.1. Further discussion of the sensitivity analysis of iPCR rank selection can be found in Appedix Section A.3.2.

2.4 Application to pan-cancer drug response data

We performed a retrospective in-silico analysis to evaluate our iPCR model on a pan-cancer dataset. Our primary scientific goals are to: 1) identify top drugs based on prediction

performance on patient tumors using cell-line data; 2) infer and interpret iPCR gene signatures associated with drug responses to identify novel drug targets; and 3) investigate the joint and individual gene co-expression networks between cell lines and patients, to identify co-ordinated mechanisms underlying drug resistance.

2.4.1 Dataset description

All datasets used in our application are publicly available. Gene expression data for 12,747 patient tumors from >40 cancer types were downloaded from the Tumor Compendium v11 Public PolyA, a database maintained by The Treehouse Childhood Cancer Initiative at the UCSC Genomics Institute (Goldman et al., 2020). These patient tumors include samples from Treehouse clinical sites, the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) project, and The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013). Gene expression data for 1,406 cell lines from 30 different tissues were obtained from the DepMap Public 22Q2 file, named as "CCLE_expression_full.csv" (Barretina et al., 2012). The gene expression data for both patient tumors and cell lines were pre-processed using established pipelines, as described in Barretina et al. (2012); Goldman et al. (2020). Additionally, the data scaling pipeline outlined in Section 2.2.3.3 was applied due to the large difference in sample size between patients and cell lines. We restricted our analysis to 1,774 cancer-related genes (Hoogstraat et al., 2013) that were present in both tumor and cell line data, and performed a gene-level centralization and standardization.

The *in vivo* drug response data of cell lines were obtained from the Genomics of Drug Sensitivity in Cancer (GDSC) panel (Yang et al., 2013). The *in vitro* response data of patients to 13 different cancer drugs for validation instead of model fitting, were obtained from Ding et al. (2016). The drug response data of cell lines and patients contained a subset of samples in the gene expression data. The number of patients with available drug response data for each drug is listed in Table 2.1, while the cell line sample size exceeds 600 for each drug. For cell lines in the GDSC panel, drug response was measured by the area under the drug response curve. This metric integrates information about both drug efficacy and potency into a single value, and has been previously reported as the most robust metric (Kurilov et al., 2020). Note, to distinguish the name of the drug response metric with the area under the ROC curve (AUC) used for measuring prediction performance, we named the former as the drug-AUC. Large values of drug-AUC are associated with drug resistance. We classified patient drug responses into two categories: Response (including "Complete Response" and "Partial Response") and Non-response (including "Stable Disease" and "Progressive Disease").

2.4.2 Patient drug response prediction

To validate the prediction performance of our iPCR model, we first aligned the whole cell line and patient expression data based on stage I of iPCR (Section 2.2.2.1). Next, for each drug, we used the cross-validation method described in Section 2.2.3.3 to select the best combination of joint and individual ranks among the candidates. Following the rank selection guideline in Section 2.2.3.3, we chose $20, 25, \ldots, 60$ as the candidates for joint ranks. For each candidate joint rank, the corresponding individual ranks of cell lines and patients were obtained by subtracting the joint rank from 64 and 84 (the estimated ranks of denoised cell line and patient gene expression data based on parallel analysis (Dinno, 2009)), respectively. Then, for each drug, given the optimal ranks, we trained the regression model in stage II of iPCR (Section 2.2.2.2) to predict the drug-AUC values for patients. Finally, we calculated the AUC by comparing the predicted drug-AUC values between known responders and non-responders. The higher the AUC, the better the prediction performance.

The competing approaches for patient drug response prediction included the four alternative methods mentioned in Section 2.3.1, along with iPCR+XGBoost (Chen and Guestrin, 2016), and SIR+XGBoost (Chen and Guestrin, 2016; Li, 1991). For TRANSACT, we used $\gamma = 5 \times 10^{-4}$ (amount of non-linearity) as a tuning parameter, as recommended by Mourragui et al. (2021). This parameter was calibrated through transfer learning from cell lines to PDX models. The iPCR+XGBoost method begins with stage I of iPCR to reduce the number of predictors, subsequently employing an XGBoost model in stage II to predict drug responses. This aims to capture the nonlinear relationship between predictors and responses. Meanwhile, the SIR+XGBoost method applies sliced inverse regression (SIR) to cell line data, obtaining a concise set of predictors that reflect both nonlinear and linear relationships with the drug response. These predictors, derived from SIR, are then utilized to train an XGBoost model. Note that during model training, except for Elastic Net regression and SIR+XGBoost, all other methods utilized the patient gene expression data, while none of the methods used patient drug responses.

In Table 2.1, we show the prediction performance for 9 drugs that exist at least one approach with AUC > 0.6. Cells in bold indicate the highest AUC among the seven methods for predicting the patient response to a certain drug. Our iPCR approach outperforms other methods in 4 drugs (Trastuzumab, Etoposide, Vinorelbine, and Oxaliplatin), PCR outperforms in 3 drugs (Trastuzumab, Temozolomide, and Pemetrexed), TRANSACT outperforms in 2 drugs (Gemcitabine and Paclitaxel), iPCR+XGBoost outperforms in 2 drugs (Trastuzumab and Oxaliplatin), SIR+XGBoost outperforms in 1 drug (Gemcitabine), and Combat + Elastic Net outperforms in 1 drug (Cisplatin). No existing method can dominate other methods across different cancer drugs. However, except Cisplatin, the methods (PCR,

Table 2.1: **Results for human drug response prediction.** Performance results, as measured by AUC, for iPCR and six competing methods in patient drug response prediction.

Drug	Patient Samples	Elastic Net	Combat + Elastic Net	PCR	TRANSACT	SIR+XGBoost	iPCR	iPCR+XGBoost
Trastuzumab (Afatinib)	16	0.82	0.82	**1.00**	0.71	0.64	**1.00**	**1.00**
Etoposide	86	0.52	0.53	0.41	0.68	0.58	**0.70**	0.56
Vinorelbine	31	0.60	0.61	0.64	0.64	0.60	**0.67**	0.55
Cisplatin	309	0.68	**0.68**	0.67	0.64	0.54	0.66	0.64
Oxaliplatin	38	0.57	0.58	0.59	0.63	0.54	**0.65**	**0.65**
Gemcitabine	158	0.64	0.64	0.63	0.65	**0.67**	0.61	0.60
Temozolomide	94	0.51	0.50	**0.68**	0.50	0.51	0.58	0.50
Pemetrexed	39	0.57	0.57	**0.65**	0.42	0.58	0.56	0.53
Paclitaxel	157	0.50	0.51	0.54	**0.60**	0.51	0.53	0.54

Note: The drug name in the parenthesis corresponds to the matching drug from GDSC. For each drug and method, we compared the predicted continuous patient drug response to the known binary clinical response. We reported patient sample size for testing and AUC for measuring the prediction performance. Bold cells correspond to the largest AUC across the five methods for predicting the patient response to a certain drug.

TRANSACT, and iPCR) that utilize the integrated gene expression data of cell lines and patients through dimension-reduction techniques improve the patient drug response prediction at some levels, compared with Elastic Net and Combat + Elastic Net. Additionally, utilizing an XGBoost model in the second stage of iPCR instead of Elastic Net does not markedly improve overall performance. While iPCR and iPCR+XGBoost exhibit equivalent performance for Trastuzumab and Oxaliplatin, iPCR demonstrates significantly better performance for Etoposide and Vinorelbine.

Lastly, in comparing iPCR with PCR and considering ± 0.02 as a threshold for similar performance, iPCR outperforms PCR in 3 cases, PCR outperforms iPCR in 2 cases, and they exhibit similar performance in 4 cases. In those three cases where iPCR outperforms PCR, it also achieves the best performance among all competing methods, emphasizing its advantage in incorporating individual structures into the model. In those cases where iPCR does not outperform PCR, one potential reason, as our simulation study suggests, is that iPCR's advantage diminishes compared to PCR when the proportion of shared structure's variation is higher, aligning with the relatively low proportion of individual structures detected by iPCR ($< 20\%$) in real applications, as illustrated in Figure 2.4A.

Association with patient responses. For assessing the association between our predicted drug responses and actual patient outcomes, we standardized the predicted patient drug-AUC values made by iPCR for each drug (shown in Figure 2.3). Each pair of violin plots compares these standardized predicted values between responders (red) and non-responders (dark grey). The five pairs of violin plots correspond to the five drugs where iPCR achieves AUC > 0.65. Based on the one-sided Mann-Whitney U test, the significant difference (p-value < 0.05) between responders and non-responders in terms of the predicted drug-AUC values by iPCR can be found in Trastuzumab (p-value = 0.0083), Etoposide (p-value =

0.015) and Cisplatin (p-value = 4×10^{-5}). In addition, the non-responders always have a higher median of predicted drug-AUC by iPCR than the responders for the drugs shown in Figure 2.3, which is consistent with the fact that a larger drug-AUC is associated with a higher level of drug resistance.

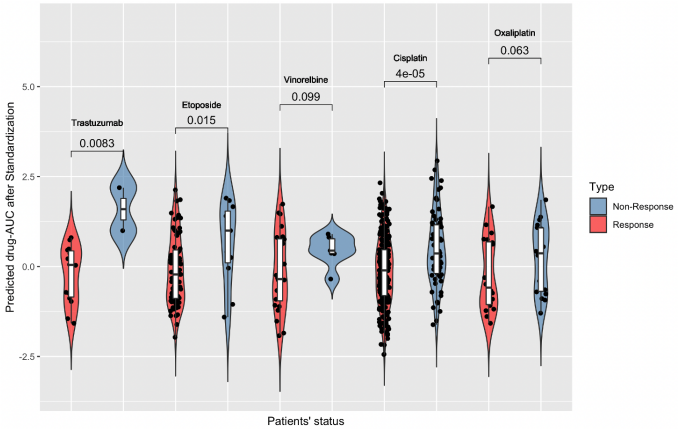


Figure 2.3: **Prediction results of iPCR for five drugs.** For each drug, a pair of violin plots compares the standardized values of predicted patient drug-AUC by iPCR between responders (red) and non-responders (dark gray). The block dots represent the predicted values. The number above each pair of violin plots corresponds to the p-value from the one-sided Mann-Whitney U test.

2.4.3 Shared variation and biological interpretations

In iPCR, given the rank set selected from Algorithm 1, the patient-cell line integration model (Section 2.2.2.1) helps quantify and compare the amount of shared and individual variation in the gene expression data between cell lines and patients. As shown in Table A.2, the patient response prediction for the drugs in Figure 2.3 uses the rank set $\{r_J, r_{I_c}, r_{I_p}\} = \{55, 9, 29\}$. Given this rank set, the percentage of variation (sum of squares) explained by the estimated joint structure, individual structure and residual noise for cell lines and patients' gene expression data is shown in Figure 2.4A. The joint structure is responsible for 59% of variation in both data, which quantify the level of similarities between the genomic profiles of the two biological model systems. The patient gene expression data have a substantial amount of structured variation (18%) that is unrelated to the cell lines, while cell lines' data only have 5% structured variation that is not shared with patients. This result is consistent with current biological understanding, as the *in vivo* environment is more complex than the *in vitro* environment (Shi et al., 2019) and various factors that govern it.

In addition, we made use of the biological interpretability of iPCR's predictors for model

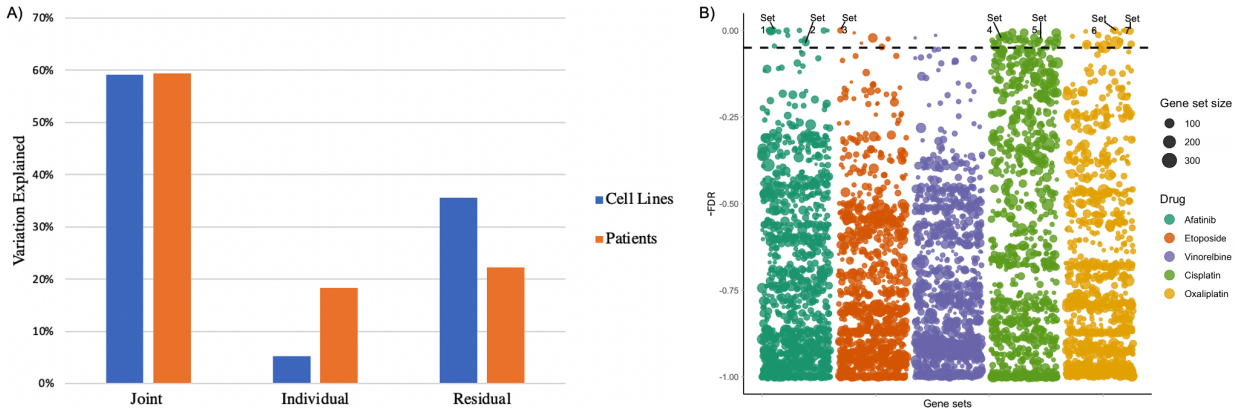


Figure 2.4: **Biological interpretations of iPCR's results. A)** Percentage of variation (sum of squares) explained by estimated joint structure (rank = 55), individual structure (rank = 9 for cell lines and rank = 29 for patients) and residual noise for cell lines and patients' gene expression data. **B)** Bubble plot shows –FDR for PreRanked gene set enrichment analysis (y-axis), based on the estimated gene effects on the patient responses to each of five drugs (in different colors). Each dot corresponds to a gene set, and those gene sets above the dashed line are considered significant (FDR < 0.05). The dot size represents the number of genes contained in a gene set. The significant gene sets that are discussed in Section 2.4.3 are highlighted. Note that Afatinib (for model training) is used as the matching drug of Trastuzumab.

validation. Our iPCR model allows an estimation of each gene's effect on the drug response and hence delineate the underlying biological mechanisms. Based on the ranked list of estimated gene effects, we conducted a PreRanked gene set enrichment analysis (Subramanian et al., 2005) for each of the five drugs where iPCR achieves AUC > 0.65, as shown in Figure 2.4B. We considered two collections of gene sets: 1) Hallmark and 2) Chemical and Genetic Perturbations (CGP). In Figure 2.4B, each dot corresponds to a gene set, and those gene sets above the dashed line are considered significant with The False Discovery Rate (FDR) < 0.05. Based on these significant gene sets identified by iPCR, we observed known drug response mechanisms determined by the expression levels of these genes.

To this end, we investigate seven relevant gene sets from our enrichment analysis that are highlighted in Figure 2.4B. First, the enrichment of two breast cancer-related gene sets, "Charafe Breast Cancer Luminal vs Mesenchymal Up" (Charafe-Jauffret et al., 2005) (Set 1 in Figure 2.4B) and "Creighton Endocrine Therapy Resistance 5" (Creighton et al., 2008) (Set 2) has a significant positive association with the sensitivity of Afatinib (FDR < 0.001 for Set 1; FDR $= 0.034$ for Set 2). Set 1 contains HER2 and HER3, and Set 2 contains HER2 and HER4. This finding is consistent with the known mechanism of Afatinib since these three genes are the main targets of Afatinib (Yonesaka et al., 2015). Second, the

overexpression of genes in the set "Cairo Hepatoblastoma Up" (Cairo et al., 2008) (Set 3) is significantly and positively associated with the sensitivity of Etoposide (FDR < 0.001). Set 3 includes the genes upregulated in hepatoblastoma, a cancer type that is typified by the upregulated MYC signaling. Third, the MYC-related gene sets, "Hallmark MYC Targets V1" (Liberzon et al., 2015) (Set 4) and "Dang Regulated By MYC Up" (Zeller et al., 2003) (Set 5) play a significant role in the sensitivity of Cisplatin with a positive direction (FDR $= 0.031$ for Set 4; FDR $= 0.024$ for Set 5). Our second and third findings are consistent with the past experimental results that Etoposide and Cisplatin provoked a strong apoptotic response in MYC-overexpressing cells (Albihn et al., 2006). Finally, the p53-related gene sets "Fischer Direct P53 Targets Meta Analysis" (Fischer et al., 2016) (Set 6), and "Hallmark P53 Pathway" (Liberzon et al., 2015) (Set 7) are significantly and positively associated with the sensitivity of oxaliplatin (FDR < 0.001 for Set 6; FDR $= 0.0016$ for Set 7); this corresponds to the past finding that colon cancer cells with a wild-type p53 exhibited a characteristic of being sensitive to Oxaliplatin (Toscano et al., 2007).

2.4.4 Gene co-expression network analyses

Using the iPCR results given the rank set $\{r_J, r_{I_c}, r_{I_p}\} = \{55, 9, 29\}$ (Table A.2), we constructed the shared gene co-expression networks (shown in Figure 2.5A) as well as cell line-specific and patient-specific networks (shown in Figures A.2 and A.3 respectively). Furthermore, the genes were clustered into six communities with different colors, based on the Louvain method (Blondel et al., 2008). We observed two highly dense communities colored in purple and green in the shared network. For downstream interpretations, we considered "hub" genes that significantly connect to more than 30 other genes. We summarize the gene connectivity levels, using a circular heatmap (Figure 2.5B) that visualizes a matrix with each entry representing the number of edges linked to a hub gene in the shared, cell line-specific or patient-specific network. The dendrogram of genes inside the heatmap shows that the connectivity pattern of genes differs across the shared and model-specific networks, and the shared network contains more hub genes with a high connectivity level (colored in red). The upset plot in Figure A.4 shows the number of hub genes in each type of networks as well as their intersections (Lex et al., 2014). The names of all hub genes are listed in Table A.3.

From our analyses, the top five hub genes in the shared network include RRM1, MCM6, RFC5, SPC25, and EXO1. The hub gene RRM1 provides the precursors that are essential for DNA synthesis, and is functionally associated with DNA replication and DNA damage repair (Kato et al., 2021). MCM6 encodes a part of the MCM complex, which is a set of proteins functioning as a helicase (Lewis et al., 2022). Helicases are essential during DNA

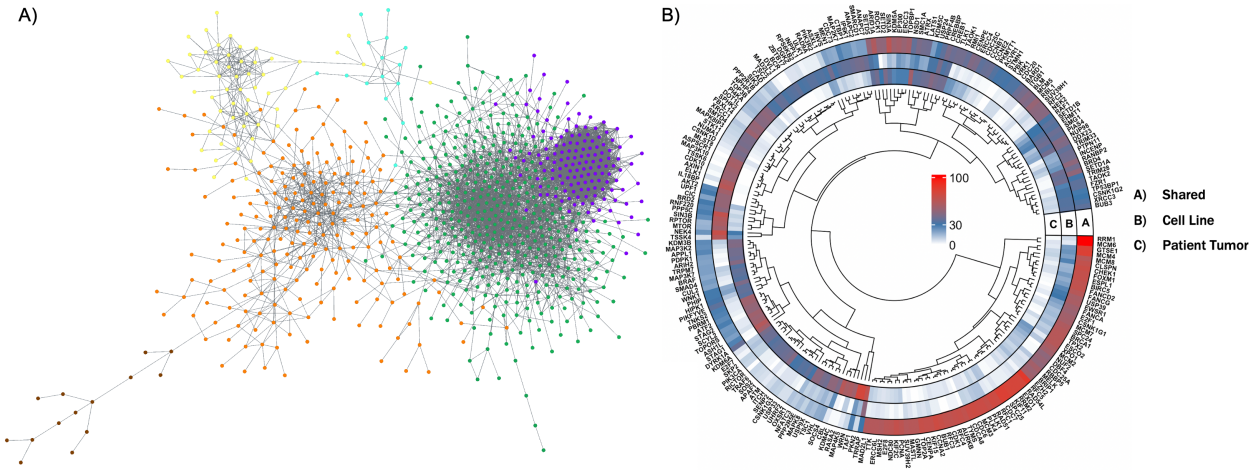


Figure 2.5: **Analysis of model-specific and shared gene co-expression networks between cell lines and patients. A)** Shared network between the two biological model systems, where each node represents a gene. Significantly associated genes are connected by edges. The gene communities in different colors are detected by an unsupervised clustering method. **B)** Circular heatmap of connection level of each gene in the shared, patient-specific, and cell line-specific networks. Color corresponds to the gene connection level, which is the number of edges linked to a gene (red, high; blue, medium; white low). Only the genes with a connection level > 30 in any of the networks are shown. The dendrogram of genes is provided inside the heatmap.

replication since they unwind the DNA so that it can be copied. RFC5 relates to homologous DNA pairing and strand exchange, which play a crucial role in repairing DNA damage during mitosis (Michiels et al., 2009). The upregulation of SPC25 can lead to genetic instability and potentially result in tumorigenesis; its overexpression has been observed in lung cancer, prostate cancer, colorectal cancer and gastric cancer (Su et al., 2021; Yang et al., 2022). EXO1 is involved in a number of DNA repair and metabolic pathways (Goellner et al., 2015; Keijzers et al., 2018). Hence, our estimated shared gene co-expression network between cell lines and patient tumors illustrates that DNA replication and repair are the two critical processes that relate to the pathogenesis of cancer.

2.5 Discussion

We present an integrative multi-system prediction model, iPCR, which serves four main objectives: 1) quantify structured similarities and differences in gene expression data between cell lines and patient tumors, 2) exploit patient-cell line similarities to predict patient drug response, 3) identify key genomic drivers for drug-specific patient response, and 4) infer both model-specific (cell lines or patients) and shared gene co-expression networks while detecting

hub genes within these networks. We demonstrate that accounting for model-specific information in the prediction model not only enhances the accuracy of drug response prediction in both simulation studies and real-world applications, but also improve the interpretability of shared biology and disparities between cancer model systems. In essence, iPCR provides a distinct advantage over existing domain adaptation methods, primarily due to its enhanced interpretability. For real-world prediction applications, iPCR represents an additional choice within the spectrum of domain adaptation methodologies for multi-source data.

Our iPCR model can be generalized in several directions. First, our framework can be potentially used to integrate multiple cancer model systems—such as cell lines, patient-derived xenografts (PDXs), and patient tumors—into a single statistical framework. This enables the investigation of their common biology and disparities while predicting patient drug responses. PDXs, derived from human tumors and cultivated in mice, are emerging as systems that more faithfully represent human tumors (Ben-David et al., 2017). However, there are still biological differences between PDXs and patient tumors, as PDXs may not consistently capture the dynamic shifts in copy number alterations found in primary human tumors. (Ben-David et al., 2017). One could reformulate the iPCR model based on Structural Learning and Integrative Decomposition of Multi-View Data (SLIDE) (Gaynanova and Li, 2019) to study the shared, partial-shared, and individual features among the three different but related cancer model systems. Second, our iPCR model could potentially be improved to allow out-of-sample prediction. In the current framework, although patient drug response data are not utilized in model training, our model only predicts drug response for patients whose genomic data are used for model training. For a new patient, the joint score can be obtained by projecting the patient's gene expression data onto the shared loading, which can then be utilized for predicting the drug response. However, both the quantification and preprocessing of the new patients' genomic data require careful consideration. Third, we acknowledge that exploring nonlinear dimension reduction in Stage I of iPCR, along with its integration with network estimation, remains a complex and largely unexplored research area. This presents a promising avenue for future investigation. Conversely, employing a nonlinear AI/ML prediction model at Stage II of iPCR, as opposed to an elastic net model, is feasible, although it could be less interpretable than a linear model. Finally, our shared gene co-expression network is constructed based on two joint sample covariance matrices. It would be of interest to refine the current sample covariance decomposition method to obtain a single joint sample covariance matrix for better interpretation. To achieve this goal, additional distributional assumptions might be required.

CHAPTER 3

Time-Varying and Covariate-Dependent (TiVAC) Correlation

3.1 Introduction

Multivariate longitudinal data often reveal that associations between outcomes are neither constant over time nor uniform across individuals. This phenomenon motivates the statistical question of modeling correlation as an explicit function of time (t) and covariates ($\mathbf{x}$), $\rho(t, \mathbf{x})$. As an example, in our psychiatry application context, depression and anxiety are commonly regarded as highly comorbid conditions(Kalin, 2020). Interestingly, in bipolar I disorder (BP-I) patients, the concurrent correlation between PHQ-9 (depression; (Kroenke et al., 2001)) and GAD-7 (anxiety; (Spitzer et al., 2006)) is heterogeneous across patients, with some exhibiting very high correlations and others low or near zero (Figure 3.1A). To further characterize this heterogeneity, we show that the correlation also varies with patient-specific covariates such as age and antipsychotic use frequency, as illustrated by the three-dimensional correlation surface in Figure 3.1B. These findings underscore the importance of accounting for both time-varying and covariate-dependent heterogeneity in correlation. This point is also highlighted by the Simpson's paradox: subgroup-level correlations can vanish or even reverse relative to a fixed population-level correlation (Blyth, 1972).

In the context of covariate-dependent correlations, Bartlett (1993) first considered correlation as a function of a discrete third variable. However, his method required large subgroup-specific sample sizes and was limited to a single discrete covariate. Subsequent extensions have allowed correlation to depend on both continuous and discrete covariates within a bivariate Gaussian framework, with parameters estimated via restricted maximum likelihood (REML) (Wilding et al., 2011) and second-order generalized estimating equations (GEE2) (Ho et al., 2010). More recently, Tu et al. (2022) introduced covariate-dependent pairwise correlation model with association size (CoCoA), which compares that REML and GEE2 estimators with maximum likelihood estimators (MLE). In contrast, other methods

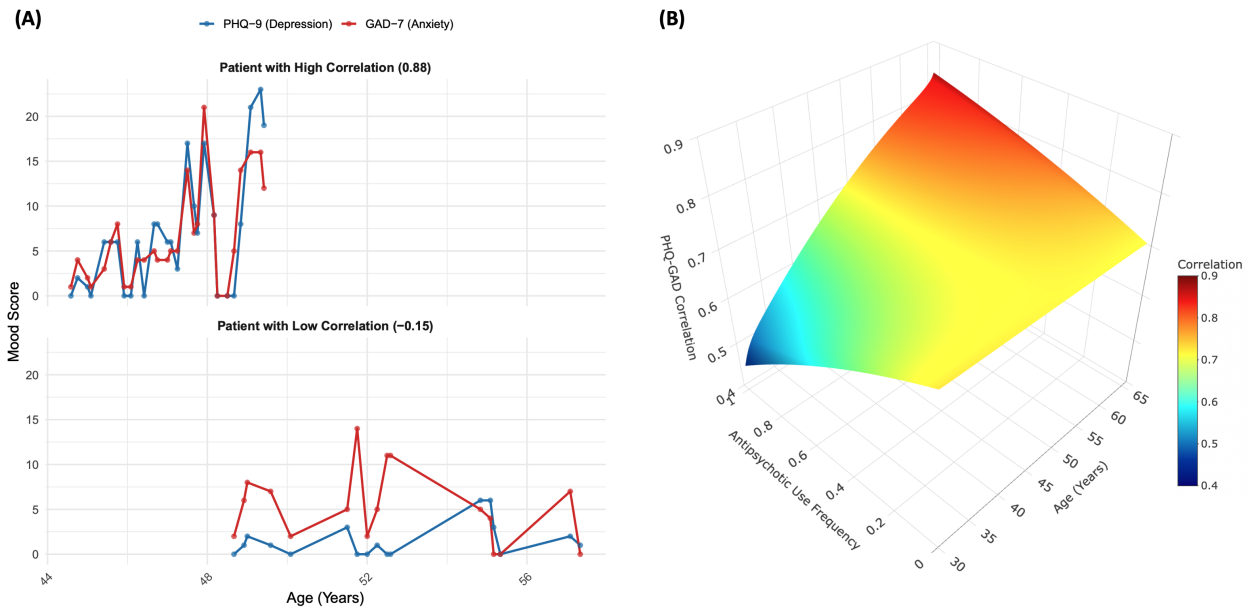


Figure 3.1: **PHQ-9 and GAD-7 trajectories and correlation heterogeneity. (A)** Irregular longitudinal trajectories of PHQ-9 (depression; blue) and GAD-7 (anxiety; red) for two representative BP-I patients. The concurrent, within-patient Pearson correlation is high for the top patient ($\rho = 0.88$) and low for the bottom patient ($\rho = -0.15$). **(B)** Three-dimensional correlation surface showing how the PHQ–GAD correlation (z-axis) varies with age (x-axis) and antipsychotic-use frequency (y-axis). Color encodes the correlation value through rainbow scale.

(Pourahmadi, 1999; Chiu et al., 1996; Hoff and Niu, 2012; Archakov and Hansen, 2021) model covariate-dependent covariance or correlation for more than two outcomes through parameterizations or transformations that ensure the covariance or correlation matrix remains positive definite. Covariate-dependent precision matrices have also been studied in graphical regression models with covariates (Zhang and Li, 2022).

Despite these advancements, the aforementioned models do not address how correlations might evolve over time. Estimating covariate-dependent correlations separately at each time point can substantially reduce statistical power, particularly when data are irregularly observed. Although one might consider treating time as a single covariate (Wilding et al., 2011; Tu et al., 2022), this strategy imposes a linearity assumption on the time-varying effects, which may obscure potentially dynamic and nonlinear patterns in the correlation structure. For instance, in our motivating application, Figure B.4 displays the monthly Pearson correlation between the depression and anxiety severity across different ages. When examining the full cohort or stratifying by gender, the estimated correlation trajectories reveal significant noise that masks the underlying true dynamic pattern, thereby motivating

the use of a flexible nonlinear modeling approach. Although several statistical approaches–including DCC-GARCH (Engle, 2002), DSEM (Asparouhov et al., 2017), and idPAC (Kundu et al., 2021)–have been proposed to capture time-varying relationships among variables, none directly or jointly model how covariates affect correlations over time and provide statistical inference on these time-varying effects.

To this end, we propose a Time-Varying and Covariate-Dependent (TiVAC) correlation model, which jointly models time-varying and covariate-dependent correlations. TiVAC focuses on concurrent correlations between two outcomes over time and facilitates statistical inference on how covariate effects on correlation might change with time. Under a bivariate Gaussian model, we apply a Fisher transformation to the correlation parameter and formulate the transformed correlation as a linear function of covariates with coefficients that vary over time. To capture the dynamic nature of these coefficients, we employ functional data analysis methods using spline-based constructions, which offer two main benefits: flexible modeling of covariate and time effects, and borrowing strength across adjacent time points. For estimation, we develop a computationally efficient strategy based on design-matrix reconstruction. We adopt penalization to ensure spline smoothness and interpretability, and maximize the penalized likelihood with a Newton–Raphson algorithm. To test time-varying and covariate functional effects, we construct simultaneous confidence bands to control for multiple testing. Through simulation studies, we compare TiVAC with competing approaches and show that it consistently outperforms them in estimating $\rho(t, \mathbf{x})$ across various time-dependent correlation functions and covariate types, under both correctly specified and misspecified models. Finally, we apply TiVAC to mood trajectory data from 291 BP-I patients enrolled in the Heinz C. Prechter Bipolar Research Project (PrBP) (McInnis et al., 2017; Yocum et al., 2023), examining how patient characteristics and clinical interventions affect the correlation between PHQ-9 and GAD-7 scores across ages 30 to 65. Our results reveal that the overall correlation between depression and anxiety increases slightly with age. Notably, males exhibit higher correlations than females at younger ages. Additionally, while increased usage of nervous system-related medications and antipsychotics is linked to lower correlations in early adulthood, this relationship reverses at older ages; in contrast, mood stabilizer use is consistently associated with lower correlations.

The remainder of the paper is organized as follows. Section 3.2 details the TiVAC model formulation. Sections 3.3.1 and 3.3.2 describe the spline representation and estimation procedure of coefficients, and Section 3.3.3 outlines the construction of simultaneous confidence bands for inference. Simulation results for correctly specified data are presented in Section 3.4.1, while Section 3.4.2 investigates model performance under incorrect specifications. Section 3.5 demonstrates the practical utility of TiVAC in the BP-I patient mood trajectory

study. Section 3.6 concludes with a discussion of our findings and future research directions.

3.2 TiVAC model

In this section, we outline the data structure and model framework. The observed data consist of two main components: a pair of concurrent and irregular longitudinal outcomes and a set of p covariates. Specifically, let n denote the number of individuals. For each individual i, let

$$\mathbf{Y}_i = \big[\{\, Y_{i1}(t_{i1}),\, Y_{i2}(t_{i1})\,\},\, \ldots,\, \{\, Y_{i1}(t_{im_i}),\, Y_{i2}(t_{im_i})\,\}\big]$$

represent the vector of observed outcomes, where $Y_{i1}(t_{ij})$ and $Y_{i2}(t_{ij})$ are the first and second outcomes for individual i at time t_{ij}, m_i is the number of observation times for individual i, and t_{ij} denotes the j-th observation time for individual i. Let $\mathbf{Y} = (\mathbf{Y}_1^T, \ldots, \mathbf{Y}_n^T)$ represent the collection of observed outcomes for all individuals. Additionally, let $\mathbf{Y}_i(t_{ij}) = \{\, Y_{i1}(t_{ij}),\, Y_{i2}(t_{ij})\,\}^T$ represent the bivariate outcome pair for individual i at time t_{ij}. The covariate information is represented by a design matrix $\mathbf{X}$ of dimensions $n \times p$, where $n \gg p$ (with p low-dimensional), and $\boldsymbol{X}_i$ is the p-dimensional covariate vector for individual i.

For each individual i at time t_{ij}, we assume the concurrently observed outcomes follow a bivariate normal distribution with a covariate-dependent and time-varying correlation:

$$\begin{bmatrix} Y_{i1}(t_{ij}) \\ Y_{i2}(t_{ij}) \end{bmatrix} \sim \mathcal{N}\big\{\mathbf{0}, \boldsymbol{\Sigma}(t_{ij}, \boldsymbol{X}_i)\big\}, \quad \text{where} \quad \boldsymbol{\Sigma}(t_{ij}, \boldsymbol{X}_i) = \begin{bmatrix} \sigma_1^2 & \rho(t_{ij}, \boldsymbol{X}_i)\,\sigma_1\,\sigma_2 \\ \rho(t_{ij}, \boldsymbol{X}_i)\,\sigma_1\,\sigma_2 & \sigma_2^2 \end{bmatrix}. \tag{3.1}$$

Here, $\rho(t_{ij}, \boldsymbol{X}_i)$ referred to as the TiVAC correlation, depends explicitly on both the observation time t_{ij} and the covariate vector $\boldsymbol{X}_i$. We emphasize the dynamic nature of concurrent correlation by allowing $\rho(t_{ij}, \boldsymbol{X}_i)$ to vary continuously over time.

Since ρ is constrained within the interval $[-1, 1]$, the Fisher transformation is employed to map it onto the real line:

$$\eta(t_{ij}, \boldsymbol{X}_i) = \ln\left\{\frac{1 + \rho(t_{ij}, \boldsymbol{X}_i)}{1 - \rho(t_{ij}, \boldsymbol{X}_i)}\right\}.$$

The transformed correlation $\eta(t_{ij}, \boldsymbol{X}_i)$ is then modeled via a varying-coefficient linear form,

$$\eta(t_{ij}, \boldsymbol{X}_i) = \boldsymbol{X}_i^T\, \boldsymbol{\beta}(t_{ij}), \tag{3.2}$$

where $\boldsymbol{\beta}(t_{ij})$ is a coefficient vector function of time. Of particular interest, the k-th element

of $\beta(t_{ij})$, denoted as $\beta_k(t_{ij})$, represents the time-varying effect of the k-th covariate on the correlation. This specification allows the concurrent correlation to depend jointly on time and covariates through interpretable and nonlinear coefficient trajectories, and can be construed as a generalization of CoCoA model (Tu et al., 2022).

Building on the specification above, we impose the following additional assumptions to ensure identifiability and interpretability of TiVAC model: (1) each outcome pair has zero mean, which can be achieved, for example, through a flexible demeaning approach such as the one proposed by Patil et al. (2022); (2) the marginal variances are constant over time, denoted by σ_1^2 and σ_2^2, reflecting our focus on the correlation rather than marginal variability; and (3) there is no autocorrelation across distinct time points, thereby focusing on dynamic heterogeneity in the concurrent correlation while balancing model complexity and computational efficiency.

3.3 Estimation

To obtain a flexible and smooth estimator $\hat{\beta}_k(t)$, we represent the coefficient functions using B-splines (de Boor, 2001), with known basis functions and spline coefficients to be estimated, as described in Section 3.3.1. Because direct maximum-likelihood estimation of spline coefficients is sensitive to the number and placement of knots, we incorporate a smoothing penalty on the spline coefficients, ensuring stability while retaining the ability to capture significant local changes (Section 3.3.2). The variances σ_1^2 and σ_2^2 are estimated separately prior to estimating $\boldsymbol{\theta}$; specifically, we pool all longitudinal observations across individuals to compute the sample variances for each outcome, although any other appropriate method for estimating variances could also be used. Finally, to address multiple comparisons arising from evaluating covariate effects simultaneously across numerous time points, we develop a bootstrap-based inference framework to construct simultaneous confidence bands, ensuring that the entire trajectory of covariate effects lies within the bands with a specified confidence level (Section 3.3.3).

3.3.1 Spline representation of functional covariate effects

Definition of B-spline. Direct estimation of $\boldsymbol{\beta}(t_{ij})$ is not feasible, as it is infinite-dimensional. To address this, we approximate it using a finite-dimensional B-spline representation (de Boor, 2001). B-splines are piecewise polynomial functions, with the order of the spline determining the degree of the polynomial; a common choice is cubic B-splines (order 4). The number and placement of knots control the flexibility of the spline–more knots allow

for greater flexibility, while fewer knots lead to smoother estimates. Specifically, let $\boldsymbol{b}(t_{ij})$ denote a q-dimensional vector of B-spline basis functions evaluated at time t_{ij}, where q depends on the number of knots and the spline order. For the i-th individual, the B-spline basis functions evaluated at all observed time points are organized into a spline basis matrix, $\mathbf{B}_i = \left\{\boldsymbol{b}(t_{i1}) \ \cdots \ \boldsymbol{b}(t_{im_i})\right\}^\top \in \mathbb{R}^{m_i \times q}$.

Incorporating B-splines into TiVAC. For each covariate k, the corresponding time-dependent component $\beta_k(t_{ij})$ of $\boldsymbol{\beta}(t_{ij})$ can then be expressed as:

$$\beta_k(t_{ij}) \approx \boldsymbol{b}(t_{ij})^T \boldsymbol{\theta}_k, \tag{3.3}$$

where $\boldsymbol{\theta}_k$ is a q-dimensional vector of spline coefficients. Substituting Equation 3.3 into Equation 3.2, the only parameters to estimate are the spline coefficients, $\boldsymbol{\theta}_k$, for $k = 1, \ldots, p$. To simplify the representation, we combine the spline bases and covariates into a single unified structure. Specifically, for the i-th individual at time t_{ij}, we define a new covariate vector $\mathbf{A}_{i,j} = \boldsymbol{X}_i^T \otimes \boldsymbol{b}(t_{ij})$ with length $q \cdot p$, where $\otimes$ denotes the Kronecker product. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T, \ldots, \boldsymbol{\theta}_p^T)^T$, a vector of length $q \cdot p$, represent all spline coefficients across the p covariates. The Fisher-transformed correlation for individual i at time t_{ij} can then be expressed as:

$$\eta(t_{ij}, \boldsymbol{X}_i) \approx \mathbf{A}_{i,j}^T \boldsymbol{\theta}.$$

After $\boldsymbol{\theta}$ is estimated, it can be substituted into Equation 3.3 to obtain the interpretable coefficient functions $\beta_k(t_{ij})$.

3.3.2 Spline coefficient estimation

Smoothness penalty on spline coefficients. To estimate the spline coefficients $\boldsymbol{\theta}$, we maximize a penalized log-likelihood function that incorporates a smoothness penalty on the B-spline coefficients. The penalty term ensures stability in the estimation process and prevents overfitting. Rather than using the traditional integral of squared derivatives, we adopt a difference penalty on adjacent spline coefficients, as proposed by Eilers and Marx (1996). This discrete penalty simplifies computation while providing a close approximation to the smoothness measure achieved by integral-based methods. In this paper, we employ the second-order difference penalty, defined as:

$$P(\boldsymbol{\theta}_k) = \frac{1}{2} \lambda_k \boldsymbol{\theta}_k^T \mathbf{D}_2^T \mathbf{D}_2 \boldsymbol{\theta}_k,$$

where $\mathbf{D}_2$ represents the second-order difference operator matrix, and $\lambda_k > 0$ is the smoothing parameter. Specifically, $\mathbf{D}_2 \boldsymbol{\theta}_k$ yields a vector whose j-th element is $\theta_{k,j} - 2\theta_{k,j+1} + \theta_{k,j+2}$. Thus,

this penalty encourages adjacent spline coefficients to remain close, penalizing curvature and promoting smoothness in the estimated coefficient functions. Higher-order difference penalties can be easily extended within this framework. As suggested by Eilers and Marx (1996), using a relatively large number of knots is advisable. This strategy eliminates the need to balance low versus high numbers of knots, which in unpenalized methods can yield drastically different smoothness levels. However, we caution against using an excessively dense grid of knots, as this can inflate the number of parameters and lead to insufficient data between adjacent knots, potentially complicating model fitting.

Estimation of spline coefficients. Incorporating the smoothness penalty into the log-likelihood framework, the penalized log-likelihood function for the spline coefficients is given by:

$$l_{\text{penalized}}(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{A}) = \log L(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{A}) - \frac{1}{2} \sum_{k=1}^{p} \lambda_k \boldsymbol{\theta}_k^T \mathbf{D}_2^T \mathbf{D}_2 \boldsymbol{\theta}_k. \tag{3.4}$$

Here, $\log L(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{A})$ represents the log-likelihood function for all individuals, incorporating their m_i bivariate outcomes:

$$\log L(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{A}) = -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \{\log(2\pi) + \log |\boldsymbol{\Sigma}(t_{ij}, \boldsymbol{X}_i)| + \mathbf{Y}_i(t_{ij})^T \boldsymbol{\Sigma}(t_{ij}, \boldsymbol{X}_i)^{-1} \mathbf{Y}_i(t_{ij})\}.$$

To estimate the spline coefficients $\boldsymbol{\theta}$, we maximize the penalized log-likelihood function defined in Equation 3.4 with respect to $\boldsymbol{\theta}$, given certain penalized parameters λ_k, $k = 1, \ldots, p$:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \, l_{\text{penalized}}(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{A}).$$

We employ the Newton–Raphson algorithm to iteratively update $\boldsymbol{\theta}$; detailed formulas are provided in Appendix Section B.1.

Selection of smoothing parameters. The algorithm described above does not select smoothing parameters. The smoothing parameters λ_k $(k = 1, \ldots, p)$ are optimized through cross-validation (default 10 folds) by maximizing the log-likelihood $l(\hat{\boldsymbol{\theta}}; \mathbf{Y}, \mathbf{A})$ on the testing data, where $\hat{\boldsymbol{\theta}}$ here is estimated using the training data. The candidate values for the smoothing parameters are chosen on a case-by-case basis, following a sequence such as $0, 10, 10^2, \ldots, 10^6$. Grid search can be employed when the number of covariates is ≥ 2; however, parallel computing is recommended to alleviate the computational burden when the number of covariates is large.

Estimation of the TiVAC correlation. After obtaining the optimal smoothing parameters, the model is fit using all available samples to estimate the optimal $\hat{\boldsymbol{\theta}}$. Let $\mathbf{B}$ be the spline basis matrix evaluated at all time points from 1 to $\max_{i,j}(t_{ij})$, having dimensions $\max_{i,j}(t_{ij}) \times q$,

where $\max_{i,j}(t_{ij})$ denotes the maximum observed time across all individuals. The estimated coefficient trajectory $\hat{\beta}_k(t)$, for $t = 1, \ldots, \max_{i,j}(t_{ij})$, is given by

$$\hat{\beta}_k(t) = \mathbf{B}\,\hat{\boldsymbol{\theta}}_k,$$

which measures the k-th covariate's time-varying effect on the correlation. The correlation trajectory for covariate $\boldsymbol{X}_i$ can then be expressed as:

$$\hat{\rho}(t, \boldsymbol{X}_i) = \tanh\left\{\sum_k X_{ik}\hat{\beta}_k(t)\right\},$$

where X_{ik} is the k^{th} covariate for individual i.

3.3.3 Simultaneous confidence band estimation

After estimating $\beta_k(t)$, it is important to conduct inference to determine whether the covariate effect on the correlation at certain time points is significant. Instead of using pointwise confidence bands, which only guarantee confidence at individual time points, we use simultaneous confidence bands that ensure the entire trajectory of a function lies within the band with a specified confidence level. Mathematically, a $100(1-\alpha)\%$ simultaneous confidence band satisfies:

$$P\left\{L(t) \leq \beta_k(t) \leq U(t),\, \forall\, t \in T\right\} \geq 1 - \alpha \text{ for all } t \in T.$$

This approach is crucial for addressing multiple comparison issues and provides broader coverage compared to pointwise confidence bands.

We construct simultaneous confidence bands for $\beta_k(t)$ using a bootstrap procedure that resamples individuals (with replacement), preserving the observed time points within each individual. Specifically, our procedure consists of two nested bootstrap loops, each serving a distinct purpose. The outer bootstrap loop quantifies the variability of the estimator $\hat{\beta}_k(t)$ by repeatedly drawing bootstrap samples $\mathbf{Y}^{(b)}$ $(b = 1, \ldots, B)$ and re-estimating the spline coefficients, yielding bootstrap estimates $\hat{\beta}_k^{(b)}(t)$. To avoid underestimating the variability of the outer bootstrap estimates, we use an inner bootstrap loop to provide robust estimates of their standard deviations. Specifically, in each outer iteration b, we perform M additional bootstrap resamples $\mathbf{Y}^{(b,m)}$ $(m = 1, \ldots, M)$ from the outer sample $\mathbf{Y}^{(b)}$, computing inner-loop estimates $\hat{\beta}_k^{(b,m)}(t)$. From these inner bootstrap samples, we compute the sample standard deviation $\tilde{s}_{k,t}$ for each outer iteration. Next, we compute the t-statistic for

each outer iteration b following Ruppert et al. (2010): $T_k^{(b)} = \max_t \left| \frac{\hat{\beta}_k^{(b)}(t) - \hat{\beta}_k(t)}{\tilde{s}_{k,t}} \right|$. The critical value T_k^{crit} is then obtained as the empirical $(1-\alpha)$-quantile of the collection of t-statistics $\{T_k^{(b)}\}_{b=1}^B$. As noted by Ruppert et al. (2010), when the sample size is large, T_k^{crit} can be approximated very accurately. Finally, the simultaneous confidence band is constructed as: $\hat{\beta}_k(t) \pm T_k^{\text{crit}} \cdot s_{k,t}$, where $s_{k,t}$ is the overall bootstrap standard deviation of $\hat{\beta}_k^{(b)}(t)$ across all outer bootstrap iterations $b = 1, \ldots, B$. A detailed summary of the entire procedure is provided in Algorithm 2.

Algorithm 2 Simultaneous confidence band estimation

Input: Observed data $\mathbf{Y}$, estimated coefficients $\hat{\beta}_k(t)$, number of outer bootstrap iterations B, number of inner bootstrap samples M, and confidence level $1-\alpha$.
Outer Bootstrap Sampling:
for $b = 1, \ldots, B$ **do**
    Draw bootstrap sample $\mathbf{Y}^{(b)}$ and compute $\hat{\beta}_k^{(b)}(t)$.
    Inner Bootstrap Sampling (estimate standard errors):
    for $m = 1, \ldots, M$ **do**
        Draw bootstrap sample $\mathbf{Y}^{(b,m)}$ based on $\mathbf{Y}^{(b)}$ and compute $\hat{\beta}_k^{(b,m)}(t)$.
    end for
    Compute the sample standard deviation $\tilde{s}_{k,t}$ of $\hat{\beta}_k^{(b,m)}(t)$ across $m = 1, \ldots, M$.
    Compute the t-statistic:

$$T_k^{(b)} = \max_t \left| \frac{\hat{\beta}_k^{(b)}(t) - \hat{\beta}_k(t)}{\tilde{s}_{k,t}} \right|.$$

end for
Compute the sample standard deviation $s_{k,t}$ of $\hat{\beta}_k^{(b)}(t)$ across $b = 1, \ldots, B$.
Determine T_k^{crit} as the empirical $(1-\alpha)$-quantile of $\{T_k^{(b)}\}_{b=1}^B$.
Construct the simultaneous confidence band:

$$\hat{\beta}_k(t) \pm T_k^{\text{crit}} \cdot s_{k,t}.$$

3.4 Simulation study

We evaluate the estimation accuracy of the TiVAC correlation under four generative models, characterized by whether the univariate covariate is binary or continuous and whether the

underlying correlation structure follows a deterministic function or is influenced by random noise. For penalty parameter selection, we employ 10-fold cross-validation to determine the optimal value.

We compare the TiVAC correlation to three alternative methods. The first is the "Empirical" correlation trajectory, calculated by computing the Pearson correlation between the two outcomes at each observed time point within each covariate group. The resulting correlations are then smoothed over time using LOESS regression, with candidate span values ranging from 0.1 to 1 in increments of 0.1. The optimal span is selected via 5-fold cross-validation. The second method is covariance regression (CovReg) (Hoff and Niu, 2012), where the covariate-dependent covariance matrix of the two outcomes is calculated at each time point, transformed into a correlation matrix, and subsequently smoothed using the same LOESS approach as in the Empirical method. The third method is CoCoA (Tu et al., 2022), implemented with three different estimation techniques: MLE, REML (Wilding et al., 2011), and GEE2 (Ho et al., 2010). The resulting estimates are also smoothed using the same LOESS approach as in the Empirical method. Following the recommendation of Tu et al. (2022) to use CoCoA-REML, we focus primarily on its performance in the main text; results for CoCoA-MLE and CoCoA-GEE2 are provided in the Appendix Section B.2.

For the simulation settings, we fix the sample size at $n = 300$ and set $\max_{i,j}(t_{ij}) = 500$. Each individual i has m_i observed time points drawn under one of four conditions: (1) a fixed 40 time points ($\mathscr{T}_{High}$); (2) a random number of time points between 3 and 40 ($\mathscr{T}_{Moderate}$), mimicking the observed PrBP data; (3) a random number of time points between 3 and 10 ($\mathscr{T}_{Low}$); and (4) an alternative setting in which the number of individuals and time points matches those of the observed PrBP data, with $\max_{i,j}(t_{ij}) = 421$ (see Section 3.5), denoted $\mathscr{T}_{PrBP}$. The outcomes $Y_{i1}(t_{ij})$ and $Y_{i2}(t_{ij})$ follow a bivariate normal distribution (Equation 3.1) with variances $\sigma_1^2 = 1$ and $\sigma_2^2 = 4$.

To generate the time- and covariate-dependent correlation $\rho(t_{ij}, X_i)$, we apply the Fisher transformation in two separate data-generating models: a deterministic model,

$$\eta(t_{ij}, X_i) \;=\; \beta_0(t_{ij}) \;+\; \beta_1(t_{ij})\, X_i,$$

and a noisy model,

$$\eta(t_{ij}, X_i) \;=\; \beta_0(t_{ij}) \;+\; \beta_1(t_{ij})\, X_i \;+\; \epsilon(t_{ij}),$$

where $\epsilon(t_{ij}) \sim N\big(0,\, \sigma_{\text{noise}}\big)$. We consider five levels of σ_{noise}: 0.1, 0.2, 0.3, 0.4, and 0.5. The coefficient functions $\beta_k(t)$ $(k = 0, 1)$ are generated using one of three trajectory patterns: linear, seasonal, or logistic. The true correlation curves, $\tanh\{\beta_0(t) + \beta_1(t)\}$, are shown as dashed lines in Figure 3.2.

For model evaluation, we compute the root mean squared error (RMSE) by comparing each method's estimated correlation to the true correlation curve over 50 replications. In the binary covariate scenario, RMSE is averaged across time points for each covariate value (0 or 1), reflecting the deviation from the true correlation trajectory in each group. In the continuous covariate scenario, RMSE is computed by averaging over both time points and a grid of 100 covariate values ranging from 0 to 1.

3.4.1 Simulations with deterministic correlation functions

In Scenario 1, the underlying correlation function is deterministic, and the covariate is binary ($X = 0$ for half the sample and $X = 1$ for the other half). Table 3.1 presents the RMSE and SD results for TiVAC, Empirical, CovReg, and CoCoA-REML across different coefficient function types and time-point settings. Overall, TiVAC achieves the highest estimation accuracy under all settings, with its advantage becoming more pronounced when fewer time points are observed. Under $\mathscr{T}_{High}$, the RMSE differences between TiVAC and the other methods are modest (up to 0.02). By contrast, in the $\mathscr{T}_{Low}$ setting, Empirical, CovReg, and CoCoA-REML exceed TiVAC's RMSE by more than 0.1 in most cases. TiVAC also exhibits smaller standard deviations (SD) than the competing methods, reflecting its stability. TiVAC's outperformance could stem from its capacity to borrow correlation information across time points via penalized splines. Under the $\mathscr{T}_{Low}$ setting, CovReg performs worse than Empirical, and CoCoA-REML exhibits even poorer performance. We suspect that CovReg and CoCoA-REML, as model-based approaches, are particularly sensitive to data sparsity. In practice, when only a few observations are available at certain time points, the iterative optimization in model-based methods may fail to converge, leading to missing estimates of correlation.

Figure 3.2 compares the estimated correlation curves from the four methods to the true curve when $X = 1$ in Scenario 1. Each subplot corresponds to a representative example from the first of 50 simulation runs (using a fixed random seed). The figure reveals two additional insights. First, TiVAC better captures the tail behavior of the true correlation. For instance, under the $\mathscr{T}_{PrBP}$ logistic trajectory, Empirical, CovReg, and CoCoA-REML exhibit an upward trend on the right tail, whereas the true curve and TiVAC flatten out. Second, TiVAC's spline-based penalty preserves a level of smoothness that aligns closely with the true trajectory, while pointwise-based estimates can become overly smoothed (e.g., under the $\mathscr{T}_{Low}$ seasonal trajectory) or fluctuate sharply (e.g., CoCoA-REML under the $\mathscr{T}_{PrBP}$ logistic trajectory). One reason is that the large variability of the pointwise estimates before smoothing makes it difficult for LOESS regression to recover the underlying correlation's

true smoothness.

Table 3.1: **Simulation results for Scenario 1**. Simulation results comparing the performance of TiVAC, Empirical, CovReg, and CoCoA-REML methods under different coefficient function types and number of observed time points per individual. The equation of correlation is assumed to be deterministic and dependent on a binary correlation function. The time-varying correlation is $\rho_0(t)$ when $X = 0$ and $\rho_1(t)$ when $X = 1$. The table presents the RMSE with SD for each method.

Coefficient	# of Time	Correlation	RMSE (SD)			
Functions	Points	Functions	TiVAC	Empirical	CovReg	CoCoA-REML
Linear	$\mathscr{T}_{Low}$	$\rho_0(t)$	0.02 (0.01)	0.16 (0.04)	0.20 (0.10)	0.36 (0.03)
		$\rho_1(t)$	0.02 (0.01)	0.15 (0.04)	0.20 (0.09)	0.34 (0.04)
	$\mathscr{T}_{Moderate}$	$\rho_0(t)$	0.01 (0.01)	0.06 (0.02)	0.04 (0.01)	0.06 (0.02)
		$\rho_1(t)$	0.01 (0.01)	0.06 (0.02)	0.04 (0.02)	0.07 (0.02)
	$\mathscr{T}_{High}$	$\rho_0(t)$	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.03 (0.01)
		$\rho_1(t)$	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.02 (0.01)
	$\mathscr{T}_{PrBP}$	$\rho_0(t)$	0.01 (0.01)	0.04 (0.02)	0.04 (0.02)	0.09 (0.02)
		$\rho_1(t)$	0.02 (0.01)	0.06 (0.02)	0.05 (0.02)	0.10 (0.02)
Seasonal	$\mathscr{T}_{Low}$	$\rho_0(t)$	0.06 (0.02)	0.12 (0.04)	0.17 (0.07)	0.08 (0.03)
		$\rho_1(t)$	0.06 (0.02)	0.16 (0.03)	0.19 (0.06)	0.22 (0.03)
	$\mathscr{T}_{Moderate}$	$\rho_0(t)$	0.04 (0.01)	0.06 (0.02)	0.07 (0.02)	0.05 (0.02)
		$\rho_1(t)$	0.03 (0.01)	0.07 (0.02)	0.07 (0.02)	0.07 (0.01)
	$\mathscr{T}_{High}$	$\rho_0(t)$	0.03 (0.01)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)
		$\rho_1(t)$	0.03 (0.01)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)
	$\mathscr{T}_{PrBP}$	$\rho_0(t)$	0.03 (0.01)	0.05 (0.02)	0.07 (0.02)	0.05 (0.02)
		$\rho_1(t)$	0.03 (0.01)	0.08 (0.02)	0.08 (0.02)	0.08 (0.02)
Logistic	$\mathscr{T}_{Low}$	$\rho_0(t)$	0.05 (0.02)	0.14 (0.04)	0.19 (0.06)	0.36 (0.04)
		$\rho_1(t)$	0.05 (0.02)	0.16 (0.05)	0.22 (0.08)	0.34 (0.04)
	$\mathscr{T}_{Moderate}$	$\rho_0(t)$	0.03 (0.01)	0.06 (0.01)	0.05 (0.01)	0.07 (0.02)
		$\rho_1(t)$	0.03 (0.01)	0.06 (0.01)	0.05 (0.01)	0.07 (0.02)
	$\mathscr{T}_{High}$	$\rho_0(t)$	0.02 (0.01)	0.03 (0.01)	0.03 (0.01)	0.03 (0.01)
		$\rho_1(t)$	0.02 (0.01)	0.03 (0.01)	0.03 (0.01)	0.03 (0.01)
	$\mathscr{T}_{PrBP}$	$\rho_0(t)$	0.03 (0.01)	0.05 (0.01)	0.04 (0.02)	0.10 (0.02)
		$\rho_1(t)$	0.03 (0.01)	0.06 (0.02)	0.06 (0.02)	0.09 (0.02)

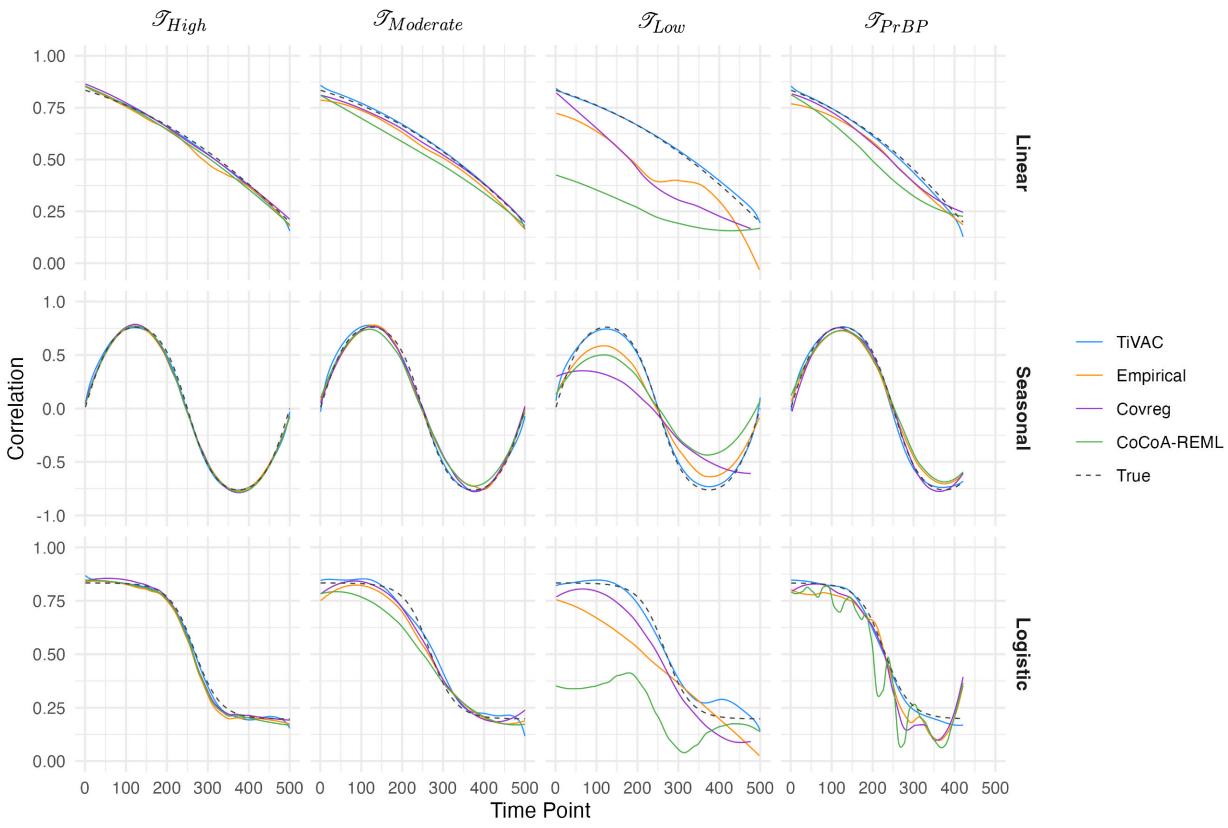


Figure 3.2: **Representative curve estimations in Scenario 1.** Comparison between the estimated correlation curves from four methods (TiVAC, Empirical, CovReg and CoCoA-REML) and the true correlation curve when covariate $X = 1$. The figure presents nine independent cases, representing three types of coefficient functions and three settings for the number of time points per individual. Each case shows a representative example from the first run of the simulation out of 50 replications (using a fixed random seed for all cases). The TiVAC, Empirical, CovReg and CoCoA-REML methods are depicted as blue, orange, purple and green curves, respectively, while the true trajectory is represented by a dashed black curve.

Furthermore, Table B.1 presents simulation results comparing the performance of CoCoA-MLE, CoCoA-REML, and CoCoA-GEE2 in Scenario 1. Under the $\mathscr{T}_{Low}$ setting, CoCoA-GEE2 outperforms the other two estimators, while in other settings their performances are similar. Nevertheless, under the $\mathscr{T}_{Low}$ setting, its performance remains inferior to that of TiVAC, Empirical, and CovReg.

Table 3.2: **Simulation results for Scenario 2.** Simulation results comparing the performance of TiVAC, CovReg, and CoCoA-REML methods under different coefficient function types and number of observed time points per individual. The equation of correlation is assumed to be deterministic and dependent on a continuous covariate. The table presents the RMSE with SD for each method.

Coefficient Functions	# of Time Points	RMSE (SD)		
		TiVAC	CovReg	CoCoA-REML
Linear	$\mathscr{T}_{Low}$	0.03 (0.01)	0.21 (0.04)	0.15 (0.02)
	$\mathscr{T}_{Moderate}$	0.02 (0.01)	0.09 (0.01)	0.06 (0.01)
	$\mathscr{T}_{High}$	0.01 (<0.01)	0.05 (0.01)	0.03 (0.01)
	$\mathscr{T}_{PrBP}$	0.02 (0.01)	0.10 (0.01)	0.09 (0.01)
Seasonal	$\mathscr{T}_{Low}$	0.04 (0.01)	0.21 (0.05)	0.12 (0.02)
	$\mathscr{T}_{Moderate}$	0.03 (0.01)	0.09 (0.01)	0.07 (0.01)
	$\mathscr{T}_{High}$	0.03 (<0.01)	0.05 (0.01)	0.04 (0.01)
	$\mathscr{T}_{PrBP}$	0.03 (0.01)	0.09 (0.01)	0.07 (0.01)
Logistic	$\mathscr{T}_{Low}$	0.05 (0.01)	0.21 (0.04)	0.16 (0.02)
	$\mathscr{T}_{Moderate}$	0.03 (0.01)	0.09 (0.01)	0.06 (0.01)
	$\mathscr{T}_{High}$	0.02 (<0.01)	0.05 (0.01)	0.03 (0.01)
	$\mathscr{T}_{PrBP}$	0.03 (0.01)	0.09 (0.02)	0.09 (0.01)

In Scenario 2, we adopt the same overall simulation design as in Scenario 1 but treat the covariate as continuous, ranging from 0 to 1. Unlike in Scenario 1, we fix each individual's number of observed time points at 10 to prevent CovReg from failing to converge under the $\mathscr{T}_{Low}$ setting. Because Empirical cannot accommodate continuous covariates, we focus on TiVAC, CovReg, and CoCoA-REML. Table 3.2 summarizes the RMSE results across different coefficient function types and time-point settings. TiVAC consistently outperforms CovReg and CoCoA-REML. Moreover, TiVAC exhibits similar performance in both Scenario 1 and Scenario 2 when the other settings are comparable, underscoring its robust performance for different covariate types. CoCoA-REML consistently outperforms CovReg, likely because its parametric assumption at each time point aligns with how the data are generated in our simulation, leading to more stable estimates than CovReg in this more complex scenario.

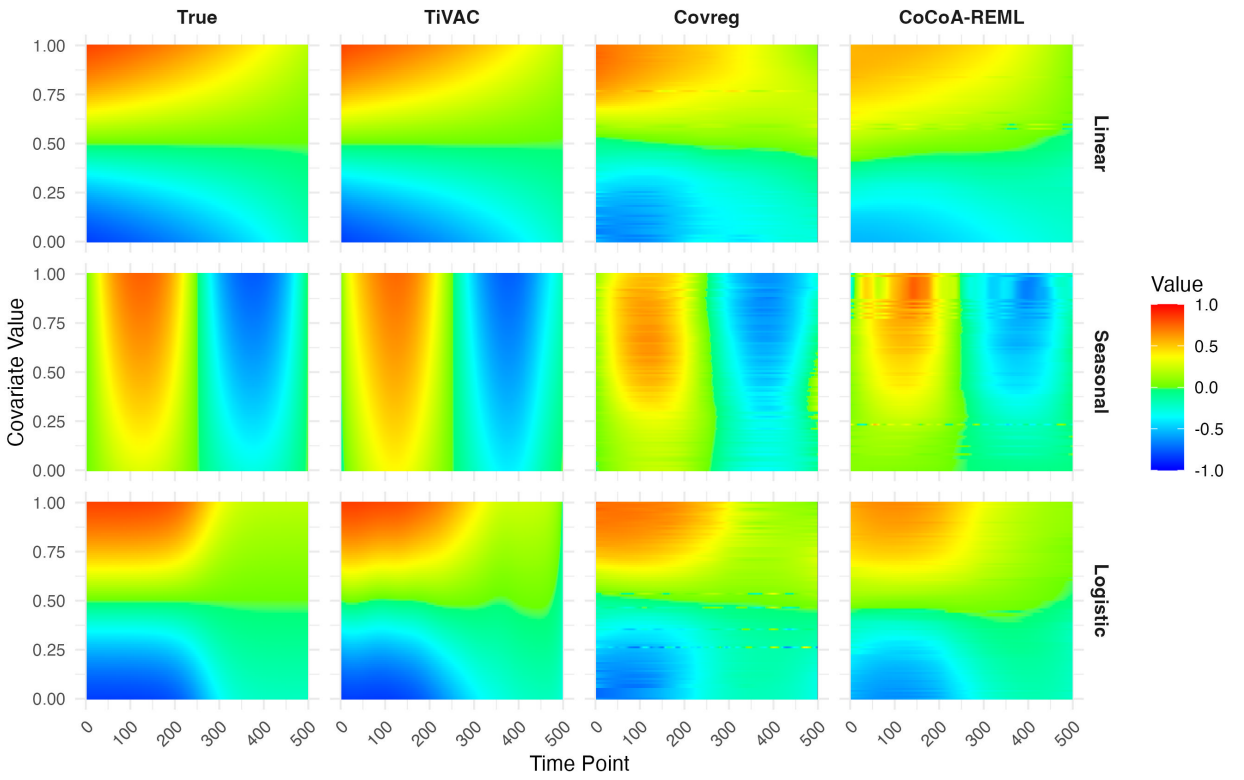


Figure 3.3: **Representative heatmap estimations under $\mathscr{T}_{Moderate}$ setting in Scenario 2.** Comparison between the estimated correlation heatmaps from three methods (TiVAC, CovReg and CoCoA-REML) and the true correlation heatmap under $\mathscr{T}_{Moderate}$ setting. In each heatmap, correlation values are color-coded, with red indicating higher correlations and blue indicating lower correlations. Each heatmap displays correlation patterns over time for different values of the continuous covariate. The figure presents three independent cases, each representing one of the three types of coefficient functions. Each case shows a representative example from the first run of the simulation (using a fixed random seed for all cases).

Figure 3.3 shows representative examples from the first run of each case (using a fixed random seed) under the $\mathscr{T}_{Moderate}$ setting in Scenario 2. The correlation heatmaps illustrate how correlation values (in rainbow colors) depend on both time points (x-axis) and covariate values (y-axis). They reveal that TiVAC's estimated heatmap is not only much closer to the true pattern but also displays smoother transitions across time and covariates. Additional comparisons of correlation heatmaps under the $\mathscr{T}_{Low}$, $\mathscr{T}_{High}$, and $\mathscr{T}_{PrBP}$ settings appear in Figures B.1, B.2, and B.3, respectively. Although CovReg and CoCoA-REML have similar RMSE under $\mathscr{T}_{Moderate}$ and $\mathscr{T}_{PrBP}$, their heatmaps indicate that these two methods recover the true patterns in $\mathscr{T}_{PrBP}$ more poorly than they do in $\mathscr{T}_{Moderate}$, illustrating how the irregular nature of the PrBP data complicates interpretability of CovReg and CoCoA-REML.

Finally, Table B.2 presents simulation results comparing CoCoA-MLE, CoCoA-REML, and CoCoA-GEE2 in Scenario 2. Unlike Scenario 1, here CoCoA-REML performs better than CoCoA-GEE2 and is comparable to CoCoA-MLE.

3.4.2 Impact of noise in correlation functions

In both Scenario 3 (binary covariate) and Scenario 4 (continuous covariate), we introduce additional randomness into the latent correlation function by letting σ_{noise} vary from 0.1 to 0.5. We use the $\mathscr{T}_{Moderate}$ setting and compute the RMSE as in their respective baseline scenarios (i.e., with no random noise in the coefficient function). As summarized in Tables B.3 (for Scenario 3) and B.5 (for Scenario 3), all methods degrade in performance as σ_{noise} increases, but TiVAC remains robust, maintaining an average RMSE below 0.1. This robustness likely stems from TiVAC's spline-based smoothing, which helps mitigate noise-induced fluctuations in the correlation estimates.

Tables B.4 (for Scenario 3) and B.6 (for Scenario 4) compare CoCoA-MLE, CoCoA-REML, and CoCoA-GEE2. In Scenario 3, CoCoA-GEE2 slightly outperforms the other CoCoA estimators for linear and logistic coefficient functions. In Scenario 4, CoCoA-MLE slightly performs better than CoCoA-REML, which in turn outperforms CoCoA-GEE2, for linear and logistic functions; whereas CoCoA-REML slightly outperforms for the seasonal functions.

3.5 Application to mood trajectories in bipolar patients

We analyzed data from patients with BP-I enrolled in PrBP (McInnis et al., 2017; Yocum et al., 2023), focusing on the PHQ-9 (Kroenke et al., 2001), a measure of depression severity, and the GAD-7 (Spitzer et al., 2006), a measure of anxiety severity, as mood outcomes. Patients' substance use disorder status and medication use were incorporated as covariates. Depression and anxiety are highly prevalent and comorbid symptoms among individuals with bipolar disorder (Kalin, 2020). However, our empirical analysis (Figure 3.1A) indicates heterogeneity in their association over time. This motivates the use of the TiVAC model to characterize the dynamic correlation between depression and anxiety and to evaluate how patient characteristics and clinical factors modulate this relationship. A better understanding of these patterns could elucidate disease mechanisms and inform personalized treatment strategies (Kim et al., 2024).

Mood outcomes were collected bimonthly, and medication use records were updated

yearly, with substantial between-patient variability in enrollment (baseline) dates and subsequent observation schedules. To ensure a consistent and interpretable timeline, we used each patient's measured age as the time variable (t_{ij}) in the TiVAC model. The dataset was preprocessed (Liu et al. 2025), followed by additional cleaning based on the following inclusion criteria: (1) for each patient, every observed time point must include both PHQ-9 and GAD-7 scores, and (2) the observed time points must correspond to ages between 30 and 65. PHQ-9 and GAD-7 scores were then centered separately for males and females and subsequently transformed using a quantile transformation (Tu et al., 2022) to approximate a Gaussian distribution. These preprocessing steps ensured that the assumptions of the TiVAC model were fairly met. As shown in Table B.7, the dataset used for modeling included 291 BP-I patients. Of these, 70.8% were female, 47.4% had a history of alcohol use disorder, and 28.9% had cannabis use disorder. Alcohol use disorder and cannabis use disorder status were assessed at enrollment and treated as time-invariant covariates. Medication-related covariates were defined based on (Liu et al. 2025) as follows: the average nervous system-related medications (which include all psychiatric medications) for each patient represents the mean count of those medications across observed time points, while the use frequency of each psychiatric medication reflects the proportion of observed time points where the medication was taken. These covariates are continuous, with this summarization approach necessitated by limitations in the medication data collection process.

We applied the TiVAC model to each covariate separately to measure its average effect on the correlation between PHQ-9 and GAD-7 scores over age in BP-I patients. Although the model can in principle handle multiple covariates, each covariate introduces a set of spline parameters. Because our data is highly irregular (e.g., few observations at certain time points), introducing too many parameters can lead to unstable and unreliable estimates. Age was recorded in months for modeling while visualized in years, with selected knots placed at intervals of 24 months. Without considering covariates, Figure B.5 shows that the correlation between PHQ-9 and GAD-7 scores slightly increases with age, consistent with the literature (Lenze, 2003) suggesting higher comorbidity between PHQ-9 and GAD-7 in older individuals.

Figure 3.4 presents the coefficient functions $\beta_1(t)$ in blue, representing the effect of each covariate on the correlation as age increases. The 95% simultaneous confidence bands (shown in gray) indicate the periods where the covariate effect is significant–specifically, when the bands do not overlap with the zero coefficient value (dashed line). For gender, males show significantly higher correlations than females as age increases, though this difference diminishes gradually. For the average nervous system-related medications, increased usage corresponds to significantly lower correlations until around age 50. Regarding antipsychotic use frequency,

greater usage is associated with lower correlations before age 50 and higher correlations after age 50, with the effects being nearly significant at the endpoints of the observed range (ages 30 and 65). For mood stabilizer use frequency, higher usage is significantly associated with lower correlations between ages 40 and 55. Although the estimated coefficient functions for these covariates appear roughly linear in time, their significance could still vary across the age range; by contrast, assuming a time-invariant effect would imply the significance remains constant. In addition, the effect of antidepressant use frequency exhibits a nonlinear pattern, as shown in its subplot, but there is insufficient evidence to conclude a significant influence over time. Similarly, there is no strong evidence of significant effects for alcohol use disorder or cannabis use disorder. Lastly, the relatively wider confidence bands at the tiles are likely due to less frequent observations in those periods.

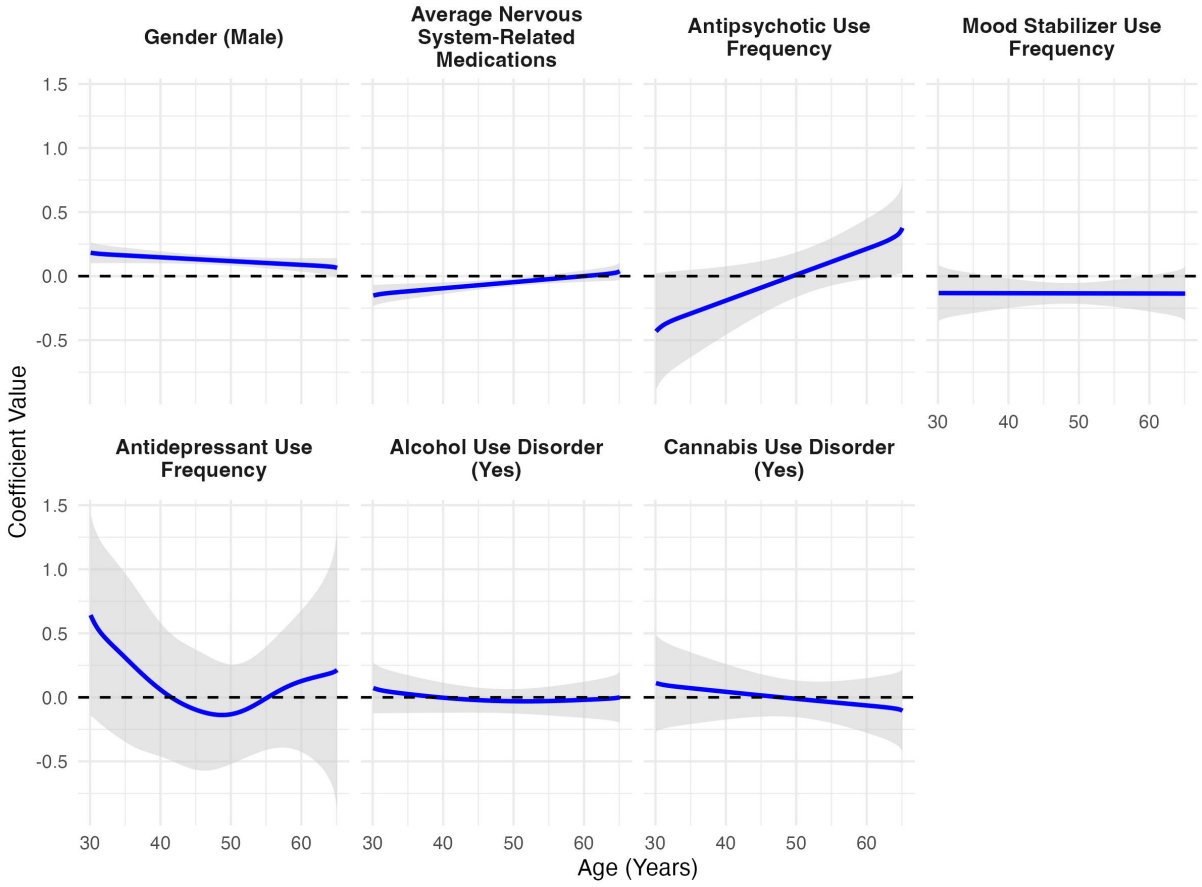


Figure 3.4: **Results of TiVAC coefficient functions on PrBP data.** TiVAC coefficient functions over age for seven covariates, measuring their effects on the correlation between PHQ-9 and GAD-7 in PrBP. For each covariate, the coefficient function $\beta(t)$ is shown in blue, with a gray band representing the 95% simultaneous confidence band. The dashed black line at zero indicates no effect.

To further interpret the TiVAC results, we visualized the TiVAC correlations for the four covariates that show significant or near-significant effects during certain time periods (Figure 3.5). Almost all TiVAC correlations across the four subplots exceed 0.5, with exceptions observed when the average nervous system-related medication usage is near 4 or the antipsychotic use frequency is near 1, particularly around ages 30–35, as indicated in blue. The consistently high TiVAC correlations highlight the well-documented comorbidity between anxiety and depressive disorders (Kalin, 2020). Genetically, these disorders share a substantial level of risk factors (Hettema, 2008). At the neural circuit level, both involve alterations in prefrontal-limbic pathways that are critical for mediating emotional regulation processes (Kalin, 2020).

In addition, the TiVAC correlation for males on average is approximately 0.1 higher than that for females at age 30 but gradually diminishes as age increases. Previous studies have shown that, in general, women with depression often exhibit greater comorbidity with anxiety disorders than men, likely due to sex differences in neural circuits and molecular mechanisms related to anxiety and depression (Bangasser and Cuarenta, 2021; Kornstein et al., 2000). However, this area remains underexplored. Interestingly, our results reveal that in the subpopulation of BP-I patients, this pattern is reversed, emphasizing the need for further research into sex-specific mechanisms underlying mood and anxiety comorbidities in BP-I.

For nervous system-related medications, increased usage is associated with lower correlations before age 40 (indicated by a shift from orange to blue) but with higher correlations between ages 60 and 65 (indicated by a shift from green to orange). A similar, though not identical, trend appears with antipsychotic use frequency, indirectly reflecting differences in drug responses across age groups. One possible explanation involves the brain's neuroplasticity, which remains relatively high at younger ages (Park and Bischof, 2013), allowing for greater adaptability. In this context, nervous system-related medications may disrupt natural emotional regulation circuits, thereby weakening correlations between anxiety and depressive symptoms.

In contrast, higher mood stabilizer use frequency is associated with slightly lower correlations (shifting from yellow to green) across all ages. Mood stabilizers are thought to modulate prefrontal-limbic circuits involved in mood regulation (Schloesser et al., 2012). This stabilizing effect may slightly decouple the neural processes underlying anxiety and depression.

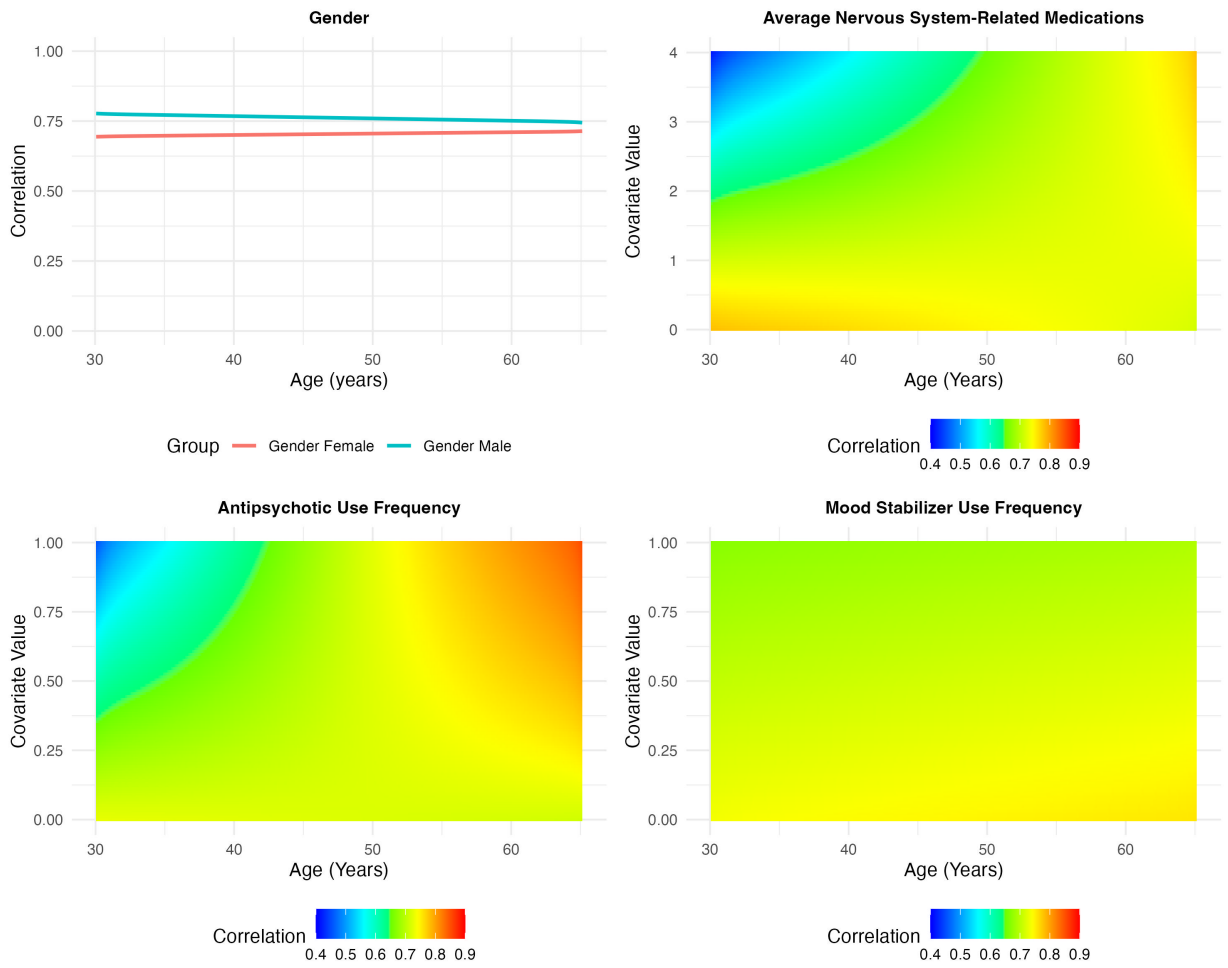


Figure 3.5: **Results of TiVAC correlations on PrBP data.** Visualization of TiVAC correlations between PHQ-9 and GAD-7 for four covariates with significant or near-significant age periods. For gender (top-left subplot), the TiVAC correlation curve over age is shown in red for females and blue for males. For the other three heatmaps, the corresponding covariates are continuous, with the covariate values on the y-axis, age on the x-axis, and heatmap colors representing the TiVAC correlation values.

Overall, by applying our TiVAC model to the mood data of BP-I patients, we revealed that the relationship between anxiety and depression depends on gender, medication use, and age, highlighting the complexity of mood dynamics in BP-I patients. While our results provide data-driven evidence, further validation is necessary to confirm these observations.

3.6 Discussion

In this work, we propose the Time-Varying and Covariate-Dependent (TiVAC) correlation model, which generalizes existing covariate-dependent correlation frameworks to account for the dynamic effects of covariates on correlation over time. Its applications span diverse fields such as biomedical research and economics. TiVAC accommodates nonlinear, time-varying covariate effects and localized periods of statistical inference by modeling latent correlation regression coefficients as smooth functions of time. In addition, our bootstrap-based procedure for estimating simultaneous confidence bands mitigates concerns of multiple testing across time points. Simulation studies demonstrated that TiVAC consistently outperforms competing methods in terms of correlation estimation accuracy across various settings. The application of TiVAC to longitudinal mood data from BP-I patients highlighted its ability to uncover meaningful clinical patterns in the relationship between depression and anxiety.

This paper primarily focuses on inference for TiVAC correlation, although its estimates could also be used to refine mean-model estimation. However, jointly modeling both mean and complex covariance structures in longitudinal data can be computationally demanding and potentially prone to convergence issues. Currently, TiVAC assumes zero autocorrelation within each outcome across different time points. Extending the bivariate Gaussian framework to a multivariate or Gaussian-process setting would allow for correlations across time points, but necessitates deeper investigation of the model-building and estimation processes due to increased complexity and computational demands. Another promising direction is to accommodate non-Gaussian outcomes, given that many real-world datasets may exhibit skewness, zero-inflation, or discrete distributions. In this setting, copula-based methods may offer a flexible approach to effectively address these distributional challenges. Finally, although the current framework focuses on two correlated variables, extending TiVAC to higher-dimensional outcomes is a promising avenue for future research; such extensions will require parameterizations that guarantee positive definiteness of the correlation matrix.

CHAPTER 4

Generalized TiVAC Correlation for Zero-Inflated Non-Gaussian Data with Application to Bipolar Mood Dynamics

4.1 Introduction

Longitudinal biomedical studies often reveal that correlations between outcomes vary across patients and evolve over time. This phenomenon motivates modeling the correlation ρ as a function of time t and patient characteristics $\mathbf{x}$—that is, $\rho(t, \mathbf{x})$—rather than assuming it is fixed at the population level. Averaging over such heterogeneity can mask or even reverse subgroup-level relationships, a phenomenon known as Simpson's paradox (Blyth, 1972).

Bipolar mood dynamics highly motivate this statistical problem. Traditionally, bipolar mood has been viewed as either (i) one-dimensional, where increases in mania naturally align with decreases in depression and vice versa, or (ii) multistable, where the illness tends to settle into distinct and persistent states (e.g., mania and depression) (Cochran et al., 2018). Statistically, both views imply a negative concurrent correlation between mania and depression. Of particular interest, a recent analysis of the Heinz C. Prechter Bipolar Research Project (PrBP) cohort found that the concurrent correlation between mania and depression can be negative or positive across individuals, with values ranging from about -0.60 to 0.75 (Figure 4.1A), revealing a more complex pattern (McInnis et al., 2017; Cochran et al., 2018). A positive concurrent correlation is consistent with mixed states or episodes in bipolar I disorder (BP-I), as formally defined in DSM-5 (Muneer, 2017). Clinically, mixed states correspond to dysphoric mania or agitated depression, in which patients feel depressed yet activated or irritable; such states often carry a higher risk of suicide (Muneer, 2017). However, few studies examine which clinical factors are associated with this heterogeneity in the mania-depression or mania-anxiety dynamic concurrent correlation. This gap motivates

our study of time-varying and covariate-dependent correlation.

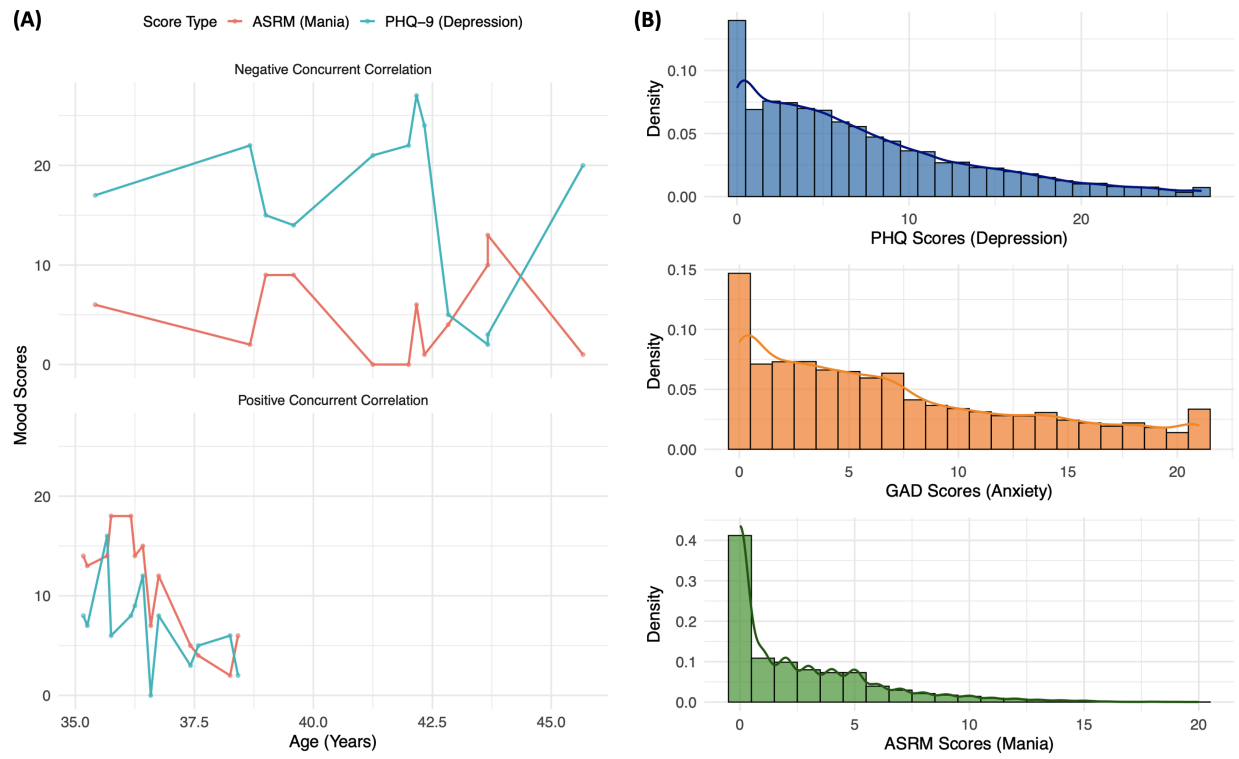


Figure 4.1: **Visualization of mood scores in BP-I. (A)** PHQ-9 (Depression) and ASRM (Mania) trajectories versus age (years) for two illustrative BP-I patients, to show the negative and positive within-person concurrent Pearson correlations. The trajectories also illustrate irregular measurement times. **(B)** Empirical distributions of PHQ-9, GAD-7 (Anxiety), and ASRM across the BP-I cohort shown as histograms with kernel density overlays; the y-axis shows the proportion of each score.

However, many biomedical outcomes, including mood scores, deviate from Gaussian assumptions. They often show zero inflation and heavy skewness. For example, in the PrBP cohort, depression scores (PHQ-9; (Kroenke et al., 2001)) and anxiety scores (GAD-7; (Spitzer et al., 2006)) contain approximately 15% zeros, while mania scores (ASRM; (Altman et al., 1997)) contain about 40% zeros and exhibit a long right tail, as shown in Figure 4.1B. Similar zero inflation arises in single-cell RNA sequencing data due to dropout (Van den Berge et al., 2018). Simple monotone transformations, such as logarithmic or empirical CDF mappings, cannot remove the excess zeros; instead, they leave a spike at the transformed value corresponding to zero. Applying Gaussian-based correlation models to such data risks biased estimates and misleading inference because the models cannot capture the discrete mass at

zero or the heavy tails. These issues motivate generalizing the TiVAC correlation model introduced in Chapter 3, which assumes bivariate normality at each time point, so that it explicitly accommodates zero inflation and skewness.

Traditional correlation models built under Gaussian assumptions do not adequately address zero inflation and skewness. For example, CoCoA (Tu et al., 2022; Ho et al., 2010; Wilding et al., 2011) was recently proposed to estimate covariate-dependent pairwise correlation, and TiVAC generalized this framework to additionally account for time variation. To handle zero-inflated data, several alternative approaches have emerged. Conditional copula methods adopt kernel-based weights to flexibly model covariate dependence (Gijbels et al., 2011, 2012). However, these methods typically lack direct inferential tools for quantifying how specific covariates influence correlation. The zero-inflated negative binomial dynamic correlation (ZENCO) model (Yang and Ho, 2021) links observed data onto latent Gaussian variables to study the underlying dependence structure. However, ZENCO relies on strict distributional assumptions, such as a negative binomial component, which may not fit the observed data in practice. Multivariate hurdle models are related but address a different question. They have been used to study conditional dependencies in zero inflated graphical models, where one outcome is modeled conditional on the others, whereas our focus is the pairwise correlation between two outcomes given covariates (McDavid et al., 2019). Importantly, all of these models except TiVAC are designed for static settings and cannot flexibly capture time-varying correlation, which is essential for understanding longitudinal outcomes.

To this end, we propose the zero–inflated Gaussian–copula–based TiVAC (Copula-TiVAC) model to learn the concurrent correlation between bivariate longitudinal outcomes and how it varies with covariates and over time. As illustrated in Figure 4.2, Copula–TiVAC assumes an underlying latent Gaussian structure that fully captures the dependence among the observed non-Gaussian, zero-inflated outcomes via a truncated Gaussian copula model (Yoon et al., 2020), providing flexibility without restrictive distributional assumptions. On the latent bivariate normal scale, we apply the Fisher z-transformation to the correlation and express the transformed correlation as a linear function of covariates with coefficients that vary smoothly over time. These time-varying coefficients capture covariate effects on correlation and are modeled with Bayesian P-splines (Lang and Brezger, 2004), which borrow strength across adjacent time points and enforce smoothness over time. Bayesian estimation proceeds via a two-stage scheme within a blocked Gibbs sampler that iteratively cycles between (i) a copula stage updating the latent Gaussian variables and segmentation thresholds given the current TiVAC correlation estimation, and (ii) a TiVAC stage updating the spline coefficients using Hamiltonian Monte Carlo. Inference on covariate effects over time is based on simultaneous confidence bands from the joint Markov Chain Monte Carlo (MCMC)

draws, avoiding the multiple testing issue of pointwise confidence intervals. Together, the benefit of Copula-TiVAC model provides a semiparametric framework well suited to zero-inflated, skewed longitudinal data while enabling inference on time- and covariate-dependent correlations.

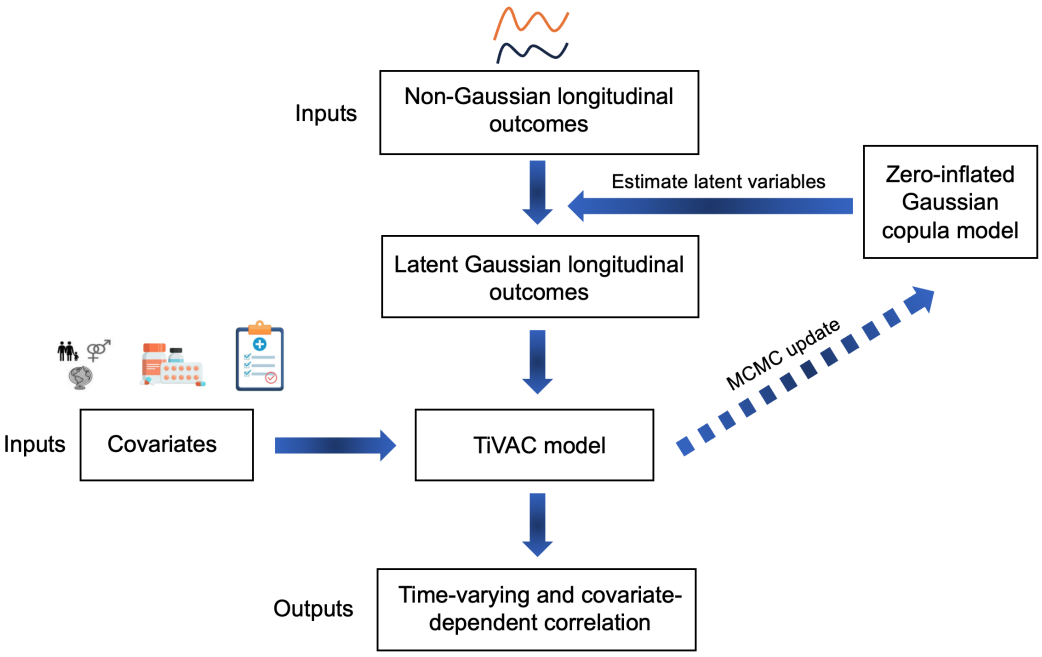


Figure 4.2: **Copula-TiVAC workflow.** Inputs consist of irregular, non-Gaussian and potentially zero-inflated longitudinal outcomes (top) together with subject-level covariates (left). The zero-inflated Gaussian copula model estimates a latent Gaussian layer that fully captures the dependence between the observed non-Gaussian outcomes. On the latent layer, TiVAC models the correlation as a function of both time and covariates. The dashed arrow indicates the blocked MCMC scheme that iterates between updating the copula model and the TiVAC model during Bayesian estimation.

Through simulation studies, we compare Copula-TiVAC against competing approaches, illustrating its advantages under various time-dependent correlation functions, different covariate setting (single binary and continuous covriate and multiple covariates), and different proportion of zeros. Finally, we apply Copula-TiVAC to longitudinal mood trajectories from around 200 BP-I participants in PrBP, focusing on pairwise correlations between mania (ASRM), depression (PHQ-9), and anxiety (GAD-7). The analysis indicates that greater exposure to psychiatric medications is associated with a decoupling of depression-mania and anxiety-mania during mid-life, cannabis use disorder tends to tighten these couplings, and gender-by-age differences also exist in depression-mania correlation.

The remainder of the paper is organized as follows. Section 4.2 presents the Copula-

TiVAC formulation. Section 4.3 details the Bayesian estimation procedure, including spline priors, the copula-based mapping, posterior inference, and simultaneous confidence band construction. Section 4.4 reports simulation results evaluating accuracy and robustness across zero-inflation levels, covariate structures, and sampling schemes. Section 4.5 applies Copula-TiVAC to the PrBP cohort and interprets the clinical implications. We conclude with a discussion of our findings and future research directions.

4.2 Copula-TiVAC model

4.2.1 Data structure and notation

In this section, we first outline the data structure and notation for irregularly observed, bivariate longitudinal outcomes. Let n denote the number of individuals. For individual $i \in \{1, \ldots, n\}$, measurements are taken at irregular times $\mathcal{T}_i = \{t_{i1}, \ldots, t_{im_i}\}$, where m_i is the total number of observed time points for individual i and t_{ij} denotes the jth time point ($j = 1, \ldots, m_i$). At time t_{ij} the concurrent bivariate outcome is $\mathbf{Y}_i(t_{ij}) = \{Y_{i1}(t_{ij}),\, Y_{i2}(t_{ij})\}^\top$. Let $\mathbf{X}_i \in \mathbb{R}^p$ denote the subject-level covariate vector with p covariates. Throughout, $k \in \{1, 2\}$ indexes the two outcome variables, j indexes times within a subject, i indexes subjects, and $r \in \{1, \ldots, p\}$ indexes covariates. We use $Y_{\cdot k}(\cdot)$ to denote a generic variable evaluated at an arbitrary subject and time; for example, $Y_{\cdot 1}(\cdot)$ is the first outcome for any individual at any time, and likewise $Y_{\cdot 2}(\cdot)$ for the second outcome. We denote the standard normal cumulative distribution function (cdf) by Φ and write $N_2(\mathbf{0}, \mathbf{\Sigma})$ for the bivariate normal with mean $\mathbf{0}$ and covariance $\mathbf{\Sigma}$. Bold symbols indicate vectors or matrices.

4.2.2 Preliminaries: copula framework

We develop the model in three sequential steps. First, we review the standard Gaussian copula model in Definition 1 (Liu et al., 2009), which assumes a latent Gaussian representation underlying the observed non-Gaussian outcomes. Second, we incorporate inflated zeros via a zero-inflated (or truncated) Gaussian copula, as illustrated in Definition 2 (Yoon et al., 2020). Finally, we allow the latent correlation to depend on both time and covariates, yielding the Copula–TiVAC model.

Definition 1 (Gaussian copula model). *A random bivariate vector* $\mathbf{Y}_\cdot(\cdot) = \{Y_{\cdot 1}(\cdot),\, Y_{\cdot 2}(\cdot)\}^\top$ *satisfies a Gaussian copula model, also known as the nonparanormal (NPN) model, if there exist strictly increasing functions* $f_1, f_2 : \mathbb{R} \to \mathbb{R}$ *such that the transformed vector*

$\mathbf{Z}_{\cdot}(\cdot) := [f_1\{Y_{\cdot 1}(\cdot)\},\, f_2\{Y_{\cdot 2}(\cdot)\}]^\top$ *is bivariate normal with mean zero and correlation matrix* $\mathbf{\Omega} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. *We denote this by* $\{Y_{\cdot 1}(\cdot), Y_{\cdot 2}(\cdot)\} \sim \mathrm{NPN}(\mathbf{0}, \mathbf{\Omega}, \mathbf{f})$ *with* $\mathbf{f} = (f_1, f_2)$.

Definition 1 formalizes the idea that continuous outcomes can be represented on a latent Gaussian scale through unknown, order-preserving transformations f_k, thereby separating marginal distributions from the correlation parameter ρ. As shown by the conceptual diagram (Figure. 4.3) adapted from Choi et al. (2024), Definition 1 corresponds to the mapping between the upper layer of latent Gaussian variables and middle layer of latent continuous variables. However, since Definition 1 presumes continuous variables, the continuity assumption fails when the observed data contain a point mass at zero. The monotone transformations $\{f_k\}$ cannot address zero inflation: all zeros collapse to a single latent value, creating a spike on the transformed scale and violating Gaussianity. This motivates an extension using zero-inflated Gaussian copula, as illustrated by Definition 2.

Definition 2 (Zero-inflated Gaussian copula model). *A random bivariate vector* $\mathbf{Y}_{\cdot}(\cdot) = \{Y_{\cdot 1}(\cdot),\, Y_{\cdot 2}(\cdot)\}^\top$ *satisfies a zero-inflated Gaussian copula model if there exists a latent continuous vector* $\mathbf{Y}^*_{\cdot}(\cdot) = \{Y^*_{\cdot 1}(\cdot),\, Y^*_{\cdot 2}(\cdot)\}^\top$ *that follows* $\mathbf{Y}^*_{\cdot}(\cdot) \sim \mathrm{NPN}(\mathbf{0}, \mathbf{\Omega}, \mathbf{f})$ *with* $\mathbf{f} = (f_1, f_2)$, *and if there exist fixed thresholds* $d_1, d_2 \in \mathbb{R}$ *such that, for* $k \in \{1, 2\}$,

$$Y_{\cdot k}(\cdot) = \mathbf{1}\{Y^*_{\cdot k}(\cdot) \le d_k\}\ 0\ +\ \mathbf{1}\{Y^*_{\cdot k}(\cdot) > d_k\}\ Y^*_{\cdot k}(\cdot). \tag{4.1}$$

As illustrated in Figure 4.3 and formalized in Definition 2, the observed variables $Y_{\cdot k}(\cdot)$, $k \in \{1, 2\}$, are generated from a latent bivariate Gaussian vector $\mathbf{Z}_{\cdot}(\cdot)$ through a two–stage mechanism. First, the latent continuous variable $Y^*_{\cdot k}(\cdot)$ is obtained from latent Gaussian by inverting the strictly increasing transformations f_k,

$$Y^*_{\cdot k}(\cdot) = f_k^{-1}\big\{Z_{\cdot k}(\cdot)\big\}, \qquad k \in \{1, 2\},$$

which maps the latent Gaussian layer to the middle latent continuous layer in Figure 4.3. Second, the observed variable $Y_{\cdot k}(\cdot)$ is derived by truncating the latent continuous variable $Y^*_{\cdot k}(\cdot)$ at the fixed thresholds d_k as in Equation 4.1. If $Y^*_{\cdot k}(\cdot) \le d_k$ the observed value is 0, whereas if $Y^*_{\cdot k}(\cdot) > d_k$ the observed value equals $Y^*_{\cdot k}(\cdot)$. Because each f_k is strictly increasing, the continuous–layer cutoffs d_k could map to Gaussian–layer cutoffs, such that $\eta_k := f_k(d_k)$. Hence, $Y_{\cdot k}(\cdot) = 0$ precisely when $Z_{\cdot k}(\cdot) \le \eta_k$, and $Y_{\cdot k}(\cdot) = Y^*_{\cdot k}(\cdot)$ when

$Z_{\cdot k}(\cdot) > \eta_k$. Equivalently,

$$Y_{\cdot k}(\cdot) = \mathbf{1}\{Z_{\cdot k}(\cdot) \leq \eta_k\}\ 0\ +\ \mathbf{1}\{Z_{\cdot k}(\cdot) > \eta_k\}\ Y^*_{\cdot k}(\cdot), \qquad k \in \{1, 2\}.$$

Note that the zero-inflated Gaussian copula model in Definition 2 treats the latent correlation as a constant ρ in $\mathbf{\Sigma}$ and thus does not model how the association varies with time or covariates.

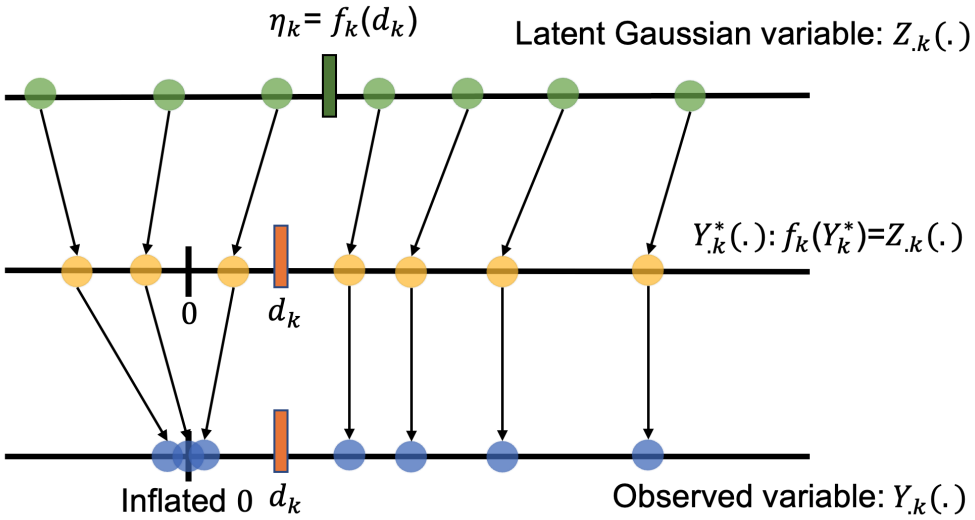


Figure 4.3: **Conceptual diagram of the zero-inflated Gaussian copula model.** Three layers illustrate the construction used in Definitions 1–2. *Top layer:* latent Gaussian variable $Z_{\cdot k}(\cdot)$ with correlation ρ between $Z_{\cdot 1}(\cdot)$ and $Z_{\cdot 2}(\cdot)$ (later generalized to $\rho(t_{ij}, \mathbf{X}_i)$ in Copula–TiVAC). *Middle layer:* latent continuous variable $Y^*_{\cdot k}(\cdot)$ linked to the Gaussian layer by the strictly increasing functions f_k, i.e., $f_k\{Y^*_{\cdot k}(\cdot)\} = Z_{\cdot k}(\cdot)$. *Bottom layer:* observed variable $Y_{\cdot k}(\cdot)$ obtained by truncating $Y^*_{\cdot k}(\cdot)$ at the fixed thresholds d_k, such that $Y_{\cdot k}(\cdot) = 0$ if $Y^*_{\cdot k}(\cdot) \leq d_k$ and $Y_{\cdot k}(\cdot) = Y^*_{\cdot k}(\cdot)$ otherwise. Thus zeros on the observed scale correspond to $Z_{\cdot k}(\cdot) \leq \eta_k = f_k(d_k)$, whereas positive values correspond to $Z_{\cdot k}(\cdot) > \eta_k$.

4.2.3 Copula–TiVAC model

Building on the zero-inflated Gaussian copula in Definition 2, we retain the same mechanism of marginal distributions for each individual i and time t_{ij}. The observed variable $Y_{ik}(t_{ij})$ is obtained from the latent continuous variable $Y^*_{ik}(t_{ij})$ by truncation at fixed thresholds d_k as in Equation 4.1, and the strictly increasing functions f_k link the latent Gaussian with $Y^*_{ik}(t_{ij})$ such that $Z_{ik}(t_{ij}) = f_k\{Y^*_{ik}(t_{ij})\}$.

Copula–TiVAC model generalizes Definition 2 by allowing the correlation matrix Ω on

the latent Gaussian scale to depend on both time t_{ij} and covariates $\mathbf{X}_i$. Specifically,

$$\mathbf{Z}_i(t_{ij}) = \{Z_{i1}(t_{ij}),\, Z_{i2}(t_{ij})\}^\top \sim N_2\Big\{\, \mathbf{0},\, \boldsymbol{\Sigma}(t_{ij}, \mathbf{X}_i)\, \Big\}, \quad \boldsymbol{\Sigma}(t_{ij}, \mathbf{X}_i) = \begin{bmatrix} 1 & \rho(t_{ij}, \mathbf{X}_i) \\ \rho(t_{ij}, \mathbf{X}_i) & 1 \end{bmatrix}. \tag{4.2}$$

Here $\rho(t_{ij}, \mathbf{X}_i) \in (-1, 1)$ denotes the concurrent latent correlation at time t_{ij} given covariates $\mathbf{X}_i$. To work on the real line, we apply the Fisher transformation such that

$$\lambda(t_{ij}, \mathbf{X}_i) = \log\Big\{ \frac{1 + \rho(t_{ij}, \mathbf{X}_i)}{1 - \rho(t_{ij}, \mathbf{X}_i)} \Big\}. \tag{4.3}$$

The transformed correlation is then modeled by a varying-coefficient linear form:

$$\lambda(t_{ij}, \mathbf{X}_i) = \mathbf{X}_i^\top \boldsymbol{\beta}(t_{ij}), \tag{4.4}$$

where $\boldsymbol{\beta}(t) = \{\beta_1(t), \ldots, \beta_p(t)\}^\top$ captures the time-varying effects of the p covariates on the latent correlation.

Model assumptions. Building on the specification above, we impose the following assumptions to ensure identifiability and interpretability of Copula-TiVAC model.

- *Inflation at 0 only.* Each variable's observations may exhibit a point mass at 0, whereas positive values are treated as continuous and carry no additional inflation at any positive level. This is consistent with the empirical distribution of the mood scores (Figure. 4.1), while it can be generalized to multiple inflation points (discussed in Section 4.6).

- *Monotone functions to a latent Gaussian scale.* There exist strictly increasing functions f_k such that the latent Gaussian variables map to the latent continuous variables via $Y^*_{ik}(t_{ij}) = f_k^{-1}\{Z_{ik}(t_{ij})\}, k \in \{1, 2\}$. On this Z-scale, the pair follows the bivariate normal model with unit variances, as shown in Equation 4.2.

- *Time- and covariate-dependent latent correlation.* The concurrent correlation on the Gaussian scale is allowed to vary with time and covariates and is parameterized through the Fisher transform Equation 4.3 and the varying-coefficient representation Equation 4.4. This captures heterogeneous associations between the two outcomes across individuals and over time, while keeping the parameterization on an unconstrained scale.

- *No serial dependence.* There is no autocorrelation across distinct observation times.

Accordingly, we focus inference on the concurrent dependence at each time t_{ij}, thereby balancing model complexity and computational efficiency.

4.3 Estimation

4.3.1 Bayesian TiVAC: spline representation and priors

We approximate each time–varying coefficient function $\beta_r(t)$ with a q-dimensional B–spline basis $\mathbf{b}(t) = \big\{b_1(t), \ldots, b_q(t)\big\}^\top$ and spline coefficient vector $\boldsymbol{\theta}_r = \big(\theta_{r1}, \ldots, \theta_{rq}\big)^\top$, i.e., $\beta_r(t) \approx \mathbf{b}(t)^\top \boldsymbol{\theta}_r$ for $r = 1, \ldots, p$. For individual i at time t_{ij}, the covariates and spline basis are combined into a new design vector $\mathbf{A}_{i,j} := \mathbf{X}_i^\top \otimes \mathbf{b}(t_{ij})$, where $\otimes$ denotes the Kronecker product. The parameters to estimate are the stacked spline coefficients $\boldsymbol{\theta} = \big(\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_p^\top\big)^\top$. Then the Fisher–transformed correlation can be equivalently written in linear form

$$\lambda\big(t_{ij}, \mathbf{X}_i\big) = \mathbf{X}_i^\top \, \boldsymbol{\beta}\big(t_{ij}\big) \;\approx\; \mathbf{A}_{i,j}\, \boldsymbol{\theta}.$$

To induce smoothness across adjacent coefficients in $\boldsymbol{\theta}_r$, and hence in $\beta_r(t)$ and $\rho(t_{ij}, \mathbf{X}_i)$, we place a Bayesian P–spline prior (Lang and Brezger, 2004) on $\boldsymbol{\theta}_r$ for each r:

$$p(\boldsymbol{\theta}_r \mid \tau_r^2) \;\propto\; \exp\Big(-\tfrac{1}{2\,\tau_r^2}\, \boldsymbol{\theta}_r^\top \mathbf{K}\, \boldsymbol{\theta}_r\Big),$$

where the first-order penalty matrix $\mathbf{K}$ can be written compactly as $\mathbf{K} = \mathrm{diag}(1, 2, \ldots, 2, 1) + \mathrm{diag}(-1, \ldots, -1)_{+1} + \mathrm{diag}(-1, \ldots, -1)_{-1}$, with $\mathrm{diag}(\cdot)_{\pm 1}$ placing entries on the the first super- and sub-diagonals.

Independent inverse-Gamma hyperpriors are placed on the smoothing penalties,

$$\tau_r^2 \sim \mathrm{IG}(a_r, b_r),$$

with weakly informative defaults $a_r = 1$ and $b_r = 0.005$, as recommended by Lang and Brezger (2004). Hence, the P–spline combines a B–spline basis representation with a smoothness penalty in its prior, allowing flexible estimation of the time-varying coefficient function $\beta_r(t)$ while promoting data-driven smoothness.

4.3.2 Copula transformation to the latent Gaussian scale

For each outcome $k \in \{1, 2\}$, we define the index sets of positive and zero observations

$$\mathcal{I}_k^+ = \{(i, j) : Y_{ik}(t_{ij}) > 0\}, \qquad \mathcal{I}_k^0 = \{(i, j) : Y_{ik}(t_{ij}) = 0\},$$

and let $N_k = |\mathcal{I}_k^+| + |\mathcal{I}_k^0|$ be the total number of observations on the kth outcome. We denote $\mathbf{Z}^+ = \{\, Z_{ik}^+(t_{ij}) : (i,j) \in \mathcal{I}_k^+,\ k = 1,2 \,\}$ as the latent Gaussian values corresponding to positive observations, and $\mathbf{Z}^0 = \{\, Z_{ik}^0(t_{ij}) : (i,j) \in \mathcal{I}_k^0,\ k = 1,2 \,\}$ as the latent Gaussian values corresponding to observed zeros.

Estimation of $\mathbf{Z}^+$. We first estimate the unknown monotone marginal transformation f_k that links the positive observations with the latent Gaussian scale via $Z_{ik}^+(t_{ij}) = f_k\big\{Y_{ik}(t_{ij})\big\}$ for $(i,j) \in \mathcal{I}_k^+$. By Definition 2, let F_k be the marginal cdf of the latent continuous variable $Y_{\cdot k}^*$; then

$$F_k(a) = P\big(Y_{\cdot k}^* \le a\big) = P\big\{Z_{\cdot k} \le f_k(a)\big\} = \Phi\big\{f_k(a)\big\} \quad \Longrightarrow \quad f_k(a) = \Phi^{-1}\{F_k(a)\}.$$

Since F_k is unknown, we estimate it via the empirical cdf based on all N_k observations (Klaassen and Wellner, 1997),

$$\widehat{F}_k(x) = \frac{N_k}{N_k + 1} \sum_{(i,j) \in \mathcal{I}_k^+ \cup \mathcal{I}_k^0} \frac{1}{N_k}\, \mathbf{1}\big\{Y_{ik}(t_{ij}) \le x\big\}.$$

For every $(i,j) \in \mathcal{I}_k^+$ and given $\widehat{F}_k$, we compute the latent Gaussian values $\widehat{Z}_{ik}^+(t_{ij}) = \widehat{f}_k\big\{Y_{ik}(t_{ij})\big\} = \Phi^{-1}\Big[\widehat{F}_k\big\{Y_{ik}(t_{ij})\big\}\Big]$, which are treated as observed pseudodata (Choi et al., 2024).

4.3.3 Posterior inference

Given the observed Gaussian pseudodata $\widehat{\mathbf{Z}}^+$, we use a blocked Gibbs sampler that cycles between the copula layer and the TiVAC layer, as summarized in Algorithm 3. Inflated zeros are tied to the left tail of the latent Gaussian scale via unknown thresholds $\boldsymbol{\eta} = (\eta_1, \eta_2)^\top$. For $\boldsymbol{\eta}$, we use independent improper uniform priors such that $p(\eta_k) \propto 1$, $k \in \{1,2\}$. Let $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_p)^\top$ denote the vector of smoothing parameters. In particular, at each iteration we update

$$\text{(i) } \mathbf{Z}^0 \mid \boldsymbol{\eta}, \boldsymbol{\theta}, \widehat{\mathbf{Z}}^+; \qquad \text{(ii) } \boldsymbol{\eta} \mid \mathbf{Z}^0, \widehat{\mathbf{Z}}^+; \qquad \text{(iii) } \boldsymbol{\theta} \mid \mathbf{Z}^0, \widehat{\mathbf{Z}}^+, \boldsymbol{\tau}; \qquad \text{(iv) } \boldsymbol{\tau} \mid \boldsymbol{\theta}.$$

Updating $\mathbf{Z}^0$. For each i and j, let $\boldsymbol{\Sigma}_{ij}$ be the 2×2 covariance matrix with unit variances and correlation $\rho\big(t_{ij}, \mathbf{X}_i\big)$. When $Y_{i1}(t_{ij}) = Y_{i2}(t_{ij}) = 0$, draw

$$\mathbf{Z}_i^0(t_{ij}) \sim \mathrm{N}_2(\mathbf{0}, \boldsymbol{\Sigma}_{ij}) \ \text{ truncated to } \ (-\infty, \eta_1] \times (-\infty, \eta_2].$$

When $Y_{i1}(t_{ij}) = 0$ but $Y_{i2}(t_{ij}) > 0$, sample $Z^0_{i1}(t_{ij})$ conditionally on $\widehat{Z}^+_{i2}(t_{ij})$:

$$Z^0_{i1}(t_{ij}) \sim \mathrm{TN}\Big\{-\infty,\, \eta_1;\ \mu_{1|2},\, 1-\rho^2\big(t_{ij}, \mathbf{X}_i\big)\Big\}, \quad \mu_{1|2} = \rho\big(t_{ij}, \mathbf{X}_i\big)\, \widehat{Z}^+_{i2}(t_{ij}),$$

and analogously when $Y_{i2}(t_{ij}) = 0 < Y_{i1}(t_{ij})$. Here $\mathrm{TN}(a, b; \mu, \sigma^2)$ denotes a normal $N(\mu, \sigma^2)$ truncated to $[a, b]$.

Updating $\boldsymbol{\eta}$. Given the latent Gaussian draws, the thresholds are sampled from a uniform distribution over the interval separating the latent values corresponding to inflated zeros from those corresponding to observed positives:

$$\eta_k \sim \mathrm{Unif}\Big[\max\{Z_{ik}(t_{ij}) : (i,j) \in \mathcal{I}^0_k\},\ \min\{\widehat{Z}_{ik}(t_{ij}) : (i,j) \in \mathcal{I}^+_k\}\Big], \quad k = 1, 2.$$

Updating $\boldsymbol{\theta}$. Define the working latent Gaussian pair at each (i, j) by

$$\tilde{\mathbf{Z}}_i(t_{ij}) = \big\{\tilde{Z}_{i1}(t_{ij}),\, \tilde{Z}_{i2}(t_{ij})\big\}^\top, \qquad \tilde{Z}_{ik}(t_{ij}) = \begin{cases} \widehat{Z}^+_{ik}(t_{ij}), & (i,j) \in \mathcal{I}^+_k, \\ Z^0_{ik}(t_{ij}), & (i,j) \in \mathcal{I}^0_k. \end{cases}$$

Then, based on Equation 4.2, TiVAC log-likelihood over all subjects and times can be expressed as

$$\ell(\boldsymbol{\theta}; \tilde{\mathbf{Z}}) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \log \phi_2\Big\{\tilde{\mathbf{Z}}_i(t_{ij});\, \mathbf{0},\ \boldsymbol{\Sigma}_{ij}\big(\rho_{ij}\big)\Big\},$$

where $\phi_2(\cdot; \mathbf{0}, \boldsymbol{\Sigma})$ is the $N_2(\mathbf{0}, \boldsymbol{\Sigma})$ density and $\rho_{ij} = \rho\big(t_{ij}, \mathbf{X}_i; \boldsymbol{\theta}\big)$. With the P–spline penalty from Section 4.3.1, the negative log-posterior is

$$U(\boldsymbol{\theta}) \ = \ -\ell(\boldsymbol{\theta}; \tilde{\mathbf{Z}}) \ + \ \frac{1}{2} \sum_{r=1}^{p} \frac{\boldsymbol{\theta}_r^\top \mathbf{K}\, \boldsymbol{\theta}_r}{\tau_r^2}.$$

Based on $U(\boldsymbol{\theta})$ and its gradient, we update spline parameters $\boldsymbol{\theta}$ by using Hamiltonian Monte Carlo (HMC) (Betancourt, 2017).

Updating $\boldsymbol{\tau}$. With prior $\tau_r^2 \sim \mathrm{IG}(a_r, b_r)$, the full conditional is conjugate:

$$\tau_r^2 \ \big|\ \boldsymbol{\theta}_r \ \sim \ \mathrm{IG}\Big(a_r + \tfrac{q-1}{2},\ b_r + \tfrac{1}{2}\, \boldsymbol{\theta}_r^\top \mathbf{K}\, \boldsymbol{\theta}_r\Big).$$

Algorithm 3 Blocked Gibbs sampler for Copula–TiVAC

Input: fixed $\widehat{\mathbf{Z}}^{+}$, design vectors $\{\mathbf{A}_{i,j}\}$, hyperparameters a_r and b_r.

Initialize: $\boldsymbol{\theta}^{(0)},\ \boldsymbol{\eta}^{(0)},\ \boldsymbol{\tau}^{(0)},\ \mathbf{Z}^{0\,(0)}$.

for $g = 1, \ldots, G$ **do**

Update $\mathbf{Z}^{0}$:

- If $Y_{i1}(t_{ij}) = Y_{i2}(t_{ij}) = 0$: $\mathbf{Z}_i^{0\,(g)}(t_{ij}) \sim N_2(\mathbf{0}, \boldsymbol{\Sigma}_{ij})$ truncated to $(-\infty, \eta_1^{(g-1)}] \times (-\infty, \eta_2^{(g-1)}]$, for $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$.
- If $Y_{i1}(t_{ij}) = 0 < Y_{i2}(t_{ij})$: $Z_{i1}^{0\,(g)}(t_{ij}) \sim \mathrm{TN}\big\{-\infty, \eta_1^{(g-1)};\ \rho_{ij}^{(g-1)}\, \widehat{Z}_{i2}^{+}(t_{ij}),\ 1 - \rho_{ij}^{2,(g-1)}\big\}$, and symmetrically for updating $Z_{i2}^{0\,(g)}(t_{ij})$, for $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$.

Update $\boldsymbol{\eta}$: $\eta_k^{(g)} \sim \mathrm{Unif}\Big[\ \max\{Z_{ik}^{0\,(g)}(t_{ij})\},\ \min\{\widehat{Z}_{ik}^{+}(t_{ij})\}\ \Big],\ k = 1, 2.$

Update $\boldsymbol{\theta}$: perform one HMC update based on negative log posterior:

$$U(\boldsymbol{\theta}) = -\sum_{i=1}^{n}\sum_{j=1}^{m_i} \log \phi_2\big\{\tilde{\mathbf{Z}}_i^{(g)}(t_{ij}); \mathbf{0}, \boldsymbol{\Sigma}_{ij}(\rho_{ij})\big\} + \tfrac{1}{2}\sum_{r=1}^{p} \frac{\boldsymbol{\theta}_r^{\top}\mathbf{K}\,\boldsymbol{\theta}_r}{\tau_r^{2,(g-1)}}.$$

Update $\boldsymbol{\tau}$: $\tau_r^{2,(g)} \sim \mathrm{IG}\Big\{a_r + \tfrac{q-1}{2},\ b_r + \tfrac{1}{2}\,\boldsymbol{\theta}_r^{(g)\top}\mathbf{K}\,\boldsymbol{\theta}_r^{(g)}\Big\},\ r = 1, \ldots, p.$

end for

Output: posterior draws $\{\boldsymbol{\theta}^{(g)}, \boldsymbol{\tau}^{(g)}, \boldsymbol{\eta}^{(g)}, \mathbf{Z}^{0\,(g)}\}_{g=1}^{G}$.

Simultaneous confidence band. The time-varying covariate effect is represented by $\beta_r(t)$ for rth covariate. Its statistical inference can be acquired by a simultaneous $(1-\alpha)$ confidence band, which covers the entire trajectory with probability at least $1 - \alpha$:

$$\Pr\Big\{\ L_r(t) \le \beta_r(t) \le U_r(t) \text{ for all } t \in \mathcal{T}\ \Big\}\ \ge\ 1 - \alpha.$$

To construct the band based on Meyer et al. (2015), we use the posterior draws of the spline coefficients $\{\boldsymbol{\theta}_r^{(g)}\}_{g=1}^{G}$ (after burn-in/thinning) and obtain $\beta_r^{(g)}(t) = \mathbf{b}(t)^{\top}\boldsymbol{\theta}_r^{(g)}, t \in \mathcal{T}$, where $\mathcal{T}$ denotes a finite grid of time points. From these draws we compute the pointwise posterior mean $\widehat{\beta}_r(t)$ and standard deviation $\widehat{s}_r(t)$. We then form the studentized supremum statistic as follow

$$T_r^{(g)} = \max_{t\in\mathcal{T}} \left| \frac{\beta_r^{(g)}(t) - \widehat{\beta}_r(t)}{\widehat{s}_r(t)} \right|,$$

and take $c_{r,1-\alpha}$ to be the empirical $(1-\alpha)$-quantile of $\{T_r^{(g)}\}_{g=1}^{G}$. Finally, the simultaneous confidence band is expressed as

$$L_r(t) = \widehat{\beta}_r(t) - c_{r,1-\alpha}\,\widehat{s}_r(t), \qquad U_r(t) = \widehat{\beta}_r(t) + c_{r,1-\alpha}\,\widehat{s}_r(t), \qquad t \in \mathcal{T}.$$

This MCMC-based band construction is the Bayesian analogue of the simultaneous band in Ruppert et al. (2010). It addresses the issue of multiple testing in pointwise intervals and requires no additional resampling such as the bootstrap.

4.4 Simulation study

We evaluate the estimation accuracy of the proposed copula–TiVAC model for time- and covariate-varying correlation in zero-inflated, right-skewed data. Scenarios vary by covariate type (binary, continuous, or multiple). Within each scenario we vary the proportion of zeros, the temporal pattern of the time-varying effects on correlation, and the sampling density of the longitudinal measurements.

We compare Copula–TiVAC to three alternative methods. The first is the TiVAC method, used as a baseline to evaluate the potential improvement of Copula–TiVAC in addressing zero-inflated data. The second is the "Empirical" correlation trajectory, computed as the Pearson correlation between the two outcomes at each observed time point within each covariate group, and then smoothed over time using LOESS regression, with candidate span values ranging from 0.1 to 1 in increments of 0.1. The optimal span is selected via 5-fold cross-validation. The third method is CoCoA–GEE (Ho et al., 2010), which estimates covariate-dependent correlation within a framework of second-order generalized estimating equations (GEE2), offering robustness to non-Gaussian outcomes. The resulting estimates are also smoothed using the same LOESS approach as in the Empirical method. For fair comparison, we apply the empirical cdf transformation (Section 4.3.2) as a preprocessing step to all competing methods.

For the simulation settings, we fix the sample size at $n = 300$ and define the time grid $\mathcal{T} = \{1, \ldots, 500\}$. The zero proportion for each outcome is controlled by $\pi \in \{0\%, 10\%, \ldots, 50\%\}$. For individual i, observation times $\{t_{ij}\}_{j=1}^{m_i}$ follow one of two designs: a design mimicking the PrBP data, $\mathscr{T}_{Real}$, with m_i randomly drawn from $\{3, \ldots, 40\}$; and a sparse design, $\mathscr{T}_{Sparse}$, with $m_i = 10$. For each individual, times are sampled without replacement from $\mathcal{T}$.

For subject i at time t_{ij}, the latent bivariate Gaussian vector $\mathbf{Z}_i(t_{ij})$ is generated according to Equations 4.2–4.4. In the single-covariate-scenario (binary or continuous X_i), we define $\lambda_{ij} = \beta_0(t_{ij}) + \beta_1(t_{ij})\, X_i$, where $X_i \in \{0, 1\}$ for the binary case and $X_i \in [0, 1]$ for the continuous case. In the multiple-covariate scenario with a continuous covariate $X_{i1} \in [0, 1]$ and a binary covariate $X_{i2} \in \{0, 1\}$, we set $\lambda_{ij} = \beta_0(t_{ij}) + \beta_1(t_{ij})\, X_{i1} + \beta_2(t_{ij})\, X_{i2} + \beta_3(t_{ij})\, X_{i1} X_{i2}$. For the single-covariate scenario, we consider trajectories in which $\beta_0(t)$ and $\beta_1(t)$ each follow linear, seasonal, or logistic patterns. For the multiple-covariate scenario, $\beta_0(t)$ is constant at 0, $\beta_1(t)$ is linear, $\beta_2(t)$ is seasonal, and $\beta_3(t)$ is logistic. Overall, we examined a total of 84

settings. To obtain zero-inflated and right-skewed observations, we map the latent variables to the observed scale via a truncated-exponential transformation,

$$Y_{ik}(t_{ij}) = \mathbf{1}\big\{ Z_{ik}(t_{ij}) \le \delta_k \big\}\, 0 + \mathbf{1}\big\{ Z_{ik}(t_{ij}) > \delta_k \big\}\, \exp\big\{ Z_{ik}(t_{ij}) \big\}, \quad \delta_k = \Phi^{-1}(\pi),$$

using the same π on both outcomes.

For model evaluation, we run 50 replications for each scenario. We use the root-mean-squared error (RMSE) to compare the estimation accuracy of the different methods in calculating the true correlation function $\rho(t, x)$,

$$\text{RMSE} = \left[\frac{1}{|\mathcal{T}|\,|\mathcal{X}|} \sum_{t \in \mathcal{T}} \sum_{x \in \mathcal{X}} \big\{ \widehat{\rho}(t, x) - \rho(t, x) \big\}^2 \right]^{1/2},$$

where $\mathcal{T}$ is the time grid and $\mathcal{X}$ is the covariate grid used for averaging. In the binary-covariate scenario, we average RMSE over time within each group $x \in \{0, 1\}$ and then across groups. In the continuous-covariate scenario, we average over time and a grid $x \in \{0, 0.25, 0.5, 0.75, 1\}$. In the multiple-covariate scenario, we average over time and the Cartesian grid $\{0, 0.25, 0.5, 0.75, 1\} \times \{0, 1\}$ for (X_{i1}, X_{i2}).

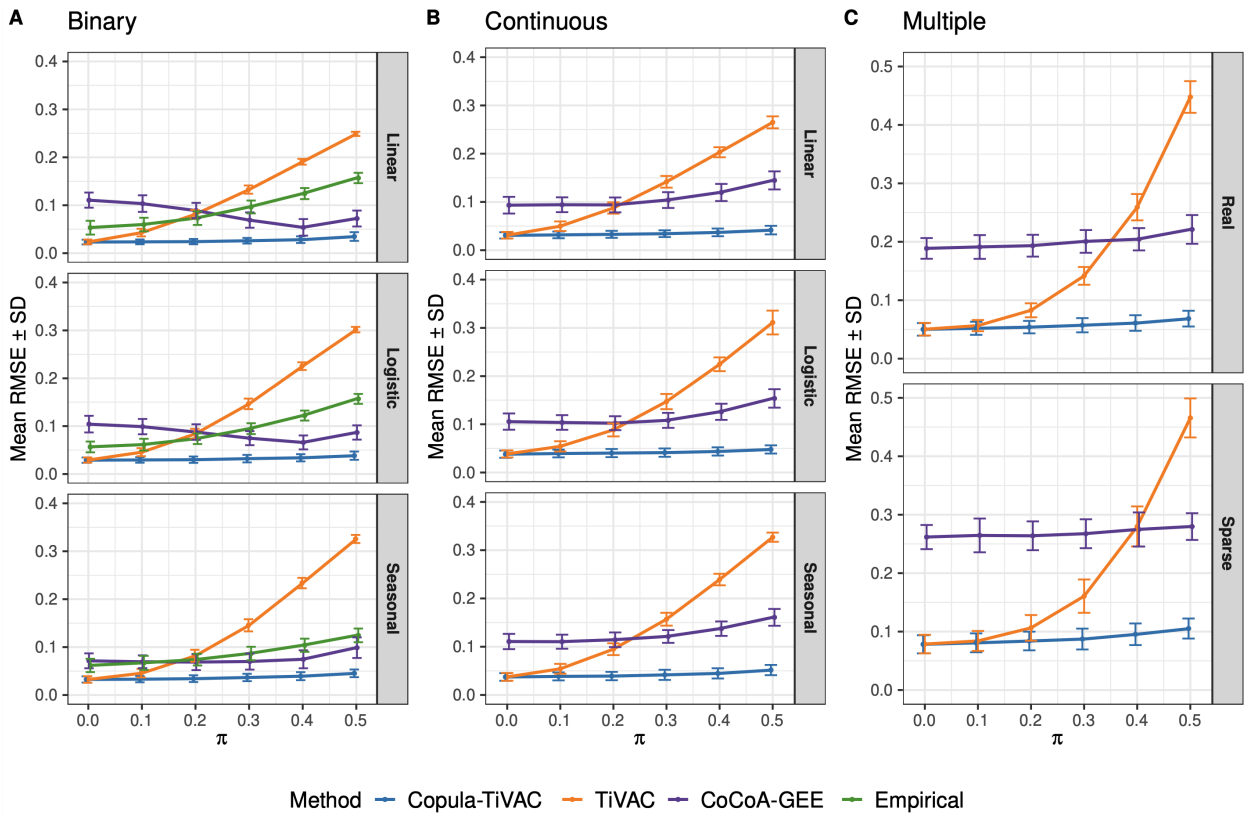


Figure 4.4: **RMSE of estimated correlations across zero proportions.** Mean RMSE ($\pm$ SD) of the estimated correlations $\rho(t, x)$ for $\pi \in \{0, 0.1, \ldots, 0.5\}$, under different scenarios and settings. Methods: Copula-TiVAC (blue), TiVAC (orange), CoCoA-GEE (purple), Empirical (green). **A**: Binary-covariate scenario under $\mathscr{T}_{\text{Real}}$ for the three trajectory patterns; **B**: Continuous-covariate scenario under $\mathscr{T}_{\text{Real}}$ for the three trajectory patterns (Empirical not applicable); **C**: Multiple-covariate scenario comparing $\mathscr{T}_{\text{Real}}$ vs. $\mathscr{T}_{\text{Sparse}}$.

Binary covariate. For the binary-covariate scenario, Panel A of Figure 4.4 reports the mean RMSE ($\pm$SD) for $\pi \in \{0, 0.1, \ldots, 0.5\}$ under $\mathscr{T}_{\text{Real}}$ and the three coefficient trajectory patterns. Copula–TiVAC is uniformly most accurate, and its error remains nearly flat as π increases, with mean RMSE consistently below 0.05. This robustness arises from the zero-inflated Gaussian copula framework, which retrieves a latent Gaussian layer that fully captures the dependence structure underlying the observed non-Gaussian outcomes. Panel A of Figure C.1 shows the analogous results under $\mathscr{T}_{\text{Sparse}}$, where Copula–TiVAC's advantage over CoCoA–GEE and the Empirical method is even larger. This is due to the P-spline representation in Copula–TiVAC, which borrows strength across nearby time points and yields stable estimates when observations are sparse. By contrast, TiVAC is equivalent to Copula–TiVAC at $\pi = 0$ (both transform the positive observations identically via the empirical cdf), but its accuracy worsens rapidly as π grows. As illustrated by the estimated correlation

trajectories in Figures 4.5 ($\mathscr{T}_{\text{Real}}$) and C.2 ($\mathscr{T}_{\text{Sparse}}$) for $\pi \in \{0.3, 0.5\}$ and $X \in \{0, 1\}$, TiVAC (yellow) tends to overestimate when $\rho(t, x) > 0$ and underestimate when $\rho(t, x) < 0$. This behavior arises because the empirical cdf in TiVAC transforms all zeros to a single latent value, creating a point mass at the lower tail; the resulting discrete–continuous mixture violates the Gaussian working assumption and biases the correlation estimates. For larger zero inflation ($\pi \gtrsim 0.3$), CoCoA–GEE and Empirical outperform TiVAC because Gaussian-based TiVAC model becomes increasingly misspecified as the point mass at zero grows, whereas CoCoA–GEE (a GEE2-based estimator) and the Empirical method do not rely on a latent Gaussian scale. Nevertheless, CoCoA–GEE and Empirical remain less accurate than Copula–TiVAC overall–and than TiVAC when π is small–because they estimate correlation pointwise in time, which does not fully exploit the temporal correlation structure. In the linear pattern (first row of Figure 4.5), the CoCoA–GEE curve (purple) shifts from slightly below the truth at $\pi = 0.3$ to above the truth at $\pi = 0.5$ for $X = 0$, reflecting the mild dip–then–rise in its RMSE curve seen in Panel A of Figure 4.4.

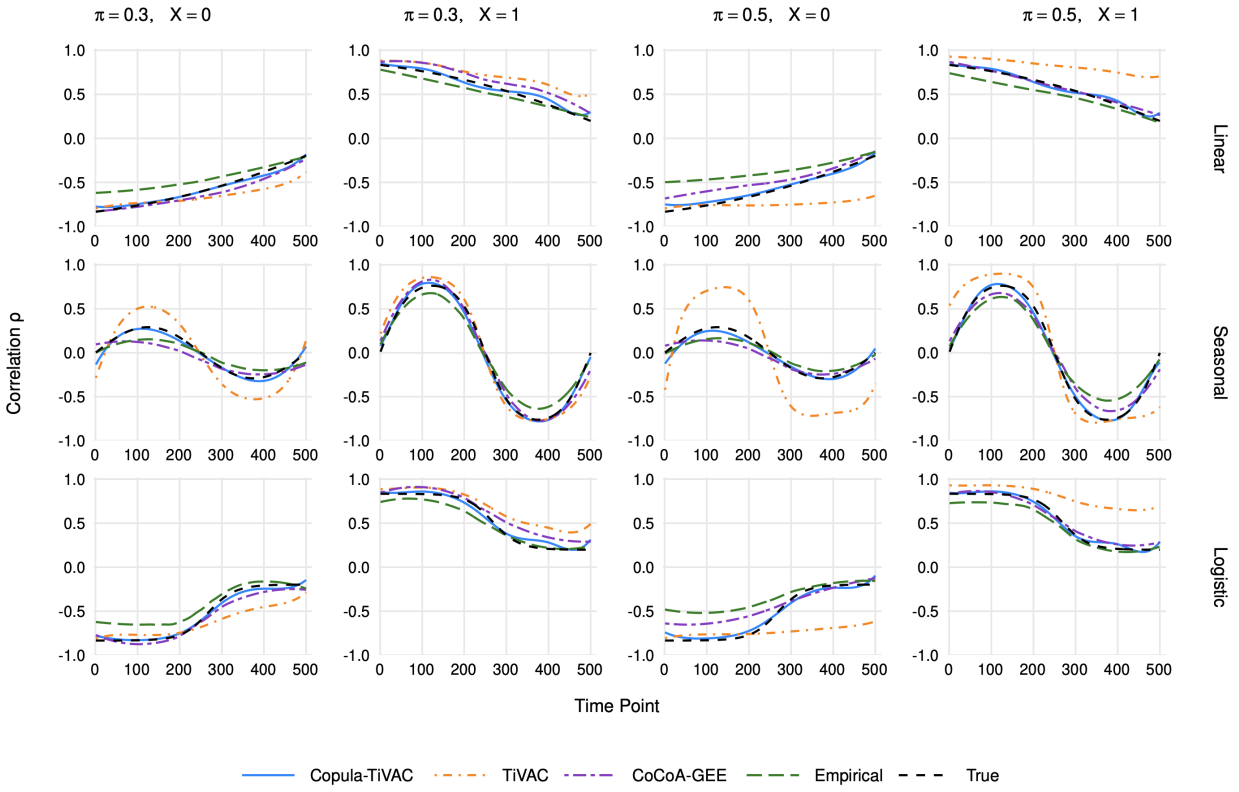


Figure 4.5: **Representative correlation curve estimations under $\mathscr{T}_{\text{Real}}$.** Comparison between the estimated correlation curves from four methods (Copula-TiVAC, TiVAC, CoCoA-GEE, and Empirical) and the true correlation curve for $X = 0$ and $X = 1$ at $\pi = 0.3$ and $\pi = 0.5$ under $\mathscr{T}_{\text{Real}}$ in binary-covariate scenario. Note that each subplot corresponds to a representative example from the first of 50 simulation runs (using a fixed random seed).

Continuous covariate. Panel B of Figure 4.4 displays the mean RMSE (±SD) for the continuous–covariate scenario under $\mathscr{T}_{\text{Real}}$, evaluated across the same three trajectory patterns values of π as in the binary case. Empirical is omitted because it does not accommodate continuous covariates. The analogous $\mathscr{T}_{\text{Sparse}}$ results are shown in Panel B of Figure C.1. Copula–TiVAC is consistently the most accurate and remains stable as π increases. From binary to continuous, the performance of Copula–TiVAC and TiVAC changes very little, whereas CoCoA–GEE worsens (RMSE approaches 0.3 under $\mathscr{T}_{\text{Sparse}}$). This pattern reflects the limitation of CoCoA–GEE's time–point-wise estimation; at each time point there are few distinct covariate values, leading to high instability of model fitting. We also visualize $\rho(t, x)$ in heatmaps for $\pi = 0.3$ and linear pattern, comparing $\mathscr{T}_{\text{Real}}$ and $\mathscr{T}_{\text{Sparse}}$ in Figure C.3. The heatmaps show that Copula–TiVAC closely recovers the truth, TiVAC exhibits the same over/underestimation pattern noted in the binary case, and CoCoA–GEE deviates notably from the truth, especially under $\mathscr{T}_{\text{Sparse}}$.

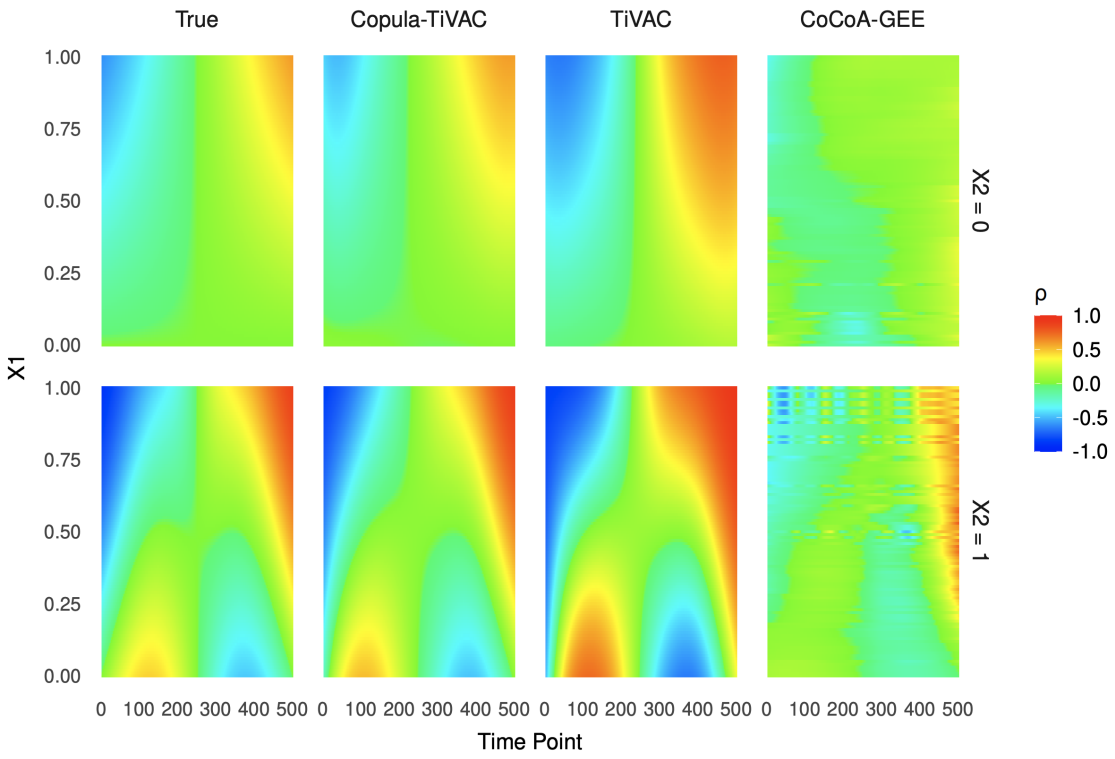


Figure 4.6: **Representative heatmap estimations under the multiple–covariate scenario ($\mathscr{T}_{\text{Real}}$).** Comparison between the true heatmap and estimates from Copula-TiVAC, TiVAC, and CoCoA-GEE. The heatmaps show how $\rho(t, x_1, x_2)$ change over time (x-axis) and the continuous covariate X_1 (y-axis); facet rows correspond to $X_2 = 0$ and $X_2 = 1$, and columns to methods. The setting includes $\mathscr{T}_{\text{Real}}$ and $\pi = 0.3$. Correlation is color-coded. Each heatmap corresponds to a representative example from the first of 50 simulation runs (fixed seed).

Multiple covariates. We include one continuous covariate X_1, one binary covariate X_2, and their interaction. Panel C of Figure 4.4 compares mean RMSE ($\pm$SD) under both $\mathscr{T}_{\text{Real}}$ and $\mathscr{T}_{\text{Sparse}}$. Relative to the single-covariate cases under $\mathscr{T}_{\text{Real}}$, Copula–TiVAC shows little change in error, indicating robustness to the addition of a few covariates. By contrast, TiVAC's RMSE increases more quickly as π increases, and CoCoA–GEE's RMSE rises by about 0.1. Figure 4.6 displays heatmaps of $\rho(t, x_1, x_2)$ by time (x-axis), X_1 (y-axis), and X_2 (facet rows) under $\mathscr{T}_{\text{Real}}$ with $\pi = 0.3$. Copula–TiVAC closely matches the true heatmaps, TiVAC shows the same over/underestimation pattern noted earlier, and CoCoA–GEE fails to recover the true structure. Similar results are observed under $\mathscr{T}_{\text{Sparse}}$ in Figure C.4.

4.5 Application to mood dynamics in bipolar disorder

We applied Copula-TiVAC to the Heinz C. Prechter Bipolar Research Project (PrBP) cohort (McInnis et al., 2017; Yocum et al., 2023) to quantify how the concurrent and pairwise correlations between PHQ-9 (depression), GAD-7 (anxiety), and ASRM (mania) scores change with age, affected by patient characteristics and clinical factors. BP-I is characterized by episodic mood swings between mania and depression (SHER, 2008). Comorbid anxiety disorders are also common in BP-I, with a lifetime risk of roughly 50% (Spoorthy et al., 2019), and anxiety is closely linked to depression (Kalin, 2020). Because these three symptom domains interact in complex and dynamic ways, understanding their joint behavior is key to improving symptom control in BP-I. Additional background on mood dynamics in bipolar disorder has been discussed in Section 4.1.

In PrBP, mood outcomes were collected approximately every two months, whereas medication records were updated annually and contain substantial missingness. We began with the preprocessing described by Liu et al. (2025), followed by additional cleaning steps: (1) to place subjects on a common time scale, we used each participant's age as the time variable and restricted analyses to ages 35–60, where observations are reasonably dense; and (2) for each outcome pair, we retained only time points at which both measures were observed. Age was recorded in months for analysis and displayed in years for visualization. Alcohol and cannabis use disorder diagnoses were obtained at enrollment and modeled as time-invariant binary indicators. Medication exposure was summarized as continuous covariates in two ways. First, for each participant we computed the mean count of nervous system-related medications (which include all psychiatric agents) across observed visits. Second, for each major class (antipsychotics, antidepressants, mood stabilizers) we defined a use frequency equal to the proportion of observed visits at which that class was taken.

As shown in Table C.1, cohort sizes differ by outcome pair, PHQ-ASRM ($n = 198$), GAD-

ASRM ($n = 185$), and PHQ-GAD ($n = 230$). Across pairs, about 72% of participants are female, alcohol and cannabis use disorders are common (about 45% and 27%, respectively), the mean number of nervous system-related medications is about 1.2, and mood stabilizers are used more often than antipsychotics or antidepressants on average. As shown in Figure 4.1, PHQ-9 and GAD-7 scores exhibit about 15% zeros, whereas ASRM shows roughly 40% zeros and a long right tail, which motivates our use of Copula-TiVAC to accommodate zero-inflated, non-Gaussian outcomes.

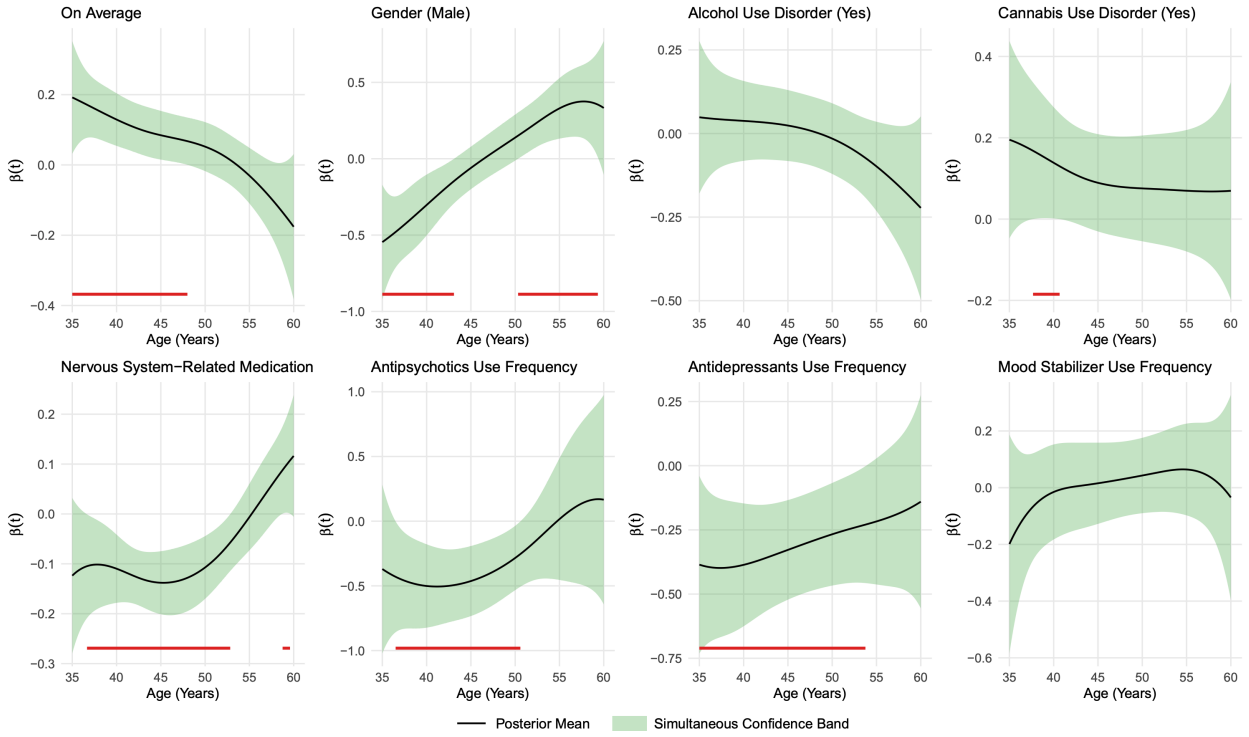


Figure 4.7: **Copula-TiVAC coefficient functions for PHQ-ASRM correlation.** Posterior mean (black) and 95% simultaneous band (green) for time-varying coefficients $\beta(t)$. The bottom bar (red) highlights the age intervals where the simultaneous confidence band does not include zero. The "On Average" panel shows the baseline trajectory without considering covariates. Remaining panels show $\beta_1(t)$ for each covariate: Gender (Male), Alcohol Use Disorder (Yes), Cannabis Use Disorder (Yes), Nervous System–Related Medication, Antipsychotics Use Frequency, Antidepressants Use Frequency, and Mood Stabilizer Use Frequency.

We fit eight Copula–TiVAC models: the first excludes covariate effects, and the remaining seven each correspond to one covariate of interest. To aid interpretation, we present two complementary sets of figures. First, Figures 4.7, C.5, and C.6 (for PHQ-ASRM, GAD-ASRM, and PHQ-GAD, respectively) display posterior means and 95% simultaneous bands for the time–varying coefficients. The top-left panel ("On Average") shows the baseline trajectory without covariate effect, and the remaining panels plot $\beta_1(t)$, the time–varying

covariate effect. Throughout, with the exception of the "On Average" panel, bands that exclude zero (corresponding to red bars) are interpreted as evidence of a significant covariate effect at the corresponding times. Second, Figures 4.8, C.7, and C.8 (again for PHQ-ASRM, GAD-ASRM, and PHQ-GAD) visualize the estimated time- and covariate-dependent correlations using trajectories (for binary covariates) and heatmaps (for continuous covariates). Thus, the first set identifies when a covariate has a significant effect on the coefficient functions, whereas the second set illustrates how the latent correlation between each outcome pair evolves with age and covariate values.

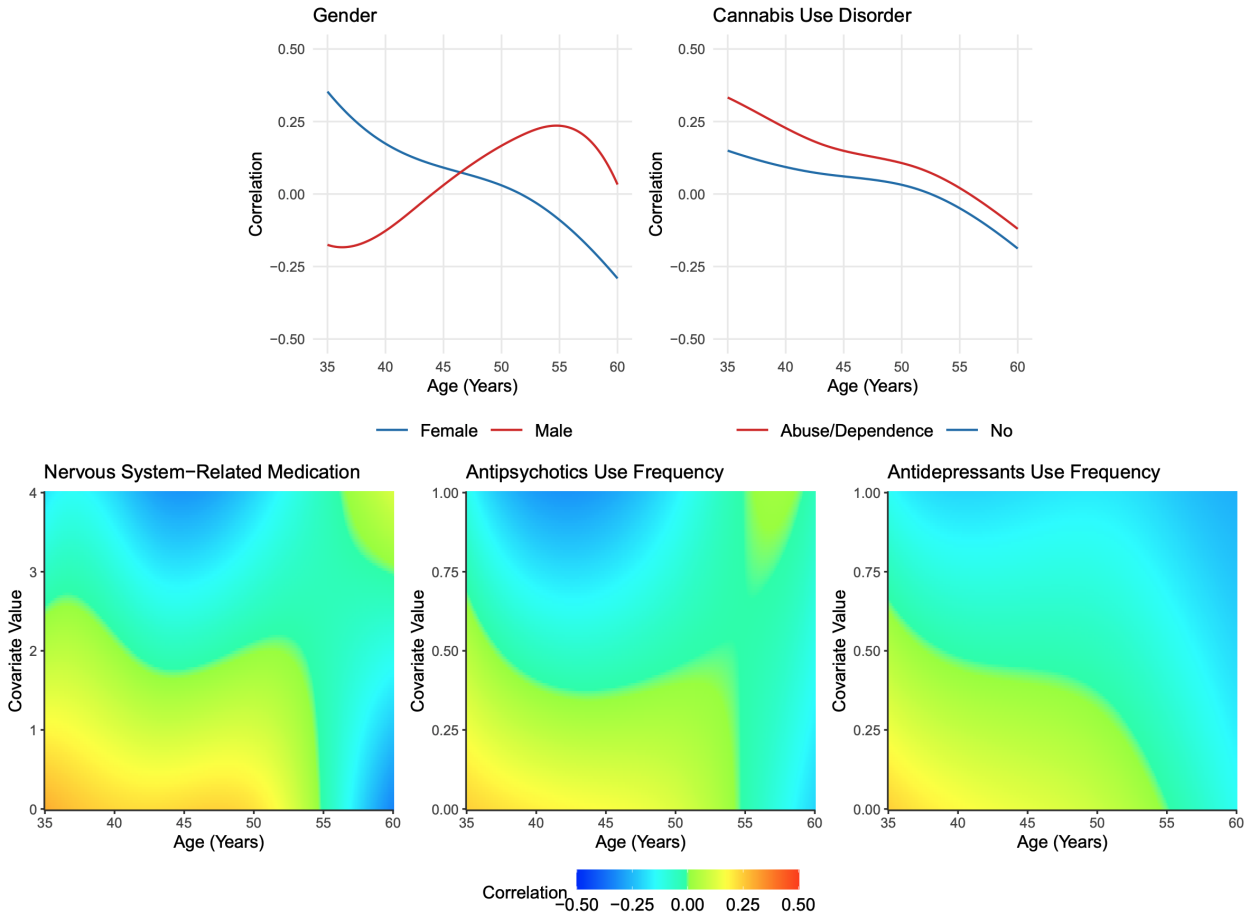


Figure 4.8: **Copula-TiVAC correlations between PHQ-9 and ASRM.** Estimated latent correlation between PHQ-9 and ASRM as a function of age and covariates. Top row shows curves for gender, and cannabis Use Disorder. Bottom row shows heatmaps for continuous covariates (Nervous System-Related Medication, Antipsychotics Use Frequency, Antidepressants Use Frequency). For the heatmaps, age is on the x-axis, covariate value on the y-axis, and the color scale ranges from -0.5 (blue) to 0.5 (red). Positive values indicate concurrent increases in depression and mania; negative values indicate inverse movement.

In Figure 4.7, higher nervous system–related medication use, higher antipsychotic use

frequency, and higher antidepressant use frequency are each significantly associated with a reduction in the concurrent PHQ–ASRM correlation around ages 35–50. As shown in Figure 4.8, before about age 55, greater exposure to these medications shifts the PHQ–ASRM correlation from positive (depressive and manic symptoms rising together) to negative. Statistically, this indicates decoupling of concurrent depressive and manic activation. Clinically, a shift from positive to negative correlation suggests better control of mixed states, consistent with the role of mood-stabilizing regimens in preventing concurrent manic–depressive symptomatology (Muneer, 2017). For the anxiety–mania pairing (GAD–ASRM), Figure C.5 shows that nervous system–related medication use and antidepressant use frequency have significant negative effects across ages 35–60, and antipsychotic use frequency is significantly negative around ages 40–50; Figure C.7 shows the GAD–ASRM correlation likewise moves from positive to negative with higher exposure. This implies that psychiatric treatment also reduces the co-occurrence of anxiety with mania.

Regarding gender, Figure 4.7 shows that male participants have a significantly lower PHQ–ASRM correlation than female participants around ages 35–45, but a significantly higher correlation around ages 50–60. In Figure 4.8, the PHQ–ASRM correlation in females decreases with age from positive to negative, whereas in males it increases from negative to positive. Prior work reports that women are somewhat more likely than men to have mixed manic episodes (depressive symptoms during mania) (Sit, 2004), consistent with our finding of a positive correlation in earlier mid-life. In females with bipolar disorder, the perimenopausal phase (around age 50) is associated with increased depressive symptoms and reduced manic symptoms (Dell'Osso et al., 2022), which can drive the PHQ–ASRM correlation negative. In males, those who begin with mania-predominant illness tend to accumulate more depressive episodes with age (Coryell et al., 2009; Dell'Osso et al., 2022), increasing the coupling of manic and depressive morbidity over time. Regarding cannabis use disorder, the time-varying coefficient plot (Figures 4.7) and the time-varying correlation plot (Figures 4.8) show that cannabis use disorder significantly raises the PHQ–ASRM correlations around age 40. This is consistent with evidence that cannabis use is related to an increased risk of mixed states (Bally et al., 2014).

In Figures C.6 and C.8, the patterns of medication and gender effects on PHQ–GAD correlation are similar to those discussed in the application section of Chapter 3. This likely reflects that PHQ-9 and GAD-7 are not highly zero-inflated and that TiVAC is robust for these slight zero-inflated outcomes. To directly compare TiVAC and Copula-TiVAC on zero-inflated data, we fit TiVAC with an empirical CDF transformation to the PHQ–ASRM pair (Figure C.9). By comparing Figure 4.8 for Copula-TiVAC with Figure C.9 for TiVAC, the trend for each covariate effect is similar between the methods, but when the true cor-

relation is positive TiVAC estimates are generally higher than Copula-TiVAC, and when the correlation is negative TiVAC estimates are generally less than Copula-TiVAC. This matches the over- and underestimation behavior of TiVAC demonstrated in our simulations (Section 4.4).

Overall, applying Copula–TiVAC to the PrBP data shows that greater medication exposure is associated with a decoupling of depression and mania (and of anxiety and mania) across mid-life, whereas cannabis use disorder tends to tighten these couplings. We also observe gender- and age-dependent patterns in the depression–mania correlation, which declines with age in females and rises in males. Together, these findings support the usefulness and validity of Copula–TiVAC for studying heterogeneity in mood dynamics.

4.6 Discussion

This chapter introduced the zero-inflated, Gaussian-copula–based Time-Varying and Covariate-Dependent correlation model (Copula-TiVAC), which generalizes the TiVAC framework to accommodate non-Gaussian outcomes with excess zeros. The key idea is to separate bivariate marginal distributions from dependence by retrieving a latent Gaussian layer that fully captures the dependence structure in the observed outcomes via a zero-inflated Gaussian copula. On the latent bivariate normal scale, the Fisher-transformed correlation is expressed as a linear function of covariates with coefficients that vary smoothly over time and are estimated with Bayesian P-splines. Bayesian estimation proceeds with a blocked Gibbs sampler that alternates between (i) updating latent Gaussian variables and zero-inflation thresholds and (ii) updating the time-varying coefficient functions using Hamiltonian Monte Carlo. Uncertainty in covariate effects over time is summarized using simultaneous confidence bands constructed from joint posterior draws.

Simulation studies showed that Copula-TiVAC consistently outperforms competing approaches in estimating time- and covariate-dependent correlation across a wide range of settings, including zero inflation, diverse temporal patterns and multiple covariates. Applied to the PrBP BP-I cohort, Copula-TiVAC revealed clinically interpretable patterns: greater exposure to psychiatric medications is associated with a decoupling of depression-mania and anxiety-mania correlations during mid-life; cannabis use disorder tends to tighten these couplings; and significant gender-by-age differences in depression-mania correlation. These findings illustrate how modeling correlation as a function of both time and covariates can uncover mechanisms that would be obscured by population-average analyses.

We outline several methodological and applied directions that build naturally on this chapter. First, introducing autocorrelation on the latent Gaussian scale would enable joint

modeling of within-subject temporal dependence alongside the concurrent correlation. Pursuing this goal via multivariate or Gaussian-process formulations will, however, increase model complexity and computational demands. Second, extending Copula-TiVAC to higher-dimensional outcomes is appealing; the focus could be on covariate effects on conditional correlations encoded in a time- and covariate-dependent precision matrix. Covariate-dependent precision matrices have been studied in graphical regression models with covariates (Zhang and Li, 2022). Third, the current zero-inflated Gaussian copula model can be generalized to the segmented Gaussian copula model, allowing multiple inflation points (Choi et al., 2024). Fourth, a natural next application is to apply Copula-TiVAC to the BP-II subgroup in PrBP to assess whether the time-varying covariate effects on concurrent correlations between depression, anxiety, and mania differ between BP-I and BP-II. At present, the effective BP-II cohort is small after preprocessing, which may limit reliable estimation. Fifth, many modern applications index observations over two-dimensional domains (e.g., spatial coordinates). Copula-TiVAC could be generalized by replacing the univariate time index with a spatial (or even spatiotemporal) coordinate to study relationships between single-cell phenotypes; a relevant application is immune-cell co-localization in the tumor microenvironment. Finally, it is of great interest to leverage unstructured clinical notes using privacy-guaranteed AI tools, and incorporate patients' genetic data, treating them as covariates to better understand mood dynamics and disease mechanisms.

CHAPTER 5

Conclusion

This dissertation develops novel AI-driven and advanced statistical learning methods for two application problems in precision medicine. The first is integrative modeling that bridges the gap between preclinical systems and patient tumors to predict patient drug response. The second is dynamic modeling that captures how correlations between mood outcomes change over time and with patient covariates in psychiatry. A common thread across these methods is the focus on modeling explained variation and co-variation in complex biomedical data. Together, they show how statistical methods can leverage complex biomedical and clinical data to address analytical and applied challenges in precision medicine. In this chapter, Section 5.1 summarizes the proposed statistical methods; Section 5.2 synthesizes key insights from real-world analyses; and Section 5.3 outlines future directions.

5.1 Summary of proposed statistical methods

Chapter 2 introduces integrative principal component regression (iPCR) for pan-cancer drug response prediction. The method decomposes gene expression from cell lines and patient tumors into shared structure, structures specific to either cell lines or patients, and noise. The shared structure contains latent features that capture common biology between the two systems, and these shared features are used to predict patient drug response using drug data measured in cell lines. The model also returns gene-level effects and infers gene co-expression networks, which support biological interpretation.

Chapter 3 proposes the Time-Varying and Covariate-Dependent (TiVAC) correlation model. The goal is to learn how the concurrent correlation between two outcomes depends on covariates across time. TiVAC works under a bivariate Gaussian framework. It applies the Fisher transform to the correlation and models the transformed parameter as a linear function with coefficients that vary smoothly over time. Nonlinearity and smoothness of coefficient curves over time are captured via B-splines, and estimation is based on a penal-

ized likelihood optimized by the Newton–Raphson algorithm. Last, a bootstrap procedure is proposed to construct simultaneous confidence bands for inference on entire coefficient curves.

Chapter 4 generalizes TiVAC to handle zero inflation and skewness in biomedical outcomes such as mood scores. To better study dynamic mood in bipolar disorder, I propose the zero-inflated Gaussian-copula–based TiVAC (Copula-TiVAC) model. This model first estimates an underlying latent Gaussian layer that, via a zero-inflated Gaussian copula, fully captures the dependence among the observed non-Gaussian, zero-inflated outcomes. Then, on the latent scale, the time-varying and covariate-dependent correlation is modeled with Bayesian P-splines. Bayesian estimation uses a blocked Gibbs sampler that cycles between two stages: one stage updates latent Gaussian variables; the other stage updates spline coefficients using Hamiltonian Monte Carlo. Finally, MCMC draws are used to construct simultaneous confidence bands that summarize uncertainty for entire coefficient curves.

To validate the proposed methods, I performed comprehensive simulations. For iPCR, I examined different similarity levels between patient tumors and cell lines in their genomic data and showed that iPCR predicts drug response with high accuracy consistently. For TiVAC and Copula-TiVAC, I considered diverse settings and showed that both methods outperform others under the Gaussian framework, while Copula-TiVAC performs best for zero-inflated and non-Gaussian data when estimating time-varying and covariate-dependent correlation.

5.2 Key insights from real-world analyses

Though cell lines have long served as foundational biological models in cancer research, the translational gap from cell lines to patient tumors is non-negligible and remains one of the main challenges in precision oncology. I applied iPCR to bridge this gap and predict patient drug response. The data included more than 40 cancer types, over 12,000 patient tumors, and more than 1,400 cell lines, covering over 1,700 cancer-related genes. iPCR outperformed six other methods on four drugs (Trastuzumab matched to Afatinib, Etoposide, Vinorelbine, and Oxaliplatin) and was competitive on the others. Methods that integrate patient and cell line genomics generally performed better than methods trained only on cell line data, highlighting the value of bridging the two biological systems. Consistent with this, iPCR showed that over half of the structured variation in genomic data is shared, while patient tumors retain notable system-specific structure, which aligns with greater *in vivo* complexity (Shi et al., 2019). In addition, gene-set enrichment analysis of iPCR-estimated gene effects recovered known mechanisms, and the shared gene co-expression networks supported identification of

hub genes. Together, these results support iPCR as a powerful model for cross-system drug response prediction and biological interpretation.

In the study of mood dynamics in bipolar disorder, traditional views suggested that increases in mania align with decreases in depression, and vice versa. However, recent work (Cochran et al., 2018) showed that the concurrent correlation between mania and depression can be negative or positive across individuals, revealing a more complex and heterogeneous pattern in mood dynamics in BP-I. Motivated by this heterogeneity, I applied Copula-TiVAC to study the time-varying and covariate-dependent concurrent correlations between PHQ-9 depression scores, GAD-7 anxiety scores, and ASRM mania scores in about 200 BP-I patients from the PrBP cohort (McInnis et al., 2017; Yocum et al., 2023). The contribution of our method is to enable examination of how patient characteristics and clinical factors shape this heterogeneous pattern over time.

Based on Copula-TiVAC's results, I found that higher exposure to psychiatric medications was associated with a shift from positive to negative concurrent correlation between depression and mania, and between anxiety and mania, during mid-life. This indicates decoupling of depressive or anxiety symptoms from manic activation (Muneer, 2017). Clinically, this pattern reflects that appropriate psychiatric medication can reduce the risk of mixed states, which are linked to lower quality of life and a higher risk of suicide. In addition, cannabis use disorder tended to increase these correlations, consistent with a greater risk of mixed states and aligned with previous research (Bally et al., 2014). The correlation between depression and mania also varies by gender and across age. In females the correlation decreased with age, while in males it increased, a pattern supported by prior studies (Sit, 2004; Coryell et al., 2009; Dell'Osso et al., 2022). For depression and anxiety, consistently high TiVAC correlations highlight the well-documented comorbidity between them (Kalin, 2020). This correlation also increased slightly with age on average, and was also influenced by gender and psychiatric medications. Together, these findings support our Copula-TiVAC as an effective and valuable approach to study heterogeneity in mood dynamics in BP-I.

5.3 Future directions

The methods in this dissertation can be generalized in many promising directions in both modeling and application.

For iPCR, several extensions are natural. First, at the sample level, we can move beyond two biological model systems by integrating cell lines, patient-derived xenografts, organoids, and patient tumors within a unified framework. Extending iPCR using the SLIDE model (Gaynanova and Li, 2019) would enable characterization of shared, partially shared, and

system-specific structure across these systems and facilitate downstream analyses. Second, at the feature level, the model can be extended from genomics to multi-omics to provide deeper insight into shared and distinct biology across preclinical models and patient tumors. Third, out-of-sample prediction for new patients is a priority. This requires standardized preprocessing and a validated projection step. This direction is especially appealing because it potentially enables real-time drug-response prediction: a patient's specimen could be profiled during a hospital visit, processed, and returned as an individualized treatment-guidance report. Finally, generalizing Stage I of iPCR with nonlinear approaches, such as deep learning, offers a path to accommodate complex omics data structures.

TiVAC and Copula–TiVAC can be generalized in several methodological directions. One direction is to add latent autocorrelation so that within-person temporal dependence is modeled together with concurrent correlation. This can better capture the dynamic pattern of mood states and can be achieved with multivariate or Gaussian-process formulations, although these increase model complexity and may reduce interpretability. Another direction is to extend beyond pairwise modeling to characterize the joint dependence among the three mood outcomes and, more generally, higher-dimensional outcomes, by modeling time-varying and covariate-dependent partial correlations. Such high-dimensional extensions would also enable applications to dynamic gene co-expression and brain networks. It is also promising to generalize from a one-dimensional time scale to two-dimensional spatial coordinates and even to a three-dimensional spatiotemporal space, supporting studies of relationships among single-cell phenotypes using spatial or spatiotemporal data. In addition, the latent linear model in TiVAC can be extended to semi- or nonparametric forms or augmented with advanced AI/ML components, and static covariates can be generalized to time-varying covariates; these gains in flexibility may reduce interpretability to varying degrees. Although our analyses are observational, examining causal effects of covariates on correlation remains possible under additional assumptions.

For applications of Copula-TiVAC to mood dynamics, several extensions are promising and high-impact, warranting focused attention. First, in practice, combined pharmacotherapy often outperforms monotherapy but may introduce unanticipated adverse effects; this motivates interaction analyses to quantify how medication classes (e.g., antidepressants, antipsychotics, mood stabilizers) jointly shape age-dependent correlations between mood outcomes. Second, mixed states are a central clinical motivation in Chapter 4. We will pursue deeper investigation by conducting individualized analyses of mood trajectories among patients exhibiting mixed-state features (e.g., positive concurrent depression-mania correlation with co-occurring symptoms). It is also of great interest to consider genetic effects in future work. If patients' DNA or RNA sequencing data can be acquired, this could provide

deeper insight into disease mechanisms and contribute to precision psychiatry. Unstructured clinical notes are another rich source for studying mood dynamics, since they contain many details of patient feedback on symptoms, treatment efficacy, and adverse events. Future work will use privacy-guaranteed AI tools to extract useful clinical data from unstructured notes. In addition, applying Copula–TiVAC to the BP-II subgroup would enable comparisons between BP-I and BP-II, quantifying subtype differences in the heterogeneity of mood dynamics. Overall, these research directions would boost progress and could be a critical game changer for studying mood dynamics in psychiatry.

In conclusion, this dissertation shows that integrative modeling and dynamic correlation modeling can turn complex biomedical data into interpretable and clinically meaningful signals. The proposed methods also lay a foundation for further methodological generalization and broader applications in future studies.

APPENDIX A

Appendix Chapter 2

In this appendix, we first summarize all the mathematical symbols and notations used throughout the main text in Section A.1. Then, we address the estimation of the gene co-expression network in Section A.2, including the proof of the uniqueness and identifiability for the covariance decomposition. Following this, Section A.3 provides additional details on our simulation studies, specifically focusing on the sensitivity analysis of rank selection for iPCR. Lastly, Section A.4 showcases additional results from our pan-cancer application.

A.1 Mathematical notations

Table A.1 summarizes the definitions and interpretations of all mathematical symbols and notations used throughout the main text.

Table A.1: **Summary of notation.**

Notation	Description	Interpretation
n_c	Number of cell lines	
n_p	Number of patients	
n	Number of cell lines and patients	
g	Number of gene features	
$\mathbf{C} \in \mathbb{R}^{n_c \times g}$	Gene expression matrix of cell lines	
$\mathbf{P} \in \mathbb{R}^{n_p \times g}$	Gene expression matrix of patients	
$\mathbf{G} \in \mathbb{R}^{n \times g}$	Gene expression matrix for cell lines and patients	
$\mathbf{D}_c \in \mathbb{R}^{n_c}$	Drug data of cell lines	
$\mathbf{D}_p^* \in \mathbb{R}^{n_p}$	Unobserved drug data of patients	
$row(\cdot)$	Row space of a matrix	
$col(\cdot)$	Column space of a matrix	
$\boldsymbol{J} \in \mathbb{R}^{n \times g}$	Joint structure matrix	Low-rank joint structure that captures the common features between cell lines and patients.
$\boldsymbol{J_c} \in \mathbb{R}^{n_c \times g}$	Sub-matrix of the joint structure matrix for cell lines	Shares the same feature space with $\boldsymbol{J_p}$.
$\boldsymbol{J_p} \in \mathbb{R}^{n_p \times g}$	Sub-matrix of the joint structure matrix for patients	Shares the same feature space with $\boldsymbol{J_c}$.
$\boldsymbol{I_c} \in \mathbb{R}^{n_c \times g}$	Individual structure matrix of cell lines	Captures the features that are not shared with patients.
$\boldsymbol{I_p} \in \mathbb{R}^{n_p \times g}$	Individual structure matrix of patients	Captures the features that are not shared with cell lines.
$\boldsymbol{\varepsilon_c} \in \mathbb{R}^{n_c \times g}$	Error matrix of cell lines	
$\boldsymbol{\varepsilon_p} \in \mathbb{R}^{n_p \times g}$	Error matrix of patients	
r_J	Joint rank	The rank of $\boldsymbol{J}$, $\boldsymbol{J_c}$ and $\boldsymbol{J_p}$. The dimension of the latent shared feature space.
r_{I_c}	Individual rank of cell lines	The rank of $\boldsymbol{I_c}$. The dimension of the latent cell line-specific feature space.
r_{I_p}	Individual rank of patients	The rank of $\boldsymbol{I_p}$. The dimension of the latent patient-specific feature space.
$\boldsymbol{S}_J \in \mathbb{R}^{n \times r_J}$	Joint score matrix of cell lines and patients	Contains r_J joint principal scores from $\boldsymbol{J}$, which can be viewed as the latent shared features of cell lines and patients, capturing their biological similarity.
$\boldsymbol{S}_{J_c} \in \mathbb{R}^{n_c \times r_J}$	Joint score matrix of cell lines	Shares the same feature space with $\boldsymbol{S}_{J_p}$.
$\boldsymbol{S}_{J_p} \in \mathbb{R}^{n_p \times r_J}$	Joint score matrix of patients	Shares the same feature space with $\boldsymbol{S}_{J_c}$.
$\boldsymbol{S}_{I_c} \in \mathbb{R}^{n_c \times r_{I_c}}$	Individual score matrix of cell lines	Contains r_{I_c} latent features individual to cell lines.
$\boldsymbol{S}_{I_p} \in \mathbb{R}^{n_p \times r_{I_p}}$	Individual score matrix of patients	Contains r_{I_p} latent features individual to patients.
$\boldsymbol{V}_J \in \mathbb{R}^{g \times r_J}$	Joint loading matrix	Quantifies the extent to which each original feature in $\boldsymbol{J}$ is related to the latent shared features in $\boldsymbol{S}_J$.
$\boldsymbol{V}_{I_c} \in \mathbb{R}^{g \times r_{I_c}}$	Individual loading matrix of cell lines	Quantifies the extent to which each original feature in $\boldsymbol{I}_c$ is related to the latent individual features in $\boldsymbol{S}_{I_c}$.
$\boldsymbol{V}_{I_p} \in \mathbb{R}^{g \times r_{I_p}}$	Individual loading matrix of patients	Quantifies the extent to which each original feature in $\boldsymbol{I}_p$ is related to the latent individual features in $\boldsymbol{S}_{I_p}$.
$\beta_0 \in \mathbb{R}$	Constant intercept in Model 3	Baseline drug response in Model 3 before considering gene effects.
$\boldsymbol{\beta}_J \in \mathbb{R}^{r_J}$	A vector of regression coefficients in Model 3	Measures the effect of the corresponding latent shared features on drug responses.
$\boldsymbol{\beta}_{genes} \in \mathbb{R}^g$	A vector of transformed $\boldsymbol{\beta}_J$ in the gene space	Measures the shared gene effects between cell lines and patients on drug responses.
$\boldsymbol{\epsilon} \in \mathbb{R}^{n_c}$	Error vector in Model 3	
V	Set of nodes (genes) in the gene co-expression network	
E	Set of edges (relationships between genes) in the gene co-expression network	
$\tilde{\boldsymbol{G}}_i \in \mathbb{R}^{n_i \times g}$	For $i \in \{c, p\}$, $\tilde{\boldsymbol{G}}_i = \hat{\boldsymbol{J}}_i + \hat{\boldsymbol{I}}_i$	Estimated denoised gene expression matrix of cell lines or patients from iPCR outputs that are used as input in gene expression estimation.
$\tilde{\boldsymbol{J}}_i \in \mathbb{R}^{n_i \times g}$	Transformed joint structure matrix for either cell lines or patients	Used to build shared gene co-expression network between cell liens and patients.
$\tilde{\boldsymbol{I}}_i \in \mathbb{R}^{n_i \times g}$	Transformed individual structure matrix for either cell lines or patients	Used to build model-specific gene co-expression network.

A.2 Gene co-expression network estimation

A.2.1 The solution of sample covariance decomposition

For all $i \in \{c, p\}$, the unique solutions of $\tilde{\boldsymbol{J}}_i$ and $\tilde{\boldsymbol{I}}_i$ that meet all the conditions in Proposition 2 are given by

$$\tilde{\boldsymbol{J}}_i = \mathbf{P}_{\Phi_i} \tilde{\boldsymbol{G}}_i \mathbf{P}_{\Theta} \tag{A.1}$$

$$\tilde{\boldsymbol{I}}_i = \mathbf{P}_{\Phi_i^c} \tilde{\boldsymbol{G}}_i, \tag{A.2}$$

where $\Theta = row(\tilde{\boldsymbol{G}}_c) \cap row(\tilde{\boldsymbol{G}}_p)$ is the shared feature space, Θ_i^c is the orthogonal complement of Θ in $row(\tilde{\boldsymbol{G}}_i)$, Φ_i is the orthogonal complement of $col(\tilde{\boldsymbol{G}}_i \mathbf{P}_{\Theta_i^c})$ in $col(\tilde{\boldsymbol{G}}_i)$, and Φ_i^c is equal to $col(\tilde{\boldsymbol{G}}_i \mathbf{P}_{\Theta_i^c})$. Based on the proof of Proposition 1 in [1], Θ is equivalent with $row(\hat{\boldsymbol{J}})$, and $\mathbf{P}_{\Theta} = \hat{\boldsymbol{V}}_J \hat{\boldsymbol{V}}_J^{\top}$. The projection matrices $\mathbf{P}_{\Phi_i}$ and $\mathbf{P}_{\Phi_i^c}$ can be easily acquired following their definitions.

Note that each column of $\tilde{\boldsymbol{G}}_i$ (a gene feature) always has a mean of 0 due to the centralization for each column of $\mathbf{C}$ and $\mathbf{P}$ as well as the JIVE estimation algorithm [2] in iPCR Stage I. Then, based on Corollary 1 in Section A.2.2, the unique solutions for $\tilde{\boldsymbol{J}}_i$ and $\tilde{\boldsymbol{I}}_i$ satisfy the property of sample covariance decomposition as described in Model 2.6.

A.2.2 Uniqueness and identifiability of the decompostion solution

Lemma 1. *Given a set of matrices $\{A,\ B\ ,C\}$, if $A = B + C$ and $rank(A) = rank(B) + rank(C)$, then $row(A) = row(B) \cup row(C)$ and $col(A) = col(B) \cup col(C)$.*

Proof. : Denote $a = rank(A)$, $b = rank(B)$, $c = rank(C)$, and $dim(A)$ as $m \times$n.
We first prove $row(A) = row(B) \cup row(C)$ by showing $row(A) \subseteq row(B) \cup row(C)$, and $row(B) \cup row(C) \subseteq row(A)$.
1): It is straightforward to show that $row(A) \subseteq row(B) \cup row(C)$ since each row of A is a linear combination of the corresponding rows of B and C.
2): To show $row(B) \cup row(C) \subseteq row(A)$, first do linear transformation of rows in B to get B' where only first b rows are nonzero, such that $B' = \begin{pmatrix} B'_{[1:b,:]} \\ \mathbf{0} \end{pmatrix}$. Then do linear transformation of rows in C to get C' such that only c rows are nonzero, and the nonzero rows in B' and C' do not have a common index (because of the rank constraint), such that $C' = \begin{pmatrix} \mathbf{0}_{b \times n} \\ C'_{[(b+1):(b+c),:]} \\ \mathbf{0} \end{pmatrix}$.
Those c rows, which form the basis for $row(C)$, belongs to $row(A)$ since the corresponding rows in B' has been zeroed out. Similarly, it can be easily shown that $row(B)$ also belongs

to $row(A)$. Hence, $row(B) \cup row(C) \subseteq row(A)$.
Same proof can be used to show $col(A) = col(B) \cup col(C)$ after transpose A, B, and C. $\square$

Proposition 2. *Given a set of matrices* $\{\tilde{\boldsymbol{G}}_i, i = c, p\}$*, there are unique sets of matrices* $\{\tilde{\boldsymbol{J}}_i, i = c, p\}$ *and* $\{\tilde{\boldsymbol{I}}_i, i = c, p\}$ *such that,*

(1) $\tilde{\boldsymbol{G}}_i = \tilde{\boldsymbol{J}}_i + \tilde{\boldsymbol{I}}_i$ *for* $i = c,\ p$

(2) $col(\tilde{\boldsymbol{J}}_i) \perp col(\tilde{\boldsymbol{I}}_i) = 0$ *for* $i = c,\ p$

(3) $row(\tilde{\boldsymbol{J}}_i) = row(\tilde{\boldsymbol{J}})$ *for* $i = c,\ p$*, where* $\tilde{\boldsymbol{J}} = \begin{bmatrix} \tilde{\boldsymbol{J}}_c \\ \tilde{\boldsymbol{J}}_p \end{bmatrix}$

(4) $row(\tilde{\boldsymbol{J}}_i) \cap row(\tilde{\boldsymbol{I}}_i) = 0$ *for* $i = c,\ p$

(5) $row(\tilde{\boldsymbol{I}}_c) \cap row(\tilde{\boldsymbol{I}}_p) = 0$

Proof. : Let $i = c, p$ for this proof. Define $row(\tilde{\boldsymbol{G}}_c) \cap row(\tilde{\boldsymbol{G}}_p) = \Theta$, and Θ_i^c as the orthogonal complement of Θ in $row(\tilde{\boldsymbol{G}}_i)$.

Existence:

When $\Theta = \{\vec{0}\}$:

It is obvious that Condition 1-5 are all satisfied as $\tilde{\boldsymbol{J}}_i = \mathbf{0}$, and $\tilde{\boldsymbol{I}}_i = \tilde{\boldsymbol{G}}_i$.

When $\Theta \neq \{\vec{0}\}$:

Define Φ_i as the orthogonal complement of $col(\tilde{\boldsymbol{G}}_i \mathbf{P}_{\Theta_i^c})$ in $col(\tilde{\boldsymbol{G}}_i)$, and Φ_i^c as $col(\tilde{\boldsymbol{G}}_i \mathbf{P}_{\Theta_i^c})$. We can derive that

$$dim(\Phi_i) = rank(\tilde{\boldsymbol{G}}_i) - rank(\tilde{\boldsymbol{G}}_i \mathbf{P}_{\Theta_i^c}) = rank(\tilde{\boldsymbol{G}}_i) - \{rank(\tilde{\boldsymbol{G}}_i) - dim(\Theta)\} = dim(\Theta) \tag{A.3}$$

Then we will show that $\tilde{\boldsymbol{J}}_i$ and $\tilde{\boldsymbol{I}}_i$ defined below satisfy Condition 1-5:

$$\tilde{\boldsymbol{J}}_i = \mathbf{P}_{\Phi_i} \tilde{\boldsymbol{G}}_i \mathbf{P}_{\Theta} \tag{A.4}$$

$$\tilde{\boldsymbol{I}}_i = \mathbf{P}_{\Phi_i^c} \tilde{\boldsymbol{G}}_i. \tag{A.5}$$

Condition 1:

$$\begin{aligned} \tilde{\boldsymbol{J}}_i + \tilde{\boldsymbol{I}}_i &= \mathbf{P}_{\Phi_i} \tilde{\boldsymbol{G}}_i \mathbf{P}_{\Theta} + \mathbf{P}_{\Phi_i^c} \tilde{\boldsymbol{G}}_i \\ &= \mathbf{P}_{\Phi_i} \tilde{\boldsymbol{G}}_i \mathbf{P}_{\Theta} + \underbrace{\mathbf{P}_{\Phi_i} \tilde{\boldsymbol{G}}_i \mathbf{P}_{\Theta_i^c}}_{\mathbf{0}} + \mathbf{P}_{\Phi_i^c} (\underbrace{\tilde{\boldsymbol{G}}_i \mathbf{P}_{\Theta} + \tilde{\boldsymbol{G}}_i \mathbf{P}_{\Theta_i^c}}_{\tilde{\boldsymbol{G}}_i}) \\ &= \mathbf{P}_{\Phi_i} (\underbrace{\tilde{\boldsymbol{G}}_i \mathbf{P}_{\Theta} + \tilde{\boldsymbol{G}}_i \mathbf{P}_{\Theta_i^c}}_{\tilde{\boldsymbol{G}}_i}) + \mathbf{P}_{\Phi_i^c} (\underbrace{\tilde{\boldsymbol{G}}_i \mathbf{P}_{\Theta} + \tilde{\boldsymbol{G}}_i \mathbf{P}_{\Theta_i^c}}_{\tilde{\boldsymbol{G}}_i}) = \tilde{\boldsymbol{G}}_i \end{aligned}$$

Condition 2: Since Φ_i is orthogonal with Φ_i^c, then Condition 2 is satisfied.

Condition 3:

$$rank(\tilde{\boldsymbol{J}}_i) = rank(\mathbf{P}_{\Phi_i}\tilde{\boldsymbol{G}}_i\mathbf{P}_\Theta) \leq rank(\mathbf{P}_\Theta) = dim(\Theta) \tag{A.6}$$

$$rank(\tilde{\boldsymbol{J}}_i) \geq rank(\tilde{\boldsymbol{G}}_i) - rank(\tilde{\boldsymbol{I}}_i) \overset{A.3}{=} rank(\tilde{\boldsymbol{G}}_i) - (rank(\tilde{\boldsymbol{G}}_i) - dim(\Theta)) = dim(\Theta) \tag{A.7}$$

Based on Equation A.6 and A.7, we can get $rank(\tilde{\boldsymbol{J}}_i) = dim(\Theta)$. Since we also know that $row(\tilde{\boldsymbol{J}}_i) \subseteq \Theta$ from Equation A.4, we can conclude that $row(\tilde{\boldsymbol{J}}_i) = \Theta$ and $row(\tilde{\boldsymbol{J}}_i) = row(\tilde{\boldsymbol{J}})$.

Condition 4: Since $rank(\tilde{\boldsymbol{J}}_i) = dim(\Theta)$ and $rank(\tilde{\boldsymbol{I}}_i) = rank(\tilde{\boldsymbol{G}}_i) - dim(\Theta)$, then

$$rank(\tilde{\boldsymbol{G}}_i) = rank(\tilde{\boldsymbol{J}}_i) + rank(\tilde{\boldsymbol{I}}_i) \tag{A.8}$$

Thus, Condition 4 is satisfied based on the theorem from [3].

Condition 5: Since $row(\tilde{\boldsymbol{J}}_i) \cap row(\tilde{\boldsymbol{I}}_i) = 0$ from Condition 4 and $row(\tilde{\boldsymbol{J}}_i) \cup row(\tilde{\boldsymbol{I}}_i) = row(\tilde{\boldsymbol{G}}_i)$ from Lemma 1, then $row(\tilde{\boldsymbol{I}}_i) = row(\tilde{\boldsymbol{G}}_i) \backslash row(\tilde{\boldsymbol{J}}_i) = row(\tilde{\boldsymbol{G}}_i) \backslash \Theta = row(\tilde{\boldsymbol{G}}_i) \backslash (row(\tilde{\boldsymbol{G}}_c) \cap row(\tilde{\boldsymbol{G}}_p))$. Hence, $row(\tilde{\boldsymbol{I}}_c) \cap row(\tilde{\boldsymbol{I}}_p) = 0$. **Uniqueness**:
Assume there exist another $\tilde{\boldsymbol{J}}_i'$ and $\tilde{\boldsymbol{I}}_i'$ satisfying Condition 1-5 such that $\tilde{\boldsymbol{J}}_i' \neq \tilde{\boldsymbol{J}}_i$, $\tilde{\boldsymbol{I}}_i' \neq \tilde{\boldsymbol{I}}_i$ from part 1. Since $\tilde{\boldsymbol{J}}_i'$ and $\tilde{\boldsymbol{I}}_i'$ satisfy Condition 1, 2 and 4, then

$$rank(\tilde{\boldsymbol{G}}_i) = rank(\tilde{\boldsymbol{J}}_i' + \tilde{\boldsymbol{I}}_i') = rank(\tilde{\boldsymbol{J}}_i') + rank(\tilde{\boldsymbol{I}}_i') \tag{A.9}$$

based on the theorem from [3].
Based on Equation A.9, Lemma 1 and Condition 3-4, we have

$$row(\tilde{\boldsymbol{J}}') \cup row(\tilde{\boldsymbol{I}}_i') = row(\tilde{\boldsymbol{G}}_i'), row(\boldsymbol{J}') \cap row(\tilde{\boldsymbol{I}}_i') = 0 \tag{A.10}$$

$$col(\tilde{\boldsymbol{J}}_i') \cup col(\tilde{\boldsymbol{I}}_i') = col(\tilde{\boldsymbol{G}}_i'), col(\tilde{\boldsymbol{J}}_i') \cap col(\tilde{\boldsymbol{I}}_i') = 0. \tag{A.11}$$

By integrating Equation A.10 with Condition 5, we can easily show $row(\tilde{\boldsymbol{J}}_i') = row(\tilde{\boldsymbol{J}}') = row(\tilde{\boldsymbol{G}}_c) \cap row(\tilde{\boldsymbol{G}}_p) = \Theta$.

Based on Equation A.11 and Condition 2, we can rewrite $\tilde{\boldsymbol{J}}_i'$ and $\tilde{\boldsymbol{I}}_i'$ as

$$\tilde{\boldsymbol{J}}_i' = \mathbf{P}_{col(\tilde{\boldsymbol{J}}_i')}\tilde{\boldsymbol{G}}_i = \mathbf{P}_{col(\tilde{\boldsymbol{J}}_i')}\tilde{\boldsymbol{G}}_i\mathbf{P}_\Theta \tag{A.12}$$

$$\tilde{\boldsymbol{I}}_i' = \mathbf{P}_{col(\tilde{\boldsymbol{I}}_i')}\tilde{\boldsymbol{G}}_i = \mathbf{P}_{col(\tilde{\boldsymbol{J}}_i')^c}\tilde{\boldsymbol{G}}_i, \tag{A.13}$$

where $col(\tilde{\boldsymbol{J}}_i')^c$ is the orthogonal complement of $col(\tilde{\boldsymbol{J}}_i')$ in $col(\tilde{\boldsymbol{G}}_i)$.
By combining Equation A.12, A.13 and Condition 1, we have

$$\begin{aligned}
\tilde{\boldsymbol{G}}_i &= \mathbf{P}_{col(\tilde{\boldsymbol{J}}_i')}\tilde{\boldsymbol{G}}_i\mathbf{P}_\Theta + \mathbf{P}_{col(\tilde{\boldsymbol{J}}_i')^c}\tilde{\boldsymbol{G}}_i \\
\tilde{\boldsymbol{G}}_i &= \mathbf{P}_{col(\tilde{\boldsymbol{J}}_i')}\tilde{\boldsymbol{G}}_i\mathbf{P}_\Theta + \mathbf{P}_{col(\tilde{\boldsymbol{J}}_i')^c}(\tilde{\boldsymbol{G}}_i\mathbf{P}_\Theta + \tilde{\boldsymbol{G}}_i\mathbf{P}_{\Theta_i^c}) \\
\tilde{\boldsymbol{G}}_i + \mathbf{P}_{col(\tilde{\boldsymbol{J}}_i')}\tilde{\boldsymbol{G}}_i\mathbf{P}_{\Theta^c} &= \mathbf{P}_{col(\tilde{\boldsymbol{J}}_i')}(\tilde{\boldsymbol{G}}_i\mathbf{P}_\Theta + \tilde{\boldsymbol{G}}_i\mathbf{P}_{\Theta_i^c}) + \mathbf{P}_{col(\tilde{\boldsymbol{J}}_i')^c}(\tilde{\boldsymbol{G}}_i\mathbf{P}_\Theta + \tilde{\boldsymbol{G}}_i\mathbf{P}_{\Theta_i^c}) \\
\tilde{\boldsymbol{G}}_i + \mathbf{P}_{col(\tilde{\boldsymbol{J}}_i')}\tilde{\boldsymbol{G}}_i\mathbf{P}_{\Theta^c} &= (\mathbf{P}_{col(\tilde{\boldsymbol{J}}_i')} + \mathbf{P}_{col(\tilde{\boldsymbol{J}}_i')^c})\tilde{\boldsymbol{G}}_i \\
\tilde{\boldsymbol{G}}_i + \mathbf{P}_{col(\tilde{\boldsymbol{J}}_i')}\tilde{\boldsymbol{G}}_i\mathbf{P}_{\Theta^c} &= \tilde{\boldsymbol{G}}_i \\
\mathbf{P}_{col(\tilde{\boldsymbol{J}}_i')}\tilde{\boldsymbol{G}}_i\mathbf{P}_{\Theta^c} &= 0.
\end{aligned} \tag{A.14}$$

Since $\mathbf{P}_{col(\tilde{\boldsymbol{J}}_i')} \subseteq col(\tilde{\boldsymbol{G}}_i), \tilde{\boldsymbol{G}}_i\mathbf{P}_{\Theta^c} \subseteq col(\tilde{\boldsymbol{G}}_i)$, $dim(col(\tilde{\boldsymbol{J}}_i')) = dim(row(\tilde{\boldsymbol{J}}_i')) = dim(\Theta)$, and $rank(\tilde{\boldsymbol{G}}_i\mathbf{P}_{\Theta^c}) = rank(\tilde{\boldsymbol{G}}_i) - dim(\Theta)$, then $col(\tilde{\boldsymbol{J}}_i')$ must be equal to the orthogonal complement of $col(\tilde{\boldsymbol{G}}_i\mathbf{P}_{\Theta^c})$ in $col(\tilde{\boldsymbol{G}}_i)$, which means $\tilde{\boldsymbol{J}}_i' = \tilde{\boldsymbol{J}}_i$ and $\tilde{\boldsymbol{I}}_i' = \tilde{\boldsymbol{I}}_i$. □

Corollary 1. *Suppose a set of matrices $\{\tilde{\boldsymbol{G}}_i, \tilde{\boldsymbol{J}}_i, \tilde{\boldsymbol{I}}_i, i = c, p\}$ that meets all the conditions in Proposition 2. For all $i \in \{c, p\}$, if $\tilde{\boldsymbol{G}}_i$ is column-centered by mean within each column (feature), and samples in $\tilde{\boldsymbol{G}}_i$ are assumed to be independent and identically distributed, then $\hat{cov}(\tilde{\boldsymbol{G}}_i) = \hat{cov}(\tilde{\boldsymbol{J}}_i) + \hat{cov}(\tilde{\boldsymbol{I}}_i)$ with respect to the sample covariance matrix between features.*

Proof. : Based on the proof of Proposition 2, we know $\tilde{\boldsymbol{I}}_i = \mathbf{P}_{\Phi_i^c}\tilde{\boldsymbol{G}}_i$.
Since $col(\tilde{\boldsymbol{I}}_i) \subseteq \Phi_i^c$ and $\Phi_i^c \subseteq col(\tilde{\boldsymbol{G}}_i)$, then $col(\tilde{\boldsymbol{I}}_i) \subseteq col(\tilde{\boldsymbol{G}}_i)$.
Since the linear combination of centered columns is still column-centered, $\tilde{\boldsymbol{I}}_i$ is column-centered. Then, $\tilde{\boldsymbol{J}}_i = \tilde{\boldsymbol{G}}_i - \tilde{\boldsymbol{I}}_i$ is also column-centered.
Define $\hat{cov}(A)^{mn}$ and $\hat{cov}(A^m, A^n)$ as the sample covariance between m^{th} feature and n^{th} feature in matrix A. The covariance equation below is always satisfied:

$$[\hat{cov}(\tilde{\boldsymbol{J}}_i + \tilde{\boldsymbol{I}}_i)]^{mn} = [\hat{cov}(\tilde{\boldsymbol{J}}_i)]^{mn} + [\hat{cov}(\tilde{\boldsymbol{I}}_i)]^{mn} + \hat{cov}(\tilde{\boldsymbol{J}}_i^m, \tilde{\boldsymbol{I}}_i^n) + \hat{cov}(\tilde{\boldsymbol{J}}_i^n, \tilde{\boldsymbol{I}}_i^m). \tag{A.15}$$

Based on Condition 4 in Proposition 2, and the column-centered $\tilde{\boldsymbol{J}}_i$ and $\tilde{\boldsymbol{I}}_i$, we have

$$\hat{cov}(\tilde{\boldsymbol{J}}_i^m, \tilde{\boldsymbol{I}}_i^n) = \hat{cov}(\tilde{\boldsymbol{J}}_i^n, \tilde{\boldsymbol{I}}_i^m) = 0 \tag{A.16}$$

Hence, $\hat{cov}(\tilde{\boldsymbol{G}}_i) = \hat{cov}(\tilde{\boldsymbol{J}}_i) + \hat{cov}(\tilde{\boldsymbol{I}}_i)$. □

A.3 Additional information of simulation studies

A.3.1 Data generating process

To construct the SVD components of $\boldsymbol{J}_c$, $\boldsymbol{J}_p$, $\boldsymbol{I}_c$, and $\boldsymbol{I}_p$ in Model 2.7 and 2.8, we first generate elements of $\boldsymbol{U}_{J_c}$ and $\boldsymbol{U}_{J_p}$ from $N(0, 4)$, and elements of $\boldsymbol{V}_J$ and $\boldsymbol{V}_J$ from $N(0, 1)$. We subsequently center and orthonormalized $\boldsymbol{U}_{J_c}$, $\boldsymbol{U}_{J_p}$, $\boldsymbol{V}_J$ and $\boldsymbol{V}_J$. Then, we set $\boldsymbol{\Sigma}_J = diag(40, \ldots, 10)$, where the 40 diagonal elements in $\boldsymbol{\Sigma}_J$ are equally spaced. Next, we generate elements of $\boldsymbol{U}_{I_c}$, $\boldsymbol{U}_{I_p}$, $\boldsymbol{V}_{I_c}$ and $\boldsymbol{V}_{I_p}$ from $N(0, 1)$, and subsequently center and orthonormalized them, where $row(\boldsymbol{V}_{I_c})$ and $row(\boldsymbol{V}_{I_p})$ are constrained to be orthogonal with each other as well as with $row(\boldsymbol{V}_J)$. After that, we set $\boldsymbol{\Sigma}_{I_c} = a \times diag(30, \ldots, 10)$ and $\boldsymbol{\Sigma}_{I_p} = b \times diag(50, \ldots, 10)$, where the 30 diagonal elements in $\boldsymbol{\Sigma}_{I_c}$ and 50 diagonal elements in $\boldsymbol{\Sigma}_{I_p}$ are equally spaced. The values of a and b can be easily calculated to meet different levels of joint proportion. Constructing in this way, we have $\|\boldsymbol{J}_c + \boldsymbol{I}_c\|_F^2 = \|\boldsymbol{J}_p + \boldsymbol{I}_p\|_F^2$, which avoids the issue of "the bigger dataset always wins" during data integration. Last, all these SVD components are scaled down by $\|\boldsymbol{J}_c + \boldsymbol{I}_c\|_F^2$.

A.3.2 Sensitivity analysis of rank selection for iPCR

To better understand the results of sensitivity analysis in Figure A.1, first we note that there are three types of underlying data structures in our simulated genomic data of cell lines and patients, including the joint variation, the model-specific variation, and the residual noise. The joint rank r_J corresponds to r_J latent joint principal components in the joint variation. The individual ranks r_{Ic} and r_{Ip} correspond to r_{Ic} and r_{Ip} latent individual principal components in the cell line-specific and patient-specific data, respectively. A value of joint proportion that is close to 1 (0) illustrates that each joint principal component explains a higher (lower) proportion of variation in the data than each individual principal component does. In addition, a SNR_{stage1} that is much less (higher) than 1 demonstrates that each of those top principal components from the noise could explain higher (lower) or comparable proportion of variation in the data than some joint and individual principal components do. We call those top principal components in the matrix of noise as residual principal components.

The sensitivity of rank selection to prediction accuracy is affected by the proportions of variation explained by joint, individual and residual principal components. Let $\{r_J, r_{I_c}, r_{I_p}\}$

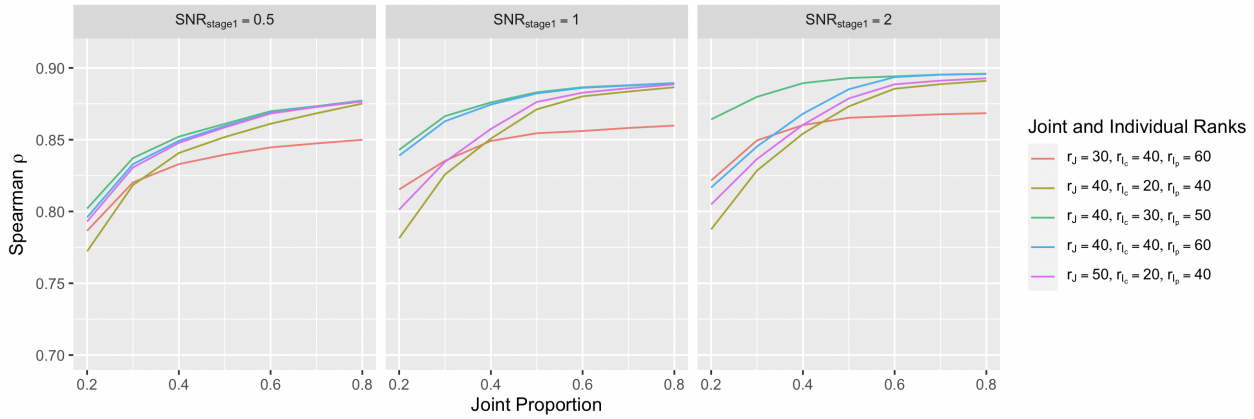


Figure A.1: **Sensitivity analysis of rank selection for iPCR.** Accuracy comparison of iPCR under different rank combinations in predicting $\mathbf{D_p}$ in simulation study. The prediction performance is measured by Spearman correlation between the true and predicted values of $\mathbf{D_p}$. The data are generated under different levels of SNR_{stage1} and joint proportion over 50 replications. The solid curve corresponds to the sample mean of Spearman correlation over replications.

be the form of a rank set. First, when joint proportion is larger than 0.6, the three subplots in Figure A.1 show that the sensitivity of rank selection to the prediction accuracy is relative low except $\{30, 40, 60\}$. This is expected since in this case the top principal components in the data correspond to the joint principal components. Hence, almost all joint principal components that relate to the drug response are easy to be detected, even though the joint rank is slightly overestimated or the individual ranks are slightly misspecified. The underestimation of joint rank could omit important predictors in the drug response prediction model, and significantly reduce the prediction accuracy. Second, when $SNR_{stage1} = 0.5$, the sensitivity of rank selection to the prediction accuracy is relative low except $\{30, 40, 60\}$ under high joint proportion. One possible reason is that since the noise is large and some residual principal components explain comparable proportion of variation in the data with the joint and individual principal components, only part of the joint principal components could be accurately detected even under the true ranks. Hence, slightly misspecified the ranks do no have large effect on the prediction accuracy. Third, when SNR_{stage1} increases from 1 to 2, the prediction accuracy of iPCR given $\{40, 40, 60\}$ (overestimation of individual ranks) decreases under low joint proportion. One possible reason is that when $SNR_{stage1} = 1$, the top residual principal components explain comparable proportion of variation in the data with those bottom individual principal components; hence the overestimation of individual ranks treats those residual principal components as individual principal components, and it does not affect the model estimation too much to detect the joint principal components. In con-

trast, when $SNR_{stage1} = 2$, the variation in the noise is relatively low, and the top residual principal components could not be appropriately treated as individual principal components during the model estimation; hence, the overestimation of individual ranks significantly affect the prediction accuracy in this case.

A.4 Additional results of pan-cancer application

A.4.1 Patient drug response by iPCR

Table A.2: **Rank selection results of iPCR models for drug response prediction in patients.**

Drug	Joint rank	Individual rank of cell line	Individual rank of patient
Trastuzumab (Afatinib)	55	9	29
Etoposide	55	9	29
Vinorelbine	55	9	29
Cisplatin	55	9	29
Oxaliplatin	55	9	29
Gemcitabine	55	9	29
Temozolomide	45	19	39
Pemetrexed	55	9	29
Paclitaxel	45	19	39

Note: The drug name in parenthesis corresponds to the matching drug from GDSC. For each drug, we report the joint and individual ranks selected by iPCR model through Algorithm 1 among the rank set candidates mentioned in Section 2.4.2

A.4.2 Gene co-expression network estimation

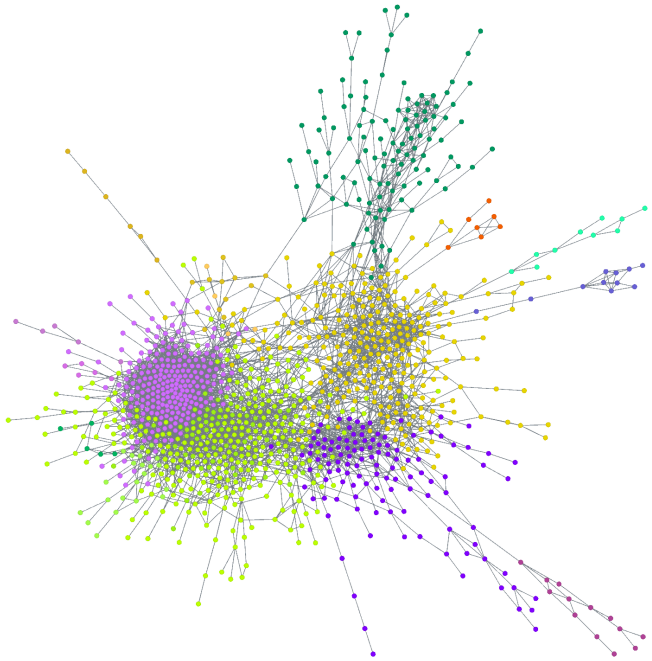


Figure A.2: **Cell line-specific gene co-expression network.** Each node represents a gene, and those significantly associated genes are connected by edges. The gene communities in different colors are detected by an unsupervised clustering method.

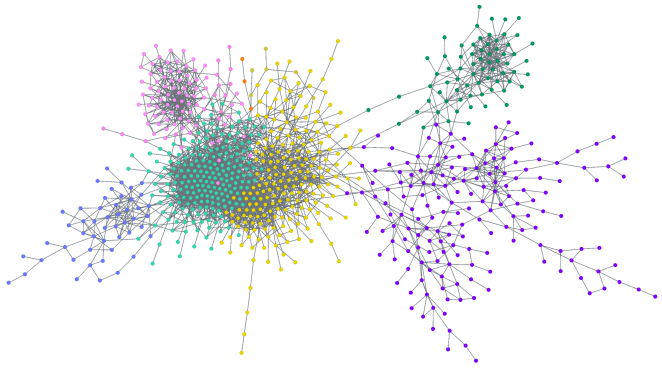


Figure A.3: **Patient-specific gene co-expression network.** Each node represents a gene, and those significantly associated genes are connected by edges. The gene communities in different colors are detected by an unsupervised clustering method.

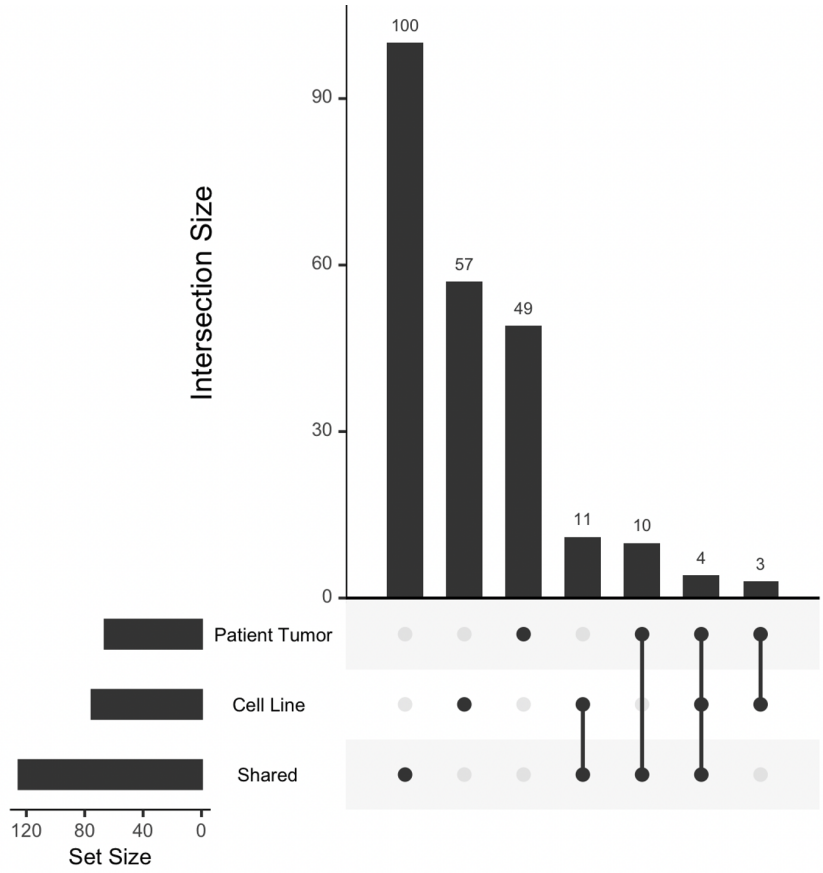


Figure A.4: **Number of hub genes.** The upset plot lists the number of hub genes (those genes that significantly connect to >30 other genes) in each type of networks and their intersections.

Table A.3: **List of hub genes.** Hub genes in shared, cell line-specific, and patient-specific gene co-expression networks.

Network Type	Hub Genes
Shared	RRM1, MCM6, RFC5, SPC25, EXO1, CDC7, MCM4, GTSE1, KIF11, RRM2, RAD54L, RAD51, RFC1, CDC6, CCNA2, MCM3, CDCA8, PLK4, CHEK1, BUB1, TOP2A, PLK1, AURKB, FOXM1, CENPA, ESPL1, KIF2C, KIF15, RFC4, CLSPN, RFC3, FANCD2, TYMS, BIRC5, CDK1, PCNA, NDC80, SIN3A, E2F8, XPO1, TPX2, GMNN, MASTL, MCM8, ESCO2, FANCG, BRCA1, TTK, SUV39H2, ERCC6L, MSH2, MAD2L1, ARID1A, SPC24, MCM2, EWSR1, USP39, EP300, MELK, MCM7, EZH2, NUF2, TOPBP1, CSNK1G1, FANCA, KDM5A, E2F1, BRIP1, RBL1, ERCC3, CDC25A, ROCK1, SETD1A, RBBP5, SETD5, DBF4, RANBP2, SETD2, SUV39H1, RFC2, NEK2, NUP98, BLM, CENPE, MCM5, DDX23, CREBBP, SMC1A, ERCC4, ITGB1, ATRX, BRD4, INCENP, TAOK1, TLK1, LATS1, PTPN11, CREB1, BARD1, SMG1, CCNE2, CCNB1, DYRK1A, RMI1, PIAS4, NSD1, TRIM33, FZR1, PRPF4B, PIKFYVE, SETD1B, TAOK2, BUB3, CDC25C, PKMYT1, EHMT1, VRK1, CDC20, XRCC3, STMN1, RAF1, ACVR1, TRIM28, HIPK1, PBK.
Cell Line	NEK4, PPP5C, IL18BP, TSSK4, SIN3B, ELK1, RPTOR, MAP3K10, ASPSCR1, AKT2, MTOR, TSSK6, XRCC1, AXIN1, CDK10, SPHK2, UPF1, CIC, BRD2, FBXL14, TOP3B, RNF220, TRRAP, NPHP4, PI4KA, DOT1L, NUMA1, MLST8, MAPK8IP3, SMYD4, SMARCD1, PRPF4B, STK11, CSNK1D, INVS, DDX23, ANAPC2, PIAS4, TRIM28, BRD4, IP6K1, ANAPC7, CAD, SETD1A, SMG1, MAP2K7, TSC2, ZBTB17, CTBP1, CREB1, NUP98, CDC73, BUB3, INPPL1, ULK1, BCR, PIK3R2, DVL1, XRCC3, CSNK1G2, EHMT1, CREBBP, PPP2R1B, SETD5, RAD9A, RPS6KB2, TP53BP1, FZR1, MAD2L2, KDM5C, DGKZ, SIK3, USP24, ASXL1, MEN1.
Patient Tumor	PKN2, TAF1, TRRAP, KDM4C, STAG1, MAP4K5, RASA2, DYRK1A, NSD1, TNKS2, ASH1L, SCYL2, WRN, PBRM1, TOPORS, ATF2, STAG2, VHL, SOCS4, MAPK8, PPP2R5E, ITCH, TRPM7, USP24, MAP3K2, WNK1, CBL, KDM3B, SETD2, PRPF4B, USP9X, KDM6A, ARID1A, CREB1, PDPK1, ATM, APPL1, CREBBP, KDM5C, TSC1, APAF1, LATS1, SKP2, E2F7, ARIH2, SOS2, TRAF6, ATRX, BRAF, HIPK1, SETD5, OXSR1, SMC1A, SMAD4, PHIP, RICTOR, SIN3A, USP33, PIKFYVE, CUL2, PIK3CA, MAP3K7, UHRF2, CSNK1G3, NFATC3, SENP2.

APPENDIX B

Appendix Chapter 3

In this appendix, Section B.1 details the estimation algorithm, Section B.2 reports additional results of simulation studies, and Section B.3 provides additional results from the analysis of the PrBP data.

B.1 Estimation algorithm

In this section, we present the details of the estimation algorithm for the spline coefficients. Before stating the algorithm, we first introduce the gradient and Hessian formulas, and then outline the estimation procedure for the spline coefficients $\boldsymbol{\theta}$ in Algorithm 4.

Notation. For individual i at time t_{ij}, let

$$\mathbf{A}_{i,j} \equiv \boldsymbol{X}_i^\top \otimes \boldsymbol{b}(t_{ij}) \in \mathbb{R}^{qp}, \qquad \eta_{ij} \equiv \mathbf{A}_{i,j}^\top \boldsymbol{\theta}, \qquad \rho_{ij} \equiv \tanh(\eta_{ij}),$$

$$u_{ij} \equiv \frac{Y_{i1}(t_{ij})}{\hat{\sigma}_1}, \qquad v_{ij} \equiv \frac{Y_{i2}(t_{ij})}{\hat{\sigma}_2}.$$

Let the block-diagonal penalty matrix be

$$\boldsymbol{\Lambda} \;\equiv\; \text{blockdiag}\big(\lambda_1 \mathbf{D}_2^\top \mathbf{D}_2, \ldots, \lambda_p \mathbf{D}_2^\top \mathbf{D}_2\big) \in \mathbb{R}^{qp \times qp}.$$

Gradient.

$$\begin{aligned} \boldsymbol{\nabla}_{\boldsymbol{\theta}}\, l_{\text{penalized}}(\boldsymbol{\theta}) &= \frac{\partial \log L(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{A})}{\partial \boldsymbol{\theta}} \;-\; \sum_{k=1}^{p} \lambda_k\, \mathbf{D}_2^\top \mathbf{D}_2\, \boldsymbol{\theta}_k \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m_i} \left[\rho_{ij} + u_{ij} v_{ij} - \rho_{ij}\, \frac{u_{ij}^2 - 2\rho_{ij} u_{ij} v_{ij} + v_{ij}^2}{1 - \rho_{ij}^2} \right] \mathbf{A}_{i,j} \;-\; \boldsymbol{\Lambda}\, \boldsymbol{\theta}. \end{aligned} \tag{B.1}$$

Hessian.

$$\mathbf{H}_{\text{pen}}(\boldsymbol{\theta}) = \frac{\partial^2 \log L(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{A})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top} \; - \; \sum_{k=1}^{p} \lambda_k \, \mathbf{D}_2^\top \mathbf{D}_2$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m_i} \left[(1 - \rho_{ij}^2) - u_{ij}^2 - v_{ij}^2 - \frac{2\rho_{ij}^2}{1 - \rho_{ij}^2} \, (u_{ij} - v_{ij})^2 \right] \mathbf{A}_{i,j} \mathbf{A}_{i,j}^\top \; - \; \boldsymbol{\Lambda}. \tag{B.2}$$

Algorithm 4 Estimation of spline coefficients $\boldsymbol{\theta}$

Input: Observed data $\mathbf{Y}$, design matrix $\mathbf{A}$, initial guess $\boldsymbol{\theta}^{(0)}$, penalty parameters $\lambda_k > 0$ for $k = 1, \ldots, p$, second-order difference matrix $\mathbf{D}_2$, and tolerance $\epsilon > 0$.

Initialize: $t \leftarrow 0$.

Repeat:

1. Compute the gradient via Equation (B.1).
2. Compute the Hessian via Equation (B.2).
3. Update coefficients:

$$\boldsymbol{\theta}^{(t+1)} \leftarrow \boldsymbol{\theta}^{(t)} - \mathbf{H}_{\text{pen}}\big\{\boldsymbol{\theta}^{(t)}\big\}^{-1} \boldsymbol{\nabla}_{\boldsymbol{\theta}} \, l_{\text{penalized}}\big\{\boldsymbol{\theta}^{(t)}\big\}.$$

4. Increment $t \leftarrow t + 1$.

Until: $\big\|\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}\big\| < \epsilon$.

Output: Estimated spline coefficients $\boldsymbol{\theta}$.

B.2 Additional simulation studies

This section presents additional simulation results evaluating the performance of TiVAC and competing methods across a range of settings.

Table B.1: **Additional simulation results of CoCoA for Scenario 1.** Simulation results comparing the performance of CoCoA-MLE, CoCoA-REML, and CoCoA-GEE2 methods under different coefficient function types and number of time points per individual. The equation of correlation is assumed to be deterministic and dependent on a binary correlation function. The time-varying correlation is $\rho_0(t)$ when $X = 0$ and $\rho_1(t)$ when $X = 1$. The table presents the RMSE with SD for each method.

Coefficient Functions	# of Time Points	Correlation Functions	RMSE (SD)		
			CoCoA-MLE	CoCoA-REML	CoCoA-GEE2
Linear	$\mathscr{T}_{Low}$	$\rho_0(t)$	0.34 (0.04)	0.36 (0.03)	0.24 (0.04)
		$\rho_1(t)$	0.36 (0.04)	0.34 (0.04)	0.25 (0.04)
	$\mathscr{T}_{Moderate}$	$\rho_0(t)$	0.05 (0.01)	0.06 (0.02)	0.04 (0.02)
		$\rho_1(t)$	0.05 (0.02)	0.07 (0.02)	0.05 (0.02)
	$\mathscr{T}_{High}$	$\rho_0(t)$	0.02 (0.01)	0.03 (0.01)	0.02 (0.01)
		$\rho_1(t)$	0.02 (0.01)	0.02 (0.01)	0.02 (0.01)
	$\mathscr{T}_{PrBP}$	$\rho_0(t)$	0.08 (0.02)	0.09 (0.02)	0.07 (0.01)
		$\rho_1(t)$	0.08 (0.02)	0.10 (0.02)	0.06 (0.02)
Seasonal	$\mathscr{T}_{Low}$	$\rho_0(t)$	0.09 (0.03)	0.08 (0.03)	0.12 (0.04)
		$\rho_1(t)$	0.23 (0.03)	0.22 (0.03)	0.22 (0.04)
	$\mathscr{T}_{Moderate}$	$\rho_0(t)$	0.07 (0.02)	0.05 (0.02)	0.07 (0.02)
		$\rho_1(t)$	0.07 (0.01)	0.07 (0.01)	0.07 (0.02)
	$\mathscr{T}_{High}$	$\rho_0(t)$	0.04 (0.02)	0.04 (0.01)	0.04 (0.02)
		$\rho_1(t)$	0.03 (0.01)	0.04 (0.01)	0.04 (0.01)
	$\mathscr{T}_{PrBP}$	$\rho_0(t)$	0.05 (0.02)	0.05 (0.02)	0.05 (0.02)
		$\rho_1(t)$	0.08 (0.02)	0.08 (0.02)	0.08 (0.02)
Logistic	$\mathscr{T}_{Low}$	$\rho_0(t)$	0.33 (0.04)	0.36 (0.04)	0.24 (0.04)
		$\rho_1(t)$	0.36 (0.04)	0.34 (0.04)	0.25 (0.04)
	$\mathscr{T}_{Moderate}$	$\rho_0(t)$	0.06 (0.02)	0.07 (0.02)	0.05 (0.02)
		$\rho_1(t)$	0.06 (0.02)	0.07 (0.02)	0.05 (0.01)
	$\mathscr{T}_{High}$	$\rho_0(t)$	0.03 (0.01)	0.03 (0.01)	0.03 (0.01)
		$\rho_1(t)$	0.03 (0.01)	0.03 (0.01)	0.03 (0.01)
	$\mathscr{T}_{PrBP}$	$\rho_0(t)$	0.09 (0.02)	0.10 (0.02)	0.08 (0.02)
		$\rho_1(t)$	0.08 (0.02)	0.09 (0.02)	0.07 (0.02)

Table B.2: **Additional simulation results of CoCoA for Scenario 2.** Simulation results comparing the performance of CoCoA-MLE, CoCoA-REML, and CoCoA-GEE2 methods under different coefficient function types and number of time points per individual. The equation of correlation is assumed to be deterministic and dependent on a continuous covariate. The table presents the RMSE with SD for each method.

Coefficient Functions	# of Time Points	RMSE (SD)		
		CoCoA-MLE	CoCoA-REML	CoCoA-GEE2
Linear	$\mathscr{T}_{Low}$	0.13 (0.02)	0.15 (0.02)	0.21 (0.02)
	$\mathscr{T}_{Moderate}$	0.05 (0.01)	0.06 (0.01)	0.10 (0.02)
	$\mathscr{T}_{High}$	0.03 (0.01)	0.03 (0.01)	0.04 (0.01)
	$\mathscr{T}_{PrBP}$	0.08 (0.01)	0.09 (0.01)	0.12 (0.02)
Seasonal	$\mathscr{T}_{Low}$	0.16 (0.02)	0.12 (0.02)	0.25 (0.02)
	$\mathscr{T}_{Moderate}$	0.09 (0.01)	0.07 (0.01)	0.13 (0.02)
	$\mathscr{T}_{High}$	0.04 (0.01)	0.04 (0.01)	0.06 (0.01)
	$\mathscr{T}_{PrBP}$	0.07 (0.01)	0.07 (0.01)	0.10 (0.01)
Logistic	$\mathscr{T}_{Low}$	0.13 (0.02)	0.16 (0.02)	0.22 (0.02)
	$\mathscr{T}_{Moderate}$	0.06 (0.01)	0.06 (0.01)	0.10 (0.02)
	$\mathscr{T}_{High}$	0.03 (0.01)	0.03 (0.01)	0.05 (0.01)
	$\mathscr{T}_{PrBP}$	0.08 (0.01)	0.09 (0.01)	0.11 (0.02)

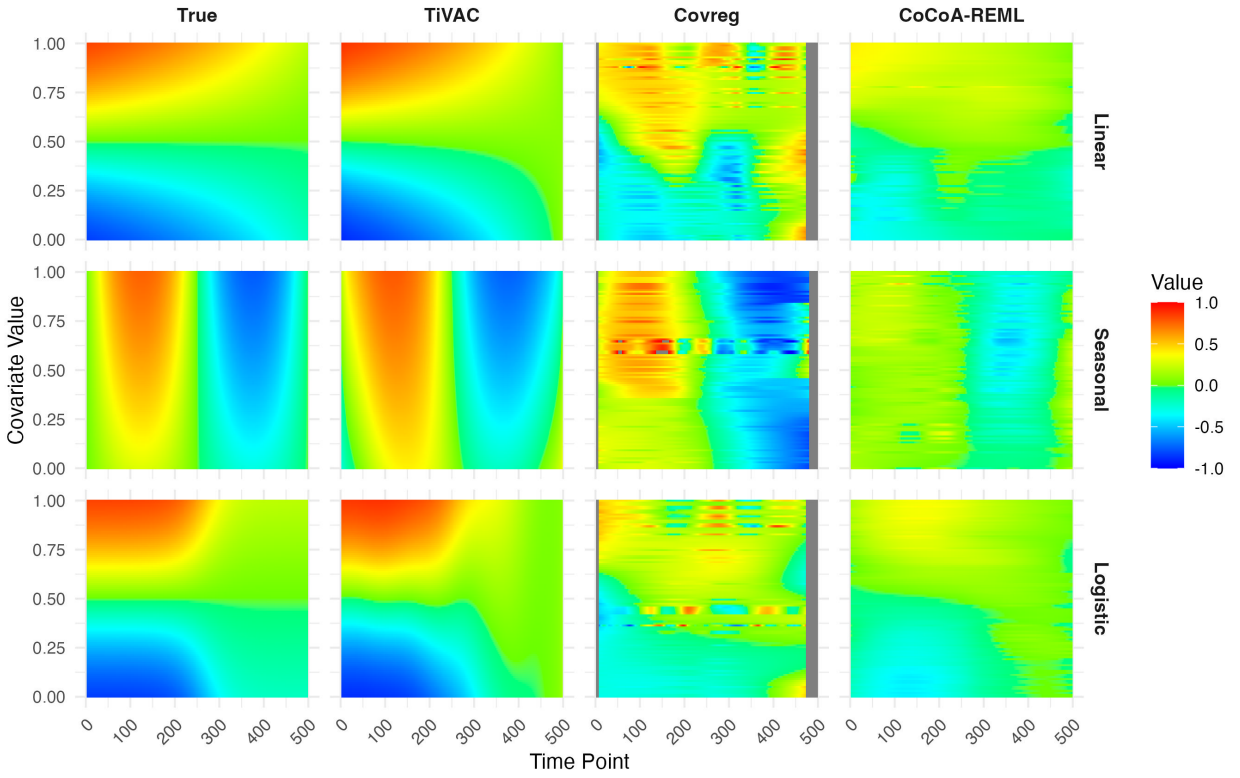


Figure B.1: **Representative heatmap estimations under $\mathscr{T}_{Low}$ setting in Scenario 2.** Comparison between the estimated correlation heatmaps from three methods (TiVAC, CovReg and CoCoA-REML) and the true correlation heatmap under $\mathscr{T}_{Low}$ setting. In each heatmap, correlation values are color-coded, with red indicating higher correlations and blue indicating lower correlations. Each heatmap displays correlation patterns over time for different values of the continuous covariate. The figure presents three independent cases, each representing one of the three types of coefficient functions. Each case shows a representative example from the first run of the simulation (using a fixed random seed for all cases).

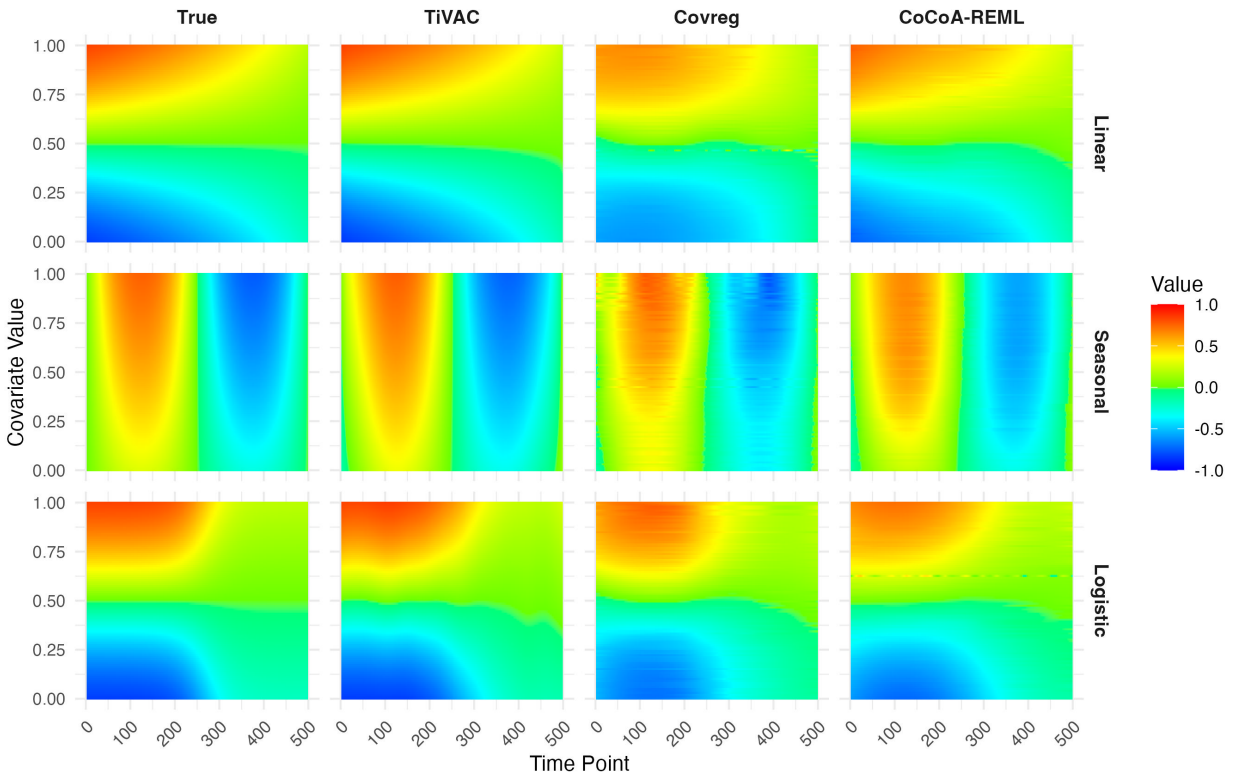


Figure B.2: **Representative heatmap estimations under $\mathscr{T}_{High}$ setting in Scenario 2.** Comparison between the estimated correlation heatmaps from three methods (TiVAC, CovReg and CoCoA-REML) and the true correlation heatmap under $\mathscr{T}_{High}$ setting. In each heatmap, correlation values are color-coded, with red indicating higher correlations and blue indicating lower correlations. Each heatmap displays correlation patterns over time for different values of the continuous covariate. The figure presents three independent cases, each representing one of the three types of coefficient functions. Each case shows a representative example from the first run of the simulation (using a fixed random seed for all cases).

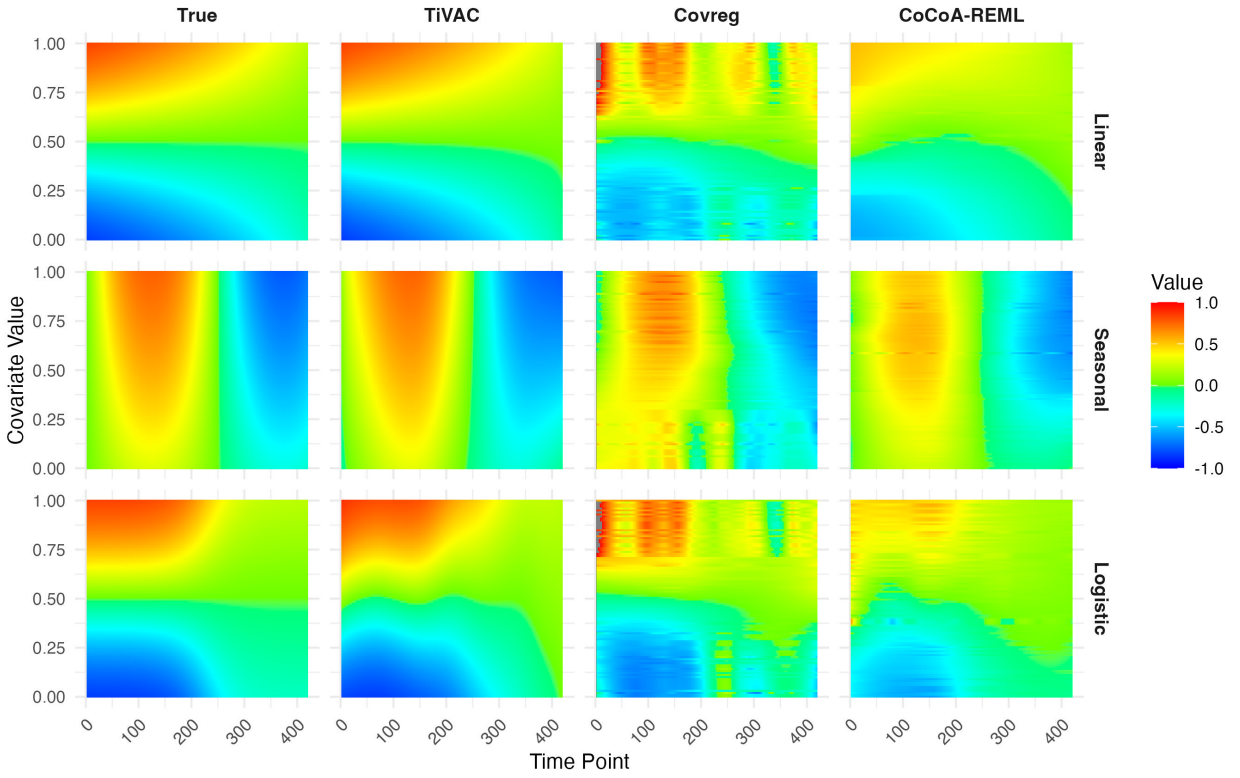


Figure B.3: **Representative heatmap estimations under $\mathscr{T}_{PrBP}$ setting in Scenario 2.** Comparison between the estimated correlation heatmaps from three methods (TiVAC, CovReg and CoCoA-REML) and the true correlation heatmap under $\mathscr{T}_{PrBP}$ setting. In each heatmap, correlation values are color-coded, with red indicating higher correlations and blue indicating lower correlations. Each heatmap displays correlation patterns over time for different values of the continuous covariate. The figure presents three independent cases, each representing one of the three types of coefficient functions. Each case shows a representative example from the first run of the simulation (using a fixed random seed for all cases).

Table B.3: **Simulation results for Scenario 3.** Simulation results comparing the performance of TiVAC, Empirical, CovReg, and CoCoA-REML methods across different coefficient function types under $\mathscr{T}_{Moderate}$ setting. The correlation equation is assumed to be influenced by varying levels of noise and dependent on a binary covariate. The time-varying correlation is $\rho_0(t)$ when $X = 0$ and $\rho_1(t)$ when $X = 1$. The table presents the RMSE with SD for each method.

Coefficient Functions	# of Time Points	σ_{noise}	Correlation Functions	RMSE (SD)			
				TiVAC	Empirical	CovReg	CoCoA-REML
Linear	$\mathscr{T}_{Moderate}$	0.1	$\rho_0(t)$	0.02 (0.01)	0.06 (0.01)	0.04 (0.02)	0.07 (0.02)
			$\rho_1(t)$	0.01 (0.01)	0.06 (0.02)	0.04 (0.02)	0.07 (0.02)
		0.2	$\rho_0(t)$	0.02 (0.01)	0.07 (0.02)	0.04 (0.02)	0.08 (0.02)
			$\rho_1(t)$	0.02 (0.01)	0.07 (0.02)	0.05 (0.02)	0.08 (0.02)
		0.3	$\rho_0(t)$	0.03 (0.01)	0.08 (0.02)	0.05 (0.02)	0.09 (0.01)
			$\rho_1(t)$	0.03 (0.01)	0.08 (0.02)	0.05 (0.02)	0.09 (0.02)
		0.4	$\rho_0(t)$	0.05 (0.01)	0.09 (0.02)	0.07 (0.02)	0.10 (0.02)
			$\rho_1(t)$	0.05 (0.02)	0.10 (0.02)	0.07 (0.02)	0.10 (0.02)
		0.5	$\rho_0(t)$	0.07 (0.01)	0.12 (0.02)	0.09 (0.02)	0.13 (0.02)
			$\rho_1(t)$	0.07 (0.01)	0.11 (0.02)	0.09 (0.02)	0.12 (0.02)
Seasonal	$\mathscr{T}_{Moderate}$	0.1	$\rho_0(t)$	0.04 (0.01)	0.06 (0.02)	0.07 (0.02)	0.06 (0.02)
			$\rho_1(t)$	0.03 (0.01)	0.07 (0.02)	0.06 (0.01)	0.07 (0.02)
		0.2	$\rho_0(t)$	0.04 (0.01)	0.06 (0.02)	0.08 (0.02)	0.06 (0.02)
			$\rho_1(t)$	0.04 (0.01)	0.08 (0.02)	0.07 (0.01)	0.08 (0.02)
		0.3	$\rho_0(t)$	0.04 (0.01)	0.06 (0.02)	0.07 (0.02)	0.06 (0.02)
			$\rho_1(t)$	0.05 (0.01)	0.09 (0.02)	0.08 (0.02)	0.09 (0.02)
		0.4	$\rho_0(t)$	0.05 (0.01)	0.07 (0.02)	0.08 (0.02)	0.07 (0.02)
			$\rho_1(t)$	0.06 (0.01)	0.10 (0.02)	0.09 (0.02)	0.10 (0.02)
		0.5	$\rho_0(t)$	0.05 (0.01)	0.07 (0.02)	0.09 (0.02)	0.08 (0.02)
			$\rho_1(t)$	0.08 (0.01)	0.12 (0.02)	0.11 (0.03)	0.12 (0.02)
Logistic	$\mathscr{T}_{Moderate}$	0.1	$\rho_0(t)$	0.03 (0.01)	0.06 (0.01)	0.05 (0.02)	0.07 (0.02)
			$\rho_1(t)$	0.03 (0.01)	0.06 (0.01)	0.05 (0.02)	0.07 (0.02)
		0.2	$\rho_0(t)$	0.04 (0.01)	0.07 (0.01)	0.05 (0.02)	0.08 (0.02)
			$\rho_1(t)$	0.03 (0.01)	0.07 (0.02)	0.05 (0.01)	0.08 (0.01)
		0.3	$\rho_0(t)$	0.04 (0.01)	0.07 (0.01)	0.05 (0.02)	0.09 (0.02)
			$\rho_1(t)$	0.04 (0.01)	0.08 (0.02)	0.06 (0.02)	0.08 (0.02)
		0.4	$\rho_0(t)$	0.05 (0.01)	0.09 (0.02)	0.07 (0.02)	0.10 (0.02)
			$\rho_1(t)$	0.05 (0.01)	0.09 (0.02)	0.07 (0.02)	0.10 (0.02)
		0.5	$\rho_0(t)$	0.07 (0.01)	0.10 (0.02)	0.08 (0.02)	0.11 (0.02)
			$\rho_1(t)$	0.07 (0.01)	0.11 (0.02)	0.08 (0.02)	0.11 (0.02)

Table B.4: **Additional simulation results of CoCoA for Scenario 3.** Simulation results comparing the performance of CoCoA-MLE, CoCoA-REML, and CoCoA-GEE2 methods across different coefficient function types under $\mathscr{T}_{Moderate}$ setting. The correlation equation is assumed to be influenced by varying levels of noise and dependent on a binary covariate. The time-varying correlation is $\rho_0(t)$ when $X = 0$ and $\rho_1(t)$ when $X = 1$. The table presents the RMSE with SD for each method.

Coefficient Functions	# of Time Points	σ_{noise}	Correlation Functions	RMSE (SD)		
				CoCoA-MLE	CoCoA-REML	CoCoA-GEE2
Linear	$\mathscr{T}_{Moderate}$	0.1	$\rho_0(t)$	0.06 (0.02)	0.07 (0.02)	0.04 (0.01)
			$\rho_1(t)$	0.06 (0.02)	0.07 (0.02)	0.05 (0.02)
		0.2	$\rho_0(t)$	0.06 (0.02)	0.08 (0.02)	0.05 (0.02)
			$\rho_1(t)$	0.06 (0.02)	0.08 (0.02)	0.05 (0.02)
		0.3	$\rho_0(t)$	0.07 (0.01)	0.09 (0.01)	0.06 (0.02)
			$\rho_1(t)$	0.07 (0.02)	0.09 (0.02)	0.06 (0.02)
		0.4	$\rho_0(t)$	0.09 (0.02)	0.10 (0.02)	0.07 (0.02)
			$\rho_1(t)$	0.09 (0.02)	0.10 (0.02)	0.08 (0.02)
		0.5	$\rho_0(t)$	0.11 (0.02)	0.13 (0.02)	0.10 (0.02)
			$\rho_1(t)$	0.10 (0.02)	0.12 (0.02)	0.09 (0.02)
Seasonal	$\mathscr{T}_{Moderate}$	0.1	$\rho_0(t)$	0.07 (0.02)	0.06 (0.02)	0.07 (0.02)
			$\rho_1(t)$	0.07 (0.02)	0.07 (0.02)	0.07 (0.02)
		0.2	$\rho_0(t)$	0.08 (0.02)	0.06 (0.02)	0.07 (0.02)
			$\rho_1(t)$	0.07 (0.02)	0.08 (0.02)	0.08 (0.01)
		0.3	$\rho_0(t)$	0.07 (0.02)	0.06 (0.02)	0.07 (0.02)
			$\rho_1(t)$	0.08 (0.02)	0.09 (0.02)	0.09 (0.02)
		0.4	$\rho_0(t)$	0.08 (0.02)	0.07 (0.02)	0.08 (0.02)
			$\rho_1(t)$	0.10 (0.02)	0.10 (0.02)	0.10 (0.02)
		0.5	$\rho_0(t)$	0.09 (0.02)	0.08 (0.02)	0.09 (0.02)
			$\rho_1(t)$	0.12 (0.03)	0.12 (0.02)	0.12 (0.02)
Logistic	$\mathscr{T}_{Moderate}$	0.1	$\rho_0(t)$	0.06 (0.02)	0.07 (0.02)	0.05 (0.01)
			$\rho_1(t)$	0.06 (0.02)	0.07 (0.02)	0.05 (0.02)
		0.2	$\rho_0(t)$	0.06 (0.02)	0.08 (0.02)	0.06 (0.02)
			$\rho_1(t)$	0.07 (0.01)	0.08 (0.01)	0.05 (0.02)
		0.3	$\rho_0(t)$	0.07 (0.02)	0.09 (0.02)	0.06 (0.02)
			$\rho_1(t)$	0.07 (0.02)	0.08 (0.02)	0.06 (0.02)
		0.4	$\rho_0(t)$	0.09 (0.02)	0.10 (0.02)	0.07 (0.02)
			$\rho_1(t)$	0.08 (0.02)	0.10 (0.02)	0.07 (0.02)
		0.5	$\rho_0(t)$	0.10 (0.02)	0.11 (0.02)	0.09 (0.02)
			$\rho_1(t)$	0.10 (0.02)	0.11 (0.02)	0.09 (0.02)

Table B.5: **Simulation results for Scenario 4.** Simulation results comparing the performance of TiVAC, CovReg, and CoCoA-REML methods across different coefficient function types under $\mathscr{T}_{Moderate}$ setting. The correlation equation is assumed to be influenced by varying levels of noise and dependent on a continuous covariate. The table presents the RMSE with SD for each method.

Coefficient Functions	# of Time Points	σ_{noise}	RMSE (SD)		
			TiVAC	CovReg	CoCoA-REML
Linear	$\mathscr{T}_{Moderate}$	0.1	0.02 (0.01)	0.09 (0.01)	0.06 (0.01)
		0.2	0.02 (0.01)	0.09 (0.02)	0.07 (0.01)
		0.3	0.03 (0.01)	0.09 (0.01)	0.08 (0.02)
		0.4	0.04 (0.01)	0.10 (0.01)	0.08 (0.02)
		0.5	0.05 (0.01)	0.11 (0.02)	0.09 (0.02)
Seasonal	$\mathscr{T}_{Moderate}$	0.1	0.03 (0.01)	0.09 (0.02)	0.07 (0.01)
		0.2	0.04 (0.01)	0.10 (0.01)	0.08 (0.01)
		0.3	0.05 (0.01)	0.11 (0.02)	0.09 (0.01)
		0.4	0.05 (0.01)	0.12 (0.01)	0.10 (0.01)
		0.5	0.07 (0.01)	0.13 (0.01)	0.11 (0.01)
Logistic	$\mathscr{T}_{Moderate}$	0.1	0.04 (0.01)	0.09 (0.01)	0.07 (0.01)
		0.2	0.04 (0.01)	0.10 (0.01)	0.07 (0.01)
		0.3	0.04 (0.01)	0.10 (0.01)	0.07 (0.01)
		0.4	0.05 (0.01)	0.10 (0.01)	0.08 (0.01)
		0.5	0.06 (0.01)	0.11 (0.02)	0.09 (0.01)

Table B.6: **Additional simulation results of CoCoA for Scenario 4.** Simulation results comparing the performance of CoCoA-MLE, CoCoA-REML, and CoCoA-GEE2 methods across different coefficient function types under $\mathscr{T}_{Moderate}$ setting. The correlation equation is assumed to be influenced by varying levels of noise and dependent on a continuous covariate. The table presents the RMSE with SD for each method.

Coefficient Functions	# of Time Points	σ_{noise}	RMSE (SD)		
			CoCoA-MLE	CoCoA-REML	CoCoA-GEE2
Linear	$\mathscr{T}_{Moderate}$	0.1	0.05 (0.01)	0.06 (0.01)	0.10 (0.02)
		0.2	0.06 (0.01)	0.07 (0.01)	0.10 (0.02)
		0.3	0.07 (0.01)	0.08 (0.02)	0.11 (0.02)
		0.4	0.07 (0.02)	0.08 (0.02)	0.12 (0.02)
		0.5	0.08 (0.01)	0.09 (0.02)	0.13 (0.02)
Seasonal	$\mathscr{T}_{Moderate}$	0.1	0.09 (0.01)	0.07 (0.01)	0.13 (0.02)
		0.2	0.09 (0.01)	0.08 (0.01)	0.13 (0.01)
		0.3	0.10 (0.01)	0.09 (0.01)	0.15 (0.02)
		0.4	0.11 (0.01)	0.10 (0.01)	0.16 (0.02)
		0.5	0.12 (0.01)	0.11 (0.01)	0.17 (0.01)
Logistic	$\mathscr{T}_{Moderate}$	0.1	0.06 (0.01)	0.07 (0.01)	0.11 (0.01)
		0.2	0.06 (0.01)	0.07 (0.01)	0.11 (0.02)
		0.3	0.06 (0.01)	0.07 (0.01)	0.12 (0.02)
		0.4	0.07 (0.01)	0.08 (0.01)	0.13 (0.02)
		0.5	0.07 (0.01)	0.09 (0.01)	0.14 (0.02)

B.3 Additional application results

This section reports the empirical analysis and additional results of TiVAC model for the PHQ–GAD correlation based on PrBP data.

Table B.7: **Descriptive statistics of the preprocessed PrBP dataset** The cohort is consist of 291 BP1 patients. The table summarizes gender, substance use disorders, and medication-related measures.

Variable	Category	Overall (N=291)
Gender	Female	206 (70.8%)
	Male	85 (29.2%)
Alcohol Use Disorder	Yes	138 (47.4%)
	No	153 (52.6%)
Cannabis Use Disorder	Yes	84 (28.9%)
	No	207 (71.1%)
Average Nervous System-Related Medications	Mean (SD)	1.25 (1.10)
Antipsychotic Use Frequency	Mean (SD)	0.137 (0.230)
Antidepressant Use Frequency	Mean (SD)	0.209 (0.303)
Mood Stabilizer Use Frequency	Mean (SD)	0.433 (0.403)

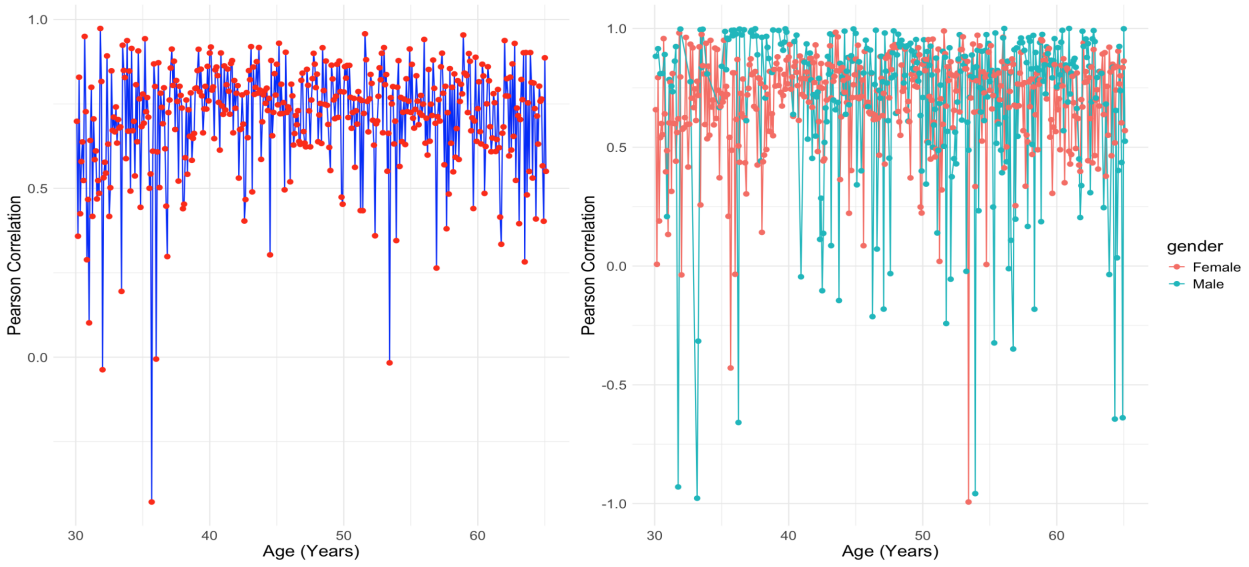


Figure B.4: **Empirical analysis on PrBP data.** Left: Pearson correlation between PHQ and GAD across age for PrBP data (overall). Right: Pearson correlation between PHQ and GAD across age stratified by gender. Each point represents the correlation measured at a specific month of age, with the x-axis converting age in months to years.

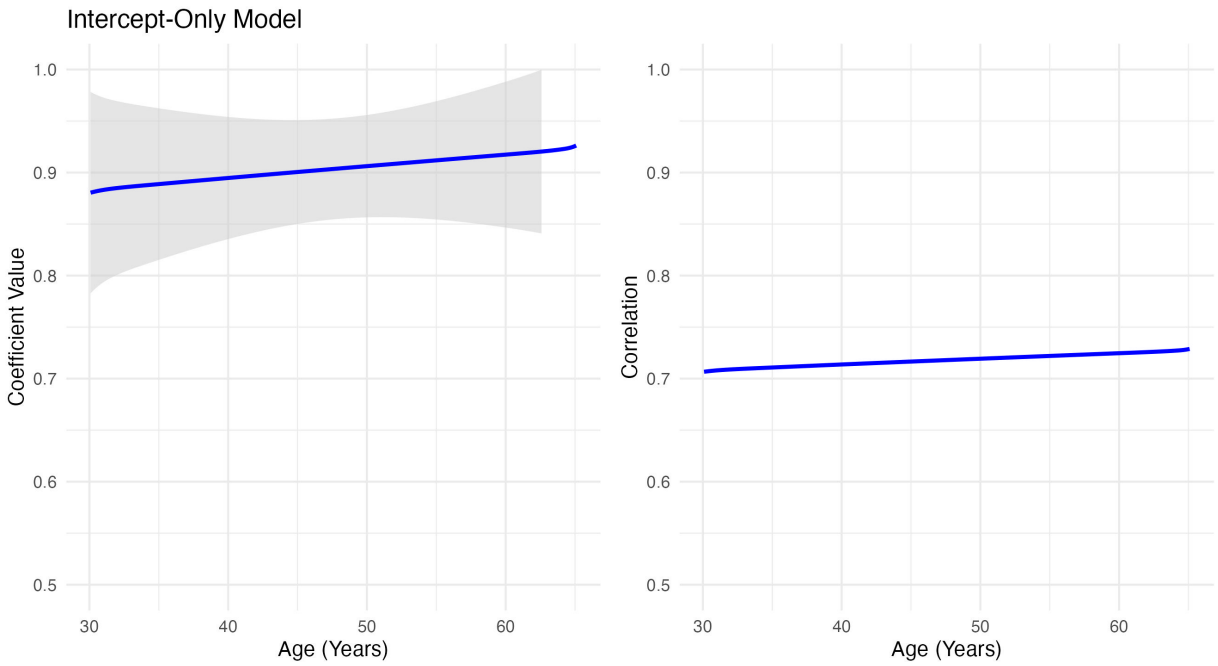


Figure B.5: **TiVAC results of intercept-only model on PrBP data.** The left panel displays the TiVAC coefficient function $\beta(t)$ in blue, with a gray band indicating the 95% simultaneous confidence interval. The right panel illustrates the TiVAC correlation between PHQ-9 and GAD-7 as a function of age, with no additional covariates considered.

APPENDIX C

Appendix Chapter 4

This appendix includes additional figures and tables for the simulation studies and the real-data application in Chapter 4.

C.1 Additional simulation results

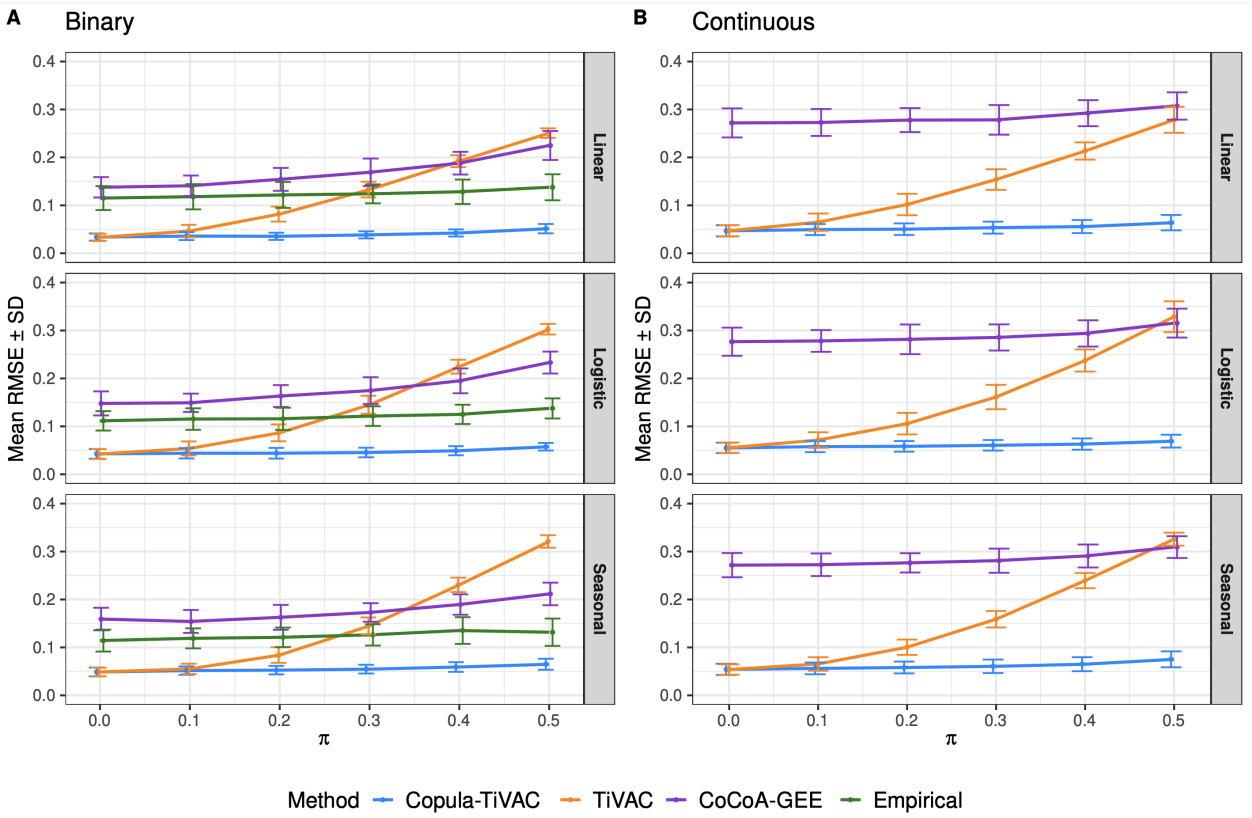


Figure C.1: **Additional simulation RMSE across zero proportions.** Mean RMSE (± SD) of the estimated correlation $\rho(t, x)$ for $\pi \in \{0, 0.1, \ldots, 0.5\}$. Methods: Copula–TiVAC (blue), TiVAC (orange), CoCoA–GEE (purple), Empirical (green). **A**: Binary covariate scenario under $\mathscr{T}_{\text{Sparse}}$ for the three trajectory patterns; **B**: Continuous covariate scenario under $\mathscr{T}_{\text{Sparse}}$ for the three trajectory patterns (Empirical not applicable).

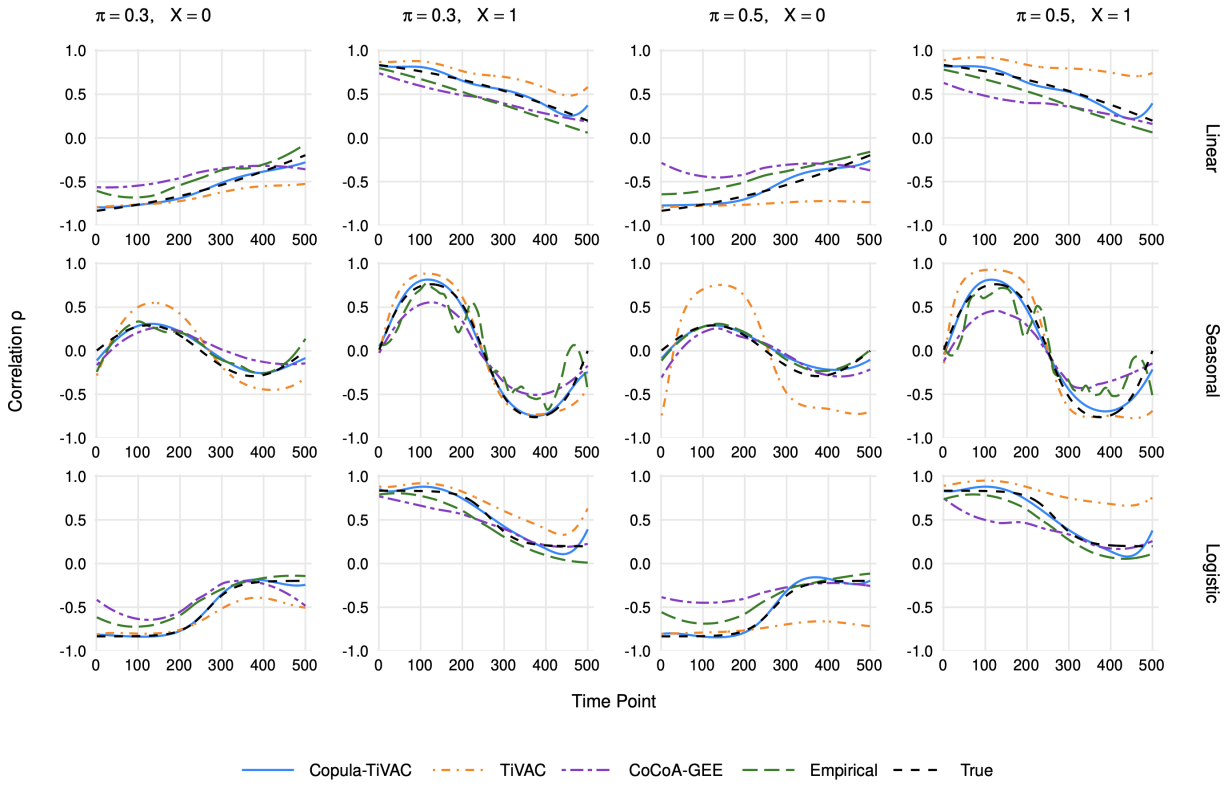


Figure C.2: **Representative correlation curve estimations under $\mathscr{T}_{\text{Sparse}}$.** Comparison between the estimated correlation curves from four methods (Copula-TiVAC, TiVAC, CoCoA-GEE, and Empirical) and the true correlation curve for $X = 0$ and $X = 1$ at $\pi = 0.3$ and $\pi = 0.5$ under $\mathscr{T}_{\text{Sparse}}$ in binary covariate scenario. Note that each subplot corresponds to a representative example from the first of 50 simulation runs (using a fixed random seed).

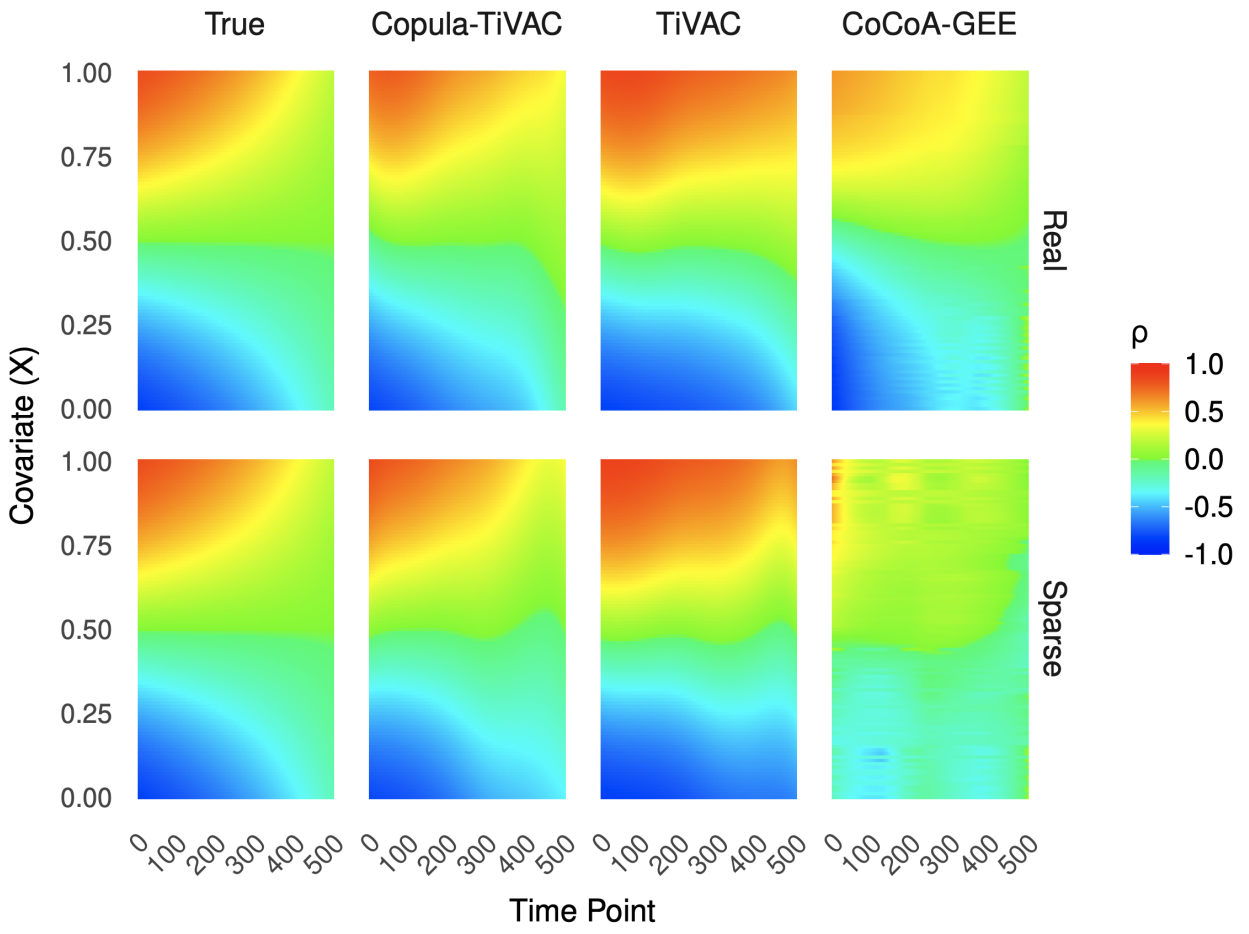


Figure C.3: **Representative heatmap estimations under the multiple–covariate scenario.** Comparison between the true heatmap and estimates from Copula-TiVAC, TiVAC, and CoCoA–GEE under the linear trajectory pattern and $\pi = 0.3$. Each heatmap displays how $\rho(t, x)$ change over time (x-axis) and the continuous covariate X (y-axis); facet rows correspond to $\mathscr{T}_{\text{Real}}$ and $\mathscr{T}_{\text{Sparse}}$, and columns to methods. Correlation is color-coded. Each estimated heatmap corresponds to a representative example from the first of 50 simulation runs (fixed seed).

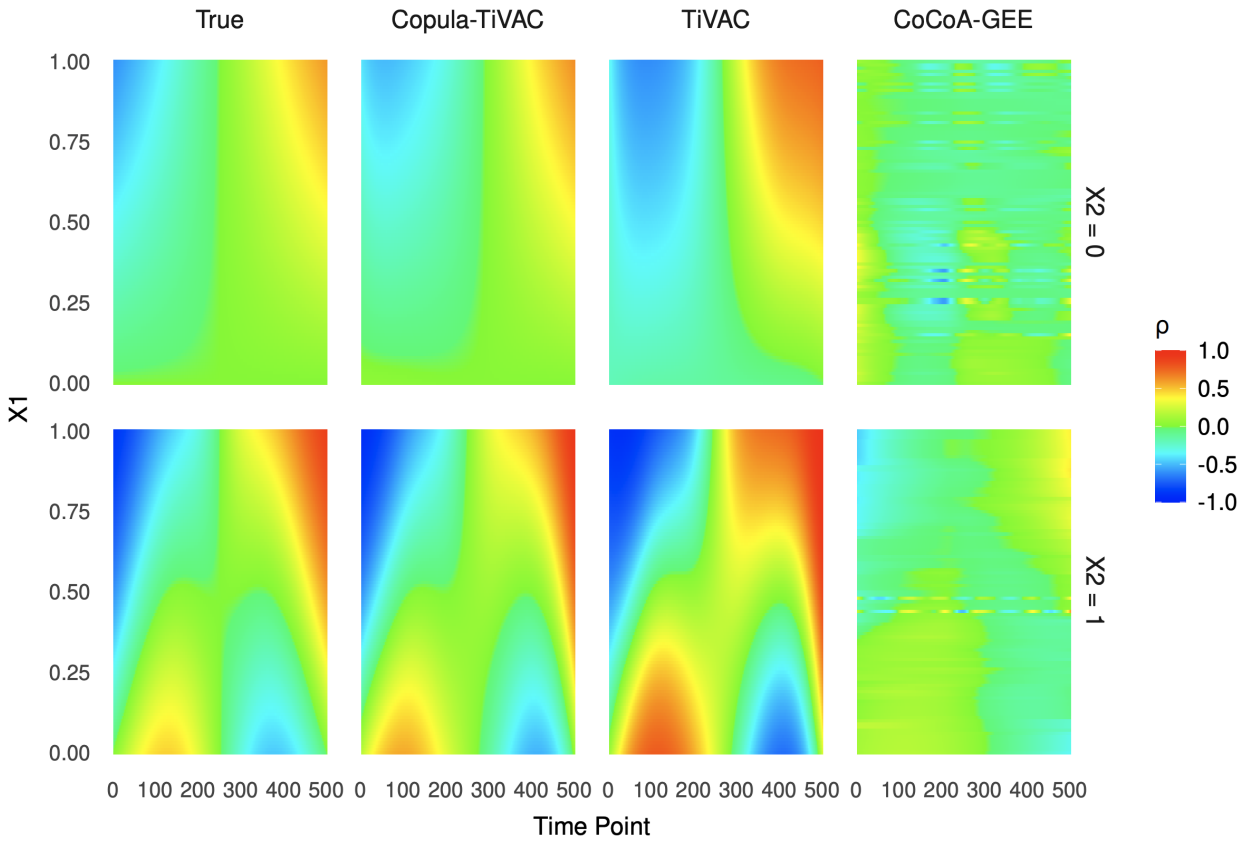


Figure C.4: **Representative heatmap estimations under the multiple–covariate scenario ($\mathscr{T}_{\text{Sparse}}$).** Comparison between the true heatmap and estimates from Copula-TiVAC, TiVAC, and CoCoA–GEE. The heatmaps show how $\rho(t, x_1, x_2)$ change over time (x-axis) and the continuous covariate X_1 (y-axis); facet rows correspond to $X_2 = 0$ and $X_2 = 1$, and columns to methods. The setting includes $\mathscr{T}_{\text{Sparse}}$ and $\pi = 0.3$. Correlation is color-coded. Each estimated heatmap corresponds to a representative example from the first of 50 simulation runs (fixed seed).

C.2 Additional application results

Table C.1: **Descriptive statistics across analysis cohorts.** After preprocessing BP-I participants from PrBP, we defined three largely overlapping analysis cohorts based on concurrent availability of each outcome pair: PHQ–ASRM ($n = 198$), GAD–ASRM ($n = 185$), and PHQ–GAD ($n = 230$). The table reports n (%) for gender and substance use disorders (alcohol, cannabis), and mean (SD) for four medication-related covariates. Values are n (%) or mean (SD).

Variable	**Category**	**PHQ–ASRM**	**GAD–ASRM**	**PHQ–GAD**
Sample Size		198	185	230
Gender	Female	141 (71.2%)	132 (71.4%)	165 (71.7%)
	Male	57 (28.8%)	53 (28.6%)	65 (28.3%)
Alcohol Use Disorder	Abuse/Dependence	89 (44.9%)	84 (45.4%)	107 (46.5%)
	No	109 (55.1%)	101 (54.6%)	123 (53.5%)
Cannabis Use Disorder	Abuse/Dependence	53 (26.8%)	50 (27.0%)	62 (27.0%)
	No	145 (73.2%)	135 (73.0%)	168 (73.0%)
Nervous System-Related Medications	Mean (SD)	1.23 (1.06)	1.21 (1.07)	1.24 (1.12)
Antipsychotic Use Frequency	Mean (SD)	0.14 (0.23)	0.14 (0.23)	0.14 (0.23)
Antidepressant Use Frequency	Mean (SD)	0.20 (0.29)	0.20 (0.29)	0.20 (0.29)
Mood Stabilizer Use Frequency	Mean (SD)	0.42 (0.40)	0.42 (0.40)	0.43 (0.40)

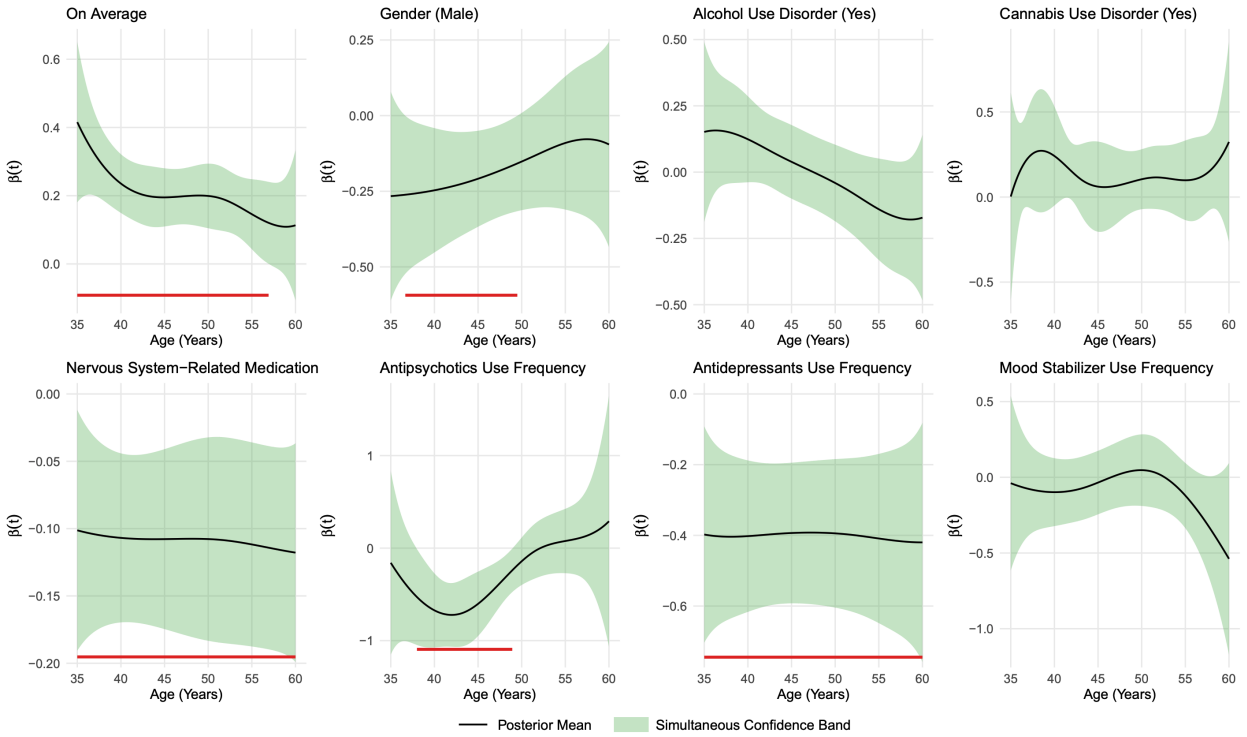


Figure C.5: **Copula-TiVAC coefficient functions for GAD-ASRM correlation.** Posterior mean (black) and 95% simultaneous band (green) for time-varying coefficients $\beta(t)$. The bottom bar (red) highlights the age intervals where the simultaneous confidence band does not include zero. The Intercept panel shows the baseline trajectory without considering covariates. Remaining panels show $\beta_1(t)$ for each covariate: Gender (Male), Alcohol Use Disorder (Yes), Cannabis Use Disorder (Yes), Nervous System–Related Medication, Antipsychotics Use Frequency, Antidepressants Use Frequency, and Mood Stabilizer Use Frequency.

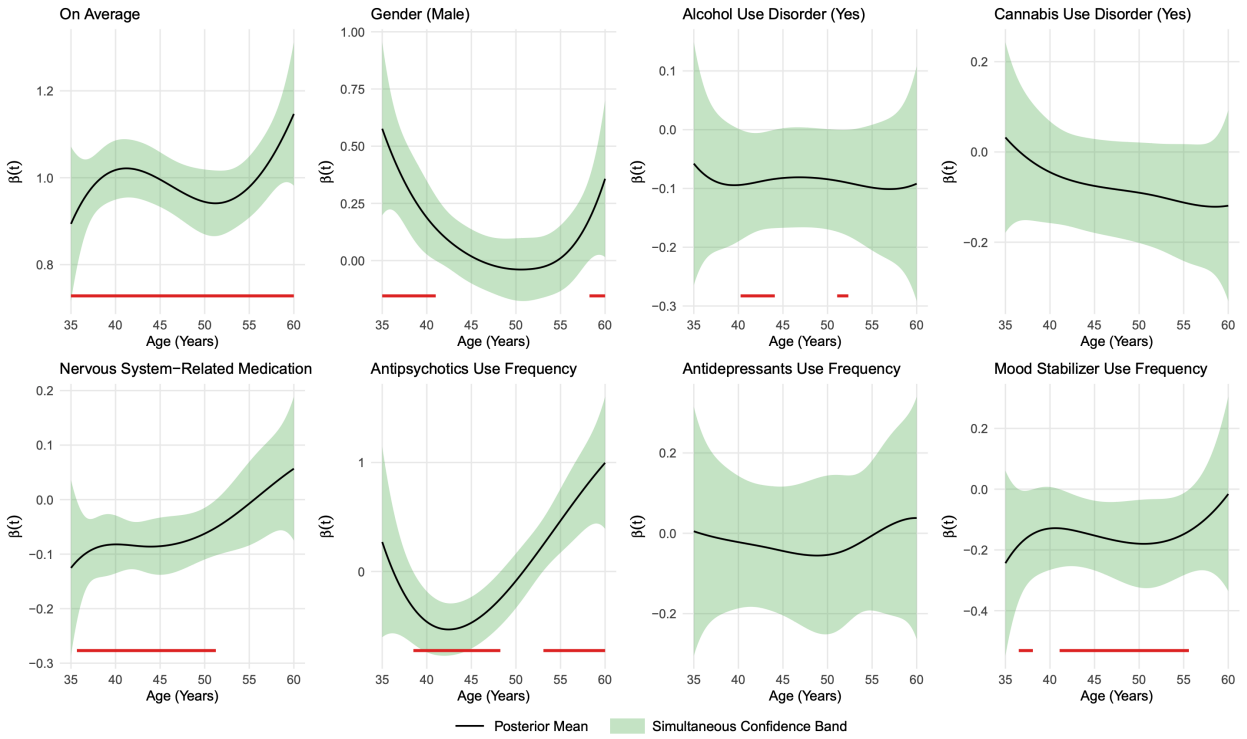


Figure C.6: **Copula-TiVAC coefficient functions for PHQ-GAD correlation.** Posterior mean (black) and 95% simultaneous band (green) for time-varying coefficients $\beta(t)$. The bottom bar (red) highlights the age intervals where the simultaneous confidence band does not include zero. The Intercept panel shows the baseline trajectory without considering covariates. Remaining panels show $\beta_1(t)$ for each covariate: Gender (Male), Alcohol Use Disorder (Yes), Cannabis Use Disorder (Yes), Nervous System–Related Medication, Antipsychotics Use Frequency, Antidepressants Use Frequency, and Mood Stabilizer Use Frequency.

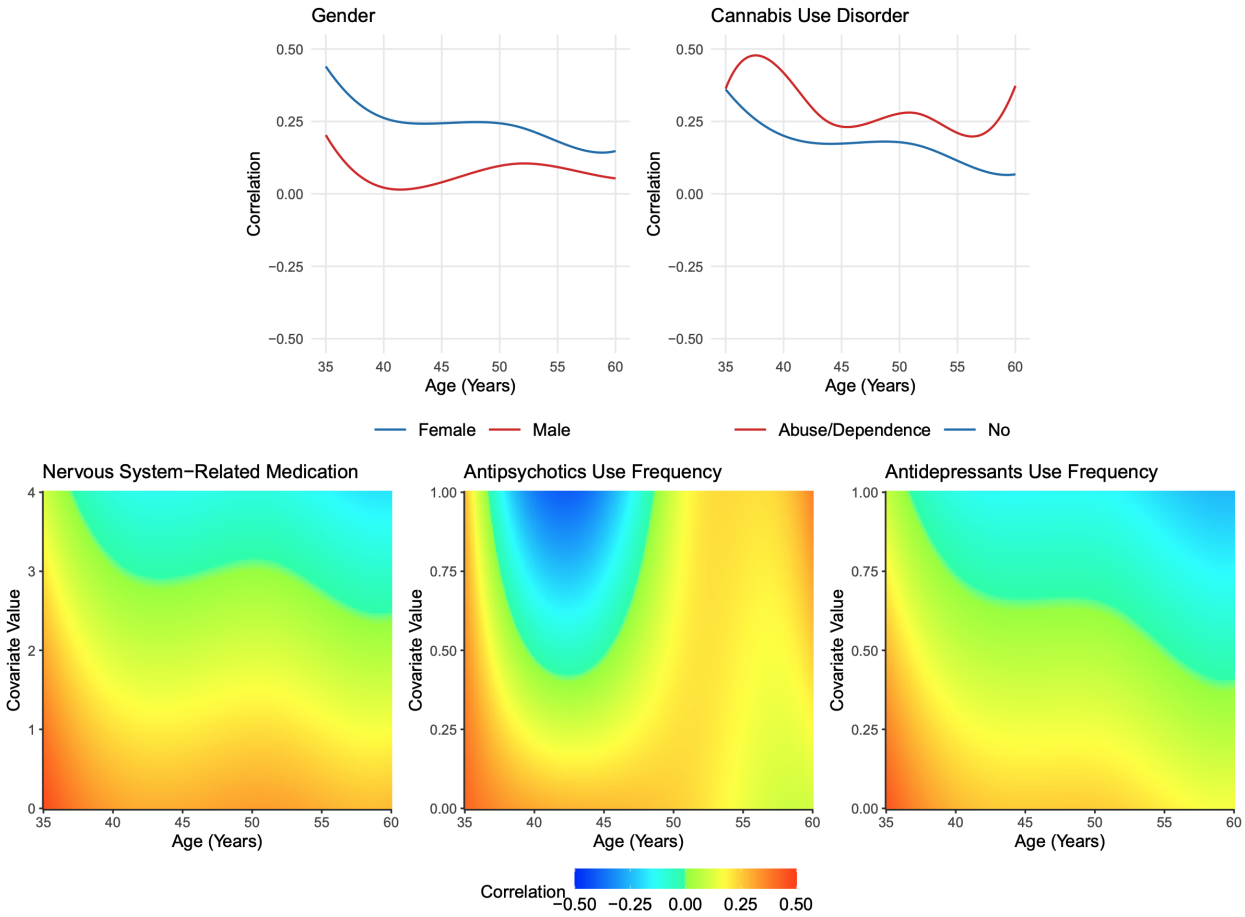


Figure C.7: **Copula-TiVAC correlations between GAD-7 and ASRM.** Estimated latent correlation between GAD-7 and ASRM as a function of age and covariates. Top row shows curves for gender, and cannabis Use Disorder. Bottom row shows heatmaps for continuous covariates (Nervous System-Related Medication, Antipsychotics Use Frequency, Antidepressants Use Frequency). For the heatmaps, age is on the x-axis, covariate value on the y-axis, and the color scale ranges from -0.5 (blue) to 0.5 (red). Positive values indicate concurrent increases in depression and mania; negative values indicate inverse movement.

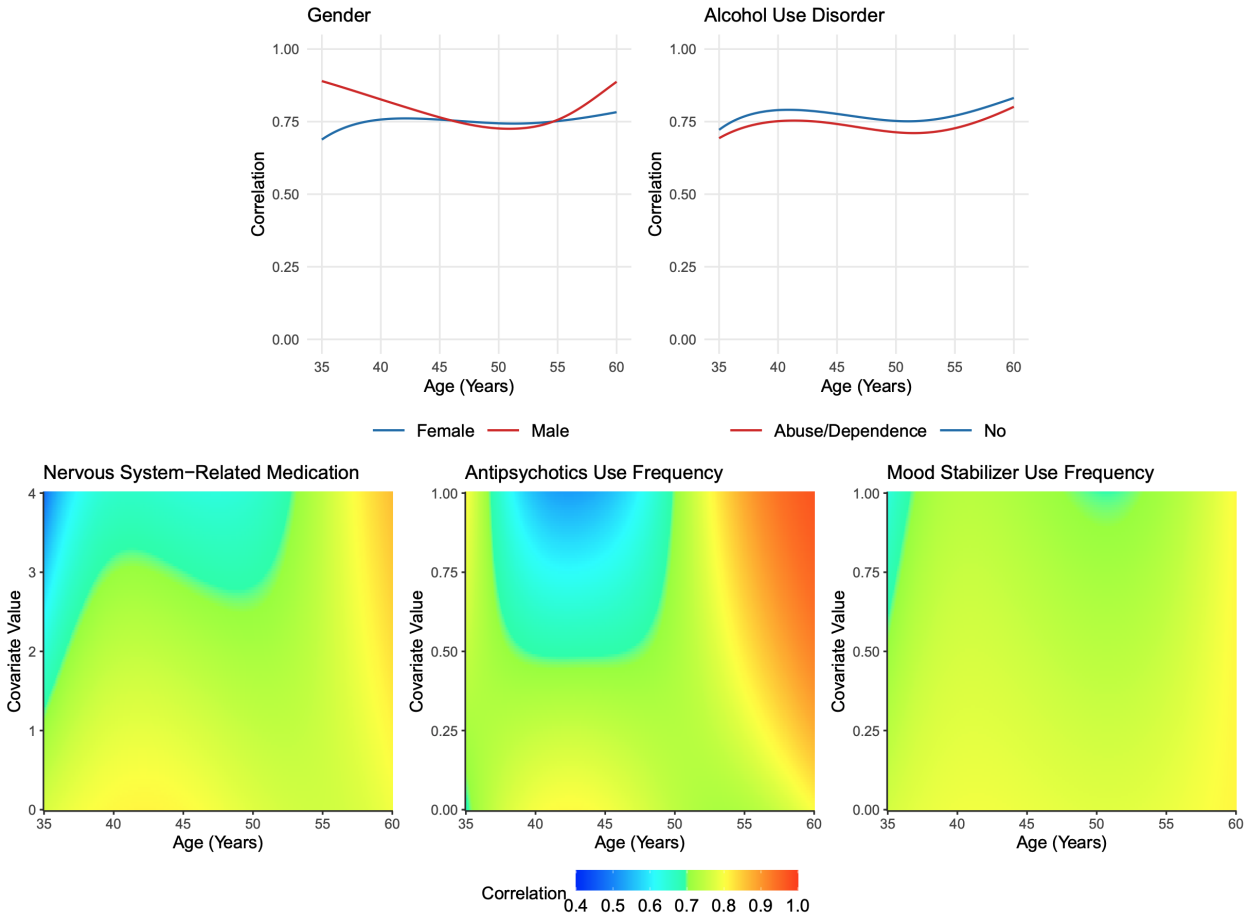


Figure C.8: **Copula-TiVAC correlations between PHQ-9 and GAD-7.** Estimated latent correlation between PHQ-9 and GAD-7 as a function of age and covariates. Top row shows curves for gender, and cannabis Use Disorder. Bottom row shows heatmaps for continuous covariates (Nervous System-Related Medication, Antipsychotics Use Frequency, Antidepressants Use Frequency). For the heatmaps, age is on the x-axis, covariate value on the y-axis, and the color scale ranges from 0.4 (blue) to 1.0 (red). Positive values indicate concurrent increases in depression and mania; negative values indicate inverse movement.

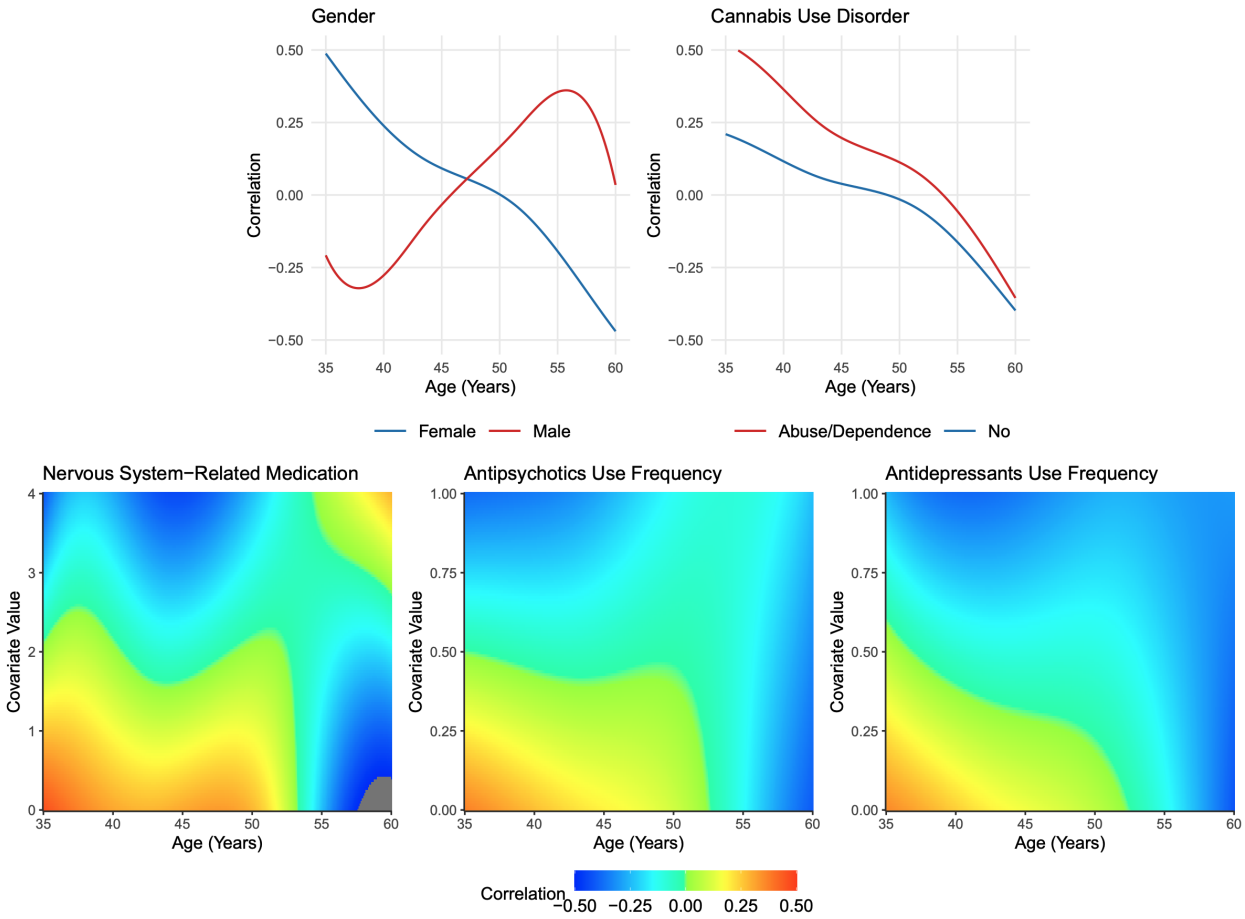


Figure C.9: **TiVAC correlations between PHQ-9 and ASRM.** Estimated latent correlation between PHQ-9 and ASRM as a function of age and covariates. Top row shows curves for gender, and cannabis Use Disorder. Bottom row shows heatmaps for continuous covariates (Nervous System-Related Medication, Antipsychotics Use Frequency, Antidepressants Use Frequency). For the heatmaps, age is on the x-axis, covariate value on the y-axis, and the color scale ranges from -0.5 (blue) to 0.5 (red). Positive values indicate concurrent increases in depression and mania; negative values indicate inverse movement.

BIBLIOGRAPHY

Acharyya, S., Zhou, X., and Baladandayuthapani, V. (2022). Spacex: Gene co-expression network estimation for spatial transcriptomics. *Bioinformatics*, 38(22):5033–5041.

Akhoon, N. (2021). Precision medicine: A new paradigm in therapeutics. *International Journal of Preventive Medicine*, 12(1).

Albihn, A., Lovén, J., Ohlsson, J., Osorio, L. M., and Henriksson, M. (2006). C-myc-dependent etoposide-induced apoptosis involves activation of bax and caspases, and pkcdelta signaling. *Journal of Cellular Biochemistry*, 98(6):1597–1614.

Alchakee, A., Ahmed, M., Eldohaji, L., Alhaj, H., and Saber-Ayad, M. (2022). Pharmacogenomics in psychiatry practice: The value and the challenges. *International Journal of Molecular Sciences*, 23(21):13485.

Altman, E. G., Hedeker, D., Peterson, J. L., and Davis, J. M. (1997). The altman self-rating mania scale. *Biological Psychiatry*, 42(10):948–955.

Anderson, I. M., Haddad, P. M., and Scott, J. (2012). Bipolar disorder. *BMJ*, 345(dec27 3).

Archakov, I. and Hansen, P. R. (2021). A new parametrization of correlation matrices. *Econometrica*, 89(4):1699–1715.

Arjmand, B., Goodarzi, P., Mohamadi-Jahani, F., Falahzadeh, K., and Larijani, B. (2017). Personalized regenerative medicine. *Acta Medica Iranica*, 55(3):144–149.

Asparouhov, T., Hamaker, E. L., and Muthén, B. (2017). Dynamic structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(3):359–388.

Bally, N., Zullino, D., and Aubry, J.-M. (2014). Cannabis use and first manic episode. *Journal of Affective Disorders*, 165:103–108.

Bangasser, D. A. and Cuarenta, A. (2021). Sex differences in anxiety and depression: Circuits and mechanisms. *Nature Reviews Neuroscience*, 22(11):674–684.

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2010). Network medicine: A network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., and et al. (2012). The cancer cell

line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607.

Bartlett, R. F. (1993). Linear modelling of pearson's product moment correlation coefficient: An application of fisher's z-transformation. *The Statistician*, 42(1):45.

Ben-David, U., Ha, G., Tseng, Y.-Y., Greenwald, N. F., Oh, C., Shih, J., McFarland, J. M., Wong, B., Boehm, J. S., Beroukhim, R., and et al. (2017). Patient-derived xenografts undergo mouse-specific tumor evolution. *Nature Genetics*, 49(11):1567–1575.

Ben-David, U., Siranosian, B., Ha, G., Tang, H., Oren, Y., Hinohara, K., Strathdee, C. A., Dempster, J., Lyons, N. J., Burns, R., and et al. (2018). Genetic and transcriptional evolution alters cancer cell line drug response. *Nature*, 560(7718):325–330.

Betancourt, M. (2017). A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*.

Bhattacharyya, R., Ha, M. J., Liu, Q., Akbani, R., Liang, H., and Baladandayuthapani, V. (2020). Personalized network modeling of the pan-cancer patient and cell line interactome. *JCO Clinical Cancer Informatics*, (4):399–411.

Birmaher, B., Axelson, D., Goldstein, B., Strober, M., Gill, M. K., Hunt, J., Houck, P., Ha, W., Iyengar, S., Kim, E., and et al. (2009). Four-year longitudinal course of children and adolescents with bipolar spectrum disorders: The course and outcome of bipolar youth (coby) study. *American Journal of Psychiatry*, 166(7):795–804.

Black, D. W. and Grant, J. E. (2014). *DSM-5 guidebook: The essential companion to the Diagnostic and statistical manual of mental disorders, fifth edition.* American Psychiatric Publishing.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10).

Blyth, C. R. (1972). On simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67(338):364.

Bonnín, C. d., Reinares, M., Martínez-Arán, A., Jiménez, E., Sánchez-Moreno, J., Solé, B., Montejo, L., and Vieta, E. (2019). Improving functioning, quality of life, and well-being in patients with bipolar disorder. *International Journal of Neuropsychopharmacology*.

Brodaczewska, K. K., Szczylik, C., Fiedorowicz, M., Porta, C., and Czarnecka, A. M. (2016). Choosing the right cell line for renal cell cancer research. *Molecular Cancer*, 15(1).

Brubaker, D. K. and Lauffenburger, D. A. (2020). Translating preclinical models to humans. *Science*, 367(6479):742–743.

Cairo, S., Armengol, C., De Reyniès, A., Wei, Y., Thomas, E., Renard, C.-A., Goga, A., Balakrishnan, A., Semeraro, M., Gresh, L., and et al. (2008). Hepatic stem-like phenotype and interplay of wnt/β-catenin and myc signaling in aggressive childhood liver cancer. *Cancer Cell*, 14(6):471–484.

Chae, Y. K., Pan, A. P., Davis, A. A., Patel, S. P., Carneiro, B. A., Kurzrock, R., and Giles, F. J. (2017). Path toward precision oncology: Review of targeted therapy studies and tools to aid in defining "actionability" of a molecular lesion and patient management support. *Molecular Cancer Therapeutics*, 16(12):2645–2655.

Charafe-Jauffret, E., Ginestier, C., Monville, F., Finetti, P., Adélaïde, J., Cervera, N., Fekairi, S., Xerri, L., Jacquemier, J., Birnbaum, D., and et al. (2005). Gene expression profiling of breast cell lines identifies potential new basal markers. *Oncogene*, 25(15):2273–2284.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM.

Chiu, T. Y., Leonard, T., and Tsui, K.-W. (1996). The matrix-logarithmic covariance model. *Journal of the American Statistical Association*, 91(433):198.

Choi, J., Chung, H. C., Gaynanova, I., and Ni, Y. (2024). Bayesian segmented gaussian copula factor model for single-cell sequencing data. *arXiv preprint arXiv:2403.15983*. https://doi.org/10.48550/arXiv.2403.15983.

Cochran, A. L., Schultz, A., McInnis, M. G., and Forger, D. B. (2018). Testing frameworks for personalizing bipolar disorder. *Translational Psychiatry*, 8(1).

Coryell, W., Fiedorowicz, J., Solomon, D., and Endicott, J. (2009). Age transitions in the course of bipolar i disorder. *Psychological Medicine*, 39(8):1247–1252.

Costello, J. C., Heiser, L. M., Georgii, E., Gönen, M., Menden, M. P., Wang, N. J., Bansal, M., Ammad-ud din, M., Hintsanen, P., Khan, S. A., and et al. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology*, 32(12):1202–1212.

Creighton, C. J., Massarweh, S., Huang, S., Tsimelzon, A., Hilsenbeck, S. G., Osborne, C. K., Shou, J., Malorni, L., and Schiff, R. (2008). Development of resistance to targeted therapies transforms the clinically associated molecular profile subtype of breast tumor xenografts. *Cancer Research*, 68(18):7493–7501.

Dadousis, C., Veerkamp, R. F., Heringstad, B., Pszczola, M., and Calus, M. P. (2014). A comparison of principal component regression and genomic reml for genomic prediction across populations. *Genetics Selection Evolution*, 46(1).

de Boor, C. (2001). *A practical guide to splines*. Springer.

Dell'Osso, B., Cremaschi, L., Macellaro, M., and Cafaro, R. (2022). Gender and sex issues in bipolar disorder. *Psychiatric Times*, 39(6). Published online: June 10, 2022.

Denny, J. C. and Collins, F. S. (2021). Precision medicine in 2030—seven ways to transform healthcare. *Cell*, 184(6):1415–1419.

Ding, Z., Zu, S., and Gu, J. (2016). Evaluating the molecule-based prediction of clinical drug responses in cancer. *Bioinformatics*, 32(19):2891–2895.

Dinno, A. (2009). Implementing horn's parallel analysis for principal component analysis and factor analysis. *The Stata Journal: Promoting communications on statistics and Stata*, 9(2):291–298.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, 11(2).

Ely, S. (2009). Personalized medicine: Individualized care of cancer patients. *Translational Research*, 154(6):303–308.

Engle, R. (2002). Dynamic conditional correlation. *Journal of Business & Economic Statistics*, 20(3):339–350.

Falgreen, S., Dybkær, K., Young, K. H., Xu-Monette, Z. Y., El-Galaly, T. C., Laursen, M. B., Bødker, J. S., Kjeldsen, M. K., Schmitz, A., Nyegaard, M., and et al. (2015). Predicting response to multidrug regimens in cancer patients using cell line experiments and regularised regression models. *BMC Cancer*, 15(1).

Feng, Q., Jiang, M., Hannig, J., and Marron, J. (2018). Angle-based joint and individual variation explained. *Journal of Multivariate Analysis*, 166:241–265.

Fernandes, B. S., Williams, L. M., Steiner, J., Leboyer, M., Carvalho, A. F., and Berk, M. (2017). The new field of 'precision psychiatry'. *BMC Medicine*, 15(1).

Fischer, M., Grossmann, P., Padi, M., and DeCaprio, J. A. (2016). Integration of tp53, dream, mmb-foxm1 and rb-e2f target gene analyses identifies cell cycle gene regulatory networks. *Nucleic Acids Research*, 44(13):6070–6086.

Franklin, S. B., Gibson, D. J., Robertson, P. A., Pohlmann, J. T., and Fralish, J. S. (1995). Parallel analysis: A method for determining significant principal components. *Journal of Vegetation Science*, 6(1):99–106.

Gaynanova, I. and Li, G. (2019). Structural learning and integrative decomposition of multi-view data. *Biometrics*, 75(4):1121–1132.

Geeleher, P., Cox, N. J., and Huang, R. S. (2014). Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biology*, 15(3).

Geeleher, P., Zhang, Z., Wang, F., Gruener, R. F., Nath, A., Morrison, G., Bhutra, S., Grossman, R. L., and Huang, R. S. (2017). Discovering novel pharmacogenomic biomarkers by imputing drug response in cancer patients from large genomics studies. *Genome Research*, 27(10):1743–1751.

German National Cohort (GNC) Consortium (2014). The german national cohort: aims, study design and organization. *European Journal of Epidemiology*, 29(5):371–382.

Gibbs, R. A. (2020). The human genome project changed everything. *Nature Reviews Genetics*, 21(10):575–576.

Gijbels, I., Omelka, M., and Veraverbeke, N. (2012). Multivariate and functional covariates and conditional copulas. *Electronic Journal of Statistics*, 6(none).

Gijbels, I., Veraverbeke, N., and Omelka, M. (2011). Conditional copulas, association measures and their applications. *Computational Statistics & Data Analysis*, 55(5):1919–1932.

Gillet, J.-P., Calcagno, A. M., Varma, S., Marino, M., Green, L. J., Vora, M. I., Patel, C., Orina, J. N., Eliseeva, T. A., Singal, V., and et al. (2011). Redefining the relevance of established cancer cell lines to the study of mechanisms of clinical anti-cancer drug resistance. *Proceedings of the National Academy of Sciences*, 108(46):18708–18713.

Gillet, J.-P., Varma, S., and Gottesman, M. M. (2013). The clinical relevance of cancer cell lines. *JNCI Journal of the National Cancer Institute*, 105(7):452–458.

Goellner, E. M., Putnam, C. D., and Kolodner, R. D. (2015). Exonuclease 1-dependent and independent mismatch repair. *DNA Repair*, 32:24–32.

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabási, A.-L. (2007). The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690.

Goldman, M. J., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., Banerjee, A., Luo, Y., Rogers, D., Brooks, A. N., and et al. (2020). Visualizing and interpreting cancer genomics data via the xena platform. *Nature Biotechnology*, 38(6):675–678.

Goodspeed, A., Heiser, L. M., Gray, J. W., and Costello, J. C. (2016). Tumor-derived cell lines as molecular models of cancer pharmacogenomics. *Molecular Cancer Research*, 14(1):3–13.

Grissinger, M. (2010). The five rights: A destination without a map. *P T*, 35(10):542.

Hettema, J. M. (2008). What is the genetic relationship between anxiety and depression? *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, 148C(2):140–146.

Ho, Y.-Y., Parmigiani, G., Louis, T. A., and Cope, L. M. (2010). Modeling liquid association. *Biometrics*, 67(1):133–141.

Hoff, P. D. and Niu, X. (2012). A covariance regression model. *Statistica Sinica*, 22(2).

Hoogstraat, M., de Pagter, M. S., Cirkel, G. A., van Roosmalen, M. J., Harkins, T. T., Duran, K., Kreeftmeijer, J., Renkens, I., Witteveen, P. O., Lee, C. C., and et al. (2013). Genomic and transcriptomic plasticity in treatment-naïve ovarian cancer. *Genome Research*, 24(2):200–211.

Iorio, F., Knijnenburg, T. A., Vis, D. J., Bignell, G. R., Menden, M. P., Schubert, M., Aben, N., Gonçalves, E., Barthorpe, S., Lightfoot, H., and et al. (2016). A landscape of pharmacogenomic interactions in cancer. *Cell*, 166(3):740–754.

Jiang, Y. and Wang, M. (2010). Personalized medicine in oncology: Tailoring the right drug to the right patient. *Biomarkers in Medicine*, 4(4):523–533.

Kalin, N. H. (2020). The critical relationship between anxiety and depression. *American Journal of Psychiatry*, 177(5):365–367.

Kaplan, A. and Lock, E. F. (2017). Prediction with dimension reduction of multiple molecular data sources for patient survival. *Cancer Informatics*, 16:117693511771851.

Kato, T., Ono, H., Fujii, M., Akahoshi, K., Ogura, T., Ogawa, K., Ban, D., Kudo, A., Tanaka, S., Tanabe, M., and et al. (2021). Cytoplasmic rrm1 activation as an acute response to gemcitabine treatment is involved in drug resistance of pancreatic cancer cells. *PLOS ONE*, 16(6).

Kaur, G. and Dufour, J. M. (2012a). Cell lines. *Spermatogenesis*, 2(1):1–5.

Kaur, G. and Dufour, J. M. (2012b). Cell lines. *Spermatogenesis*, 2(1):1–5.

Keijzers, G., Bakula, D., Petr, M., Madsen, N., Teklu, A., Mkrtchyan, G., Osborne, B., and Scheibye-Knudsen, M. (2018). Human exonuclease 1 (exo1) regulatory functions in dna replication with putative roles in cancer. *International Journal of Molecular Sciences*, 20(1):74.

Kim, H., McInnis, M. G., and Sperry, S. H. (2024). Longitudinal dynamics between anxiety and depression in bipolar spectrum disorders. *Journal of Psychopathology and Clinical Science*, 133(2):129–139.

Klaassen, C. A. J. and Wellner, J. A. (1997). Efficient estimation in the bivariate normal copula model: Normal margins are least favourable. *Bernoulli*, 3(1):55–77.

Kornstein, S., Schatzberg, A., Thase, M., Yonkers, K., McCullough, J., Keitner, G., Gelenberg, A., Ryan, C., Hess, A., Harrison, W., and et al. (2000). Gender differences in chronic major and double depression. *Journal of Affective Disorders*, 60(1):1–11.

Kroenke, K., Spitzer, R. L., and Williams, J. B. (2001). The phq-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9):606–613.

Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, 28(5).

Kumar-Sinha, C. and Chinnaiyan, A. M. (2018). Precision oncology in the age of integrative genomics. *Nature Biotechnology*, 36(1):46–60.

Kundu, S., Ming, J., Nocera, J., and McGregor, K. M. (2021). Integrative learning for population of dynamic networks with covariates. *NeuroImage*, 236:118181.

Kurilov, R., Haibe-Kains, B., and Brors, B. (2020). Assessment of modelling strategies for drug response prediction in cell lines and xenografts. *Scientific Reports*, 10(1).

Lang, S. and Brezger, A. (2004). Bayesian p-splines. *Journal of Computational and Graphical Statistics*, 13(1):183–212.

Lee, H., Park, Y. M., and Lee, S. (2015). Principal component regression by principal component selection. *Communications for Statistical Applications and Methods*, 22(2):173–180.

Lenze, E. J. (2003). Comorbidity of depression and anxiety in the elderly. *Current Psychiatry Reports*, 5(1):62–67.

Lewis, J. S., Gross, M. H., Sousa, J., Henrikus, S. S., Greiwe, J. F., Nans, A., Diffley, J. F., and Costa, A. (2022). Mechanism of replication origin melting nucleated by cmg helicase assembly. *Nature*, 606(7916):1007–1014.

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). Upset: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1983–1992.

Li, J., Lu, Y., Akbani, R., Ju, Z., Roebuck, P. L., Liu, W., Yang, J.-Y., Broom, B. M., Verhaak, R. G., Kane, D. W., and et al. (2013). Tcpa: A resource for cancer functional proteomics data. *Nature Methods*, 10(11):1046–1047.

Li, J., Zhao, W., Akbani, R., Liu, W., Ju, Z., Ling, S., Vellano, C. P., Roebuck, P., Yu, Q., Eterovic, A. K., and et al. (2017). Characterization of human cancer cell lines by reverse-phase protein arrays. *Cancer Cell*, 31(2):225–239.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J., and Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. *Cell Systems*, 1(6):417–425.

Liu, H., Lafferty, J., and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10:2295–2328.

Liu, Q., Ha, M. J., Bhattacharyya, R., Garmire, L., and Baladandayuthapani, V. (2019). Network-based matching of patients and targeted therapies for precision oncology. *Biocomputing 2020*.

Liu, Q., Srivastava, P., Rozhkov, I., Farnum, G. A., Holmes, M. J., Fritsche, L. G., Belmonte, D., Ade, A., Sun, J., Yocum, A., McInnis, M., Baladandayuthapani, V., and Athey, B. D. (2025). A linear analysis mixed model phenotyper (lamp) for bipolar disorder: Evaluating the impact of medication, substance use disorders, and comorbidities on mood trajectories. *medRxiv*. Preprint.

Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2013). Joint and individual variation explained (jive) for integrated analysis of multiple data types. *The Annals of Applied Statistics*, 7(1).

Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., and Bergmann, S. (2016). Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nature Methods*, 13(4):366–370.

Marques, L., Costa, B., Pereira, M., Silva, A., Santos, J., Saldanha, L., Silva, I., Magalhães, P., Schmidt, S., and Vale, N. (2024). Advancing precision medicine: A review of innovative in silico approaches for drug development, clinical pharmacology and personalized healthcare. *Pharmaceutics*, 16(3):332.

McCarthy, J. J., McLeod, H. L., and Ginsburg, G. S. (2013). Genomic medicine: A decade of successes, challenges, and opportunities. *Science Translational Medicine*, 5(189).

McDavid, A., Gottardo, R., Simon, N., and Drton, M. (2019). Graphical models for zero-inflated single cell gene expression. *The Annals of Applied Statistics*, 13(2).

McInnis, M. G., Assari, S., Kamali, M., Ryan, K., Langenecker, S. A., Saunders, E. F., Versha, K., Evans, S., O'Shea, K. S., Mower Provost, E., and et al. (2017). Cohort profile: The heinz c. prechter longitudinal study of bipolar disorder. *International Journal of Epidemiology*, 47(1).

Meyer, M. J., Coull, B. A., Versace, F., Cinciripini, P., and Morris, J. S. (2015). Bayesian function-on-function regression for multilevel functional data. *Biometrics*, 71(3):563–574.

Michiels, S., Laplanche, A., Boulet, T., Dessen, P., Guillonneau, B., Méjean, A., Desgrandchamps, F., Lathrop, M., Sarasin, A., Benhamou, S., and et al. (2009). Genetic polymorphisms in 85 dna repair genes and bladder cancer risk. *Carcinogenesis*, 30(5):763–768.

Mourragui, S., Loog, M., van de Wiel, M. A., Reinders, M. J., and Wessels, L. F. (2019). Precise: A domain adaptation approach to transfer predictors of drug response from preclinical models to tumors. *Bioinformatics*, 35(14):i510–i519.

Mourragui, S. M., Loog, M., Vis, D. J., Moore, K., Manjon, A. G., van de Wiel, M. A., Reinders, M. J., and Wessels, L. F. (2021). Predicting patient response with models trained on cell lines and patient-derived xenografts by nonlinear transfer learning. *Proceedings of the National Academy of Sciences*, 118(49).

Muneer, A. (2017). Mixed states in bipolar disorder: Etiology, pathogenesis and treatment. *Chonnam Medical Journal*, 53(1):1.

Nunes, A., Singh, S., Allman, J., Becker, S., Ortiz, A., Trappenberg, T., and Alda, M. (2022). A critical evaluation of dynamical systems models of bipolar disorder. *Translational Psychiatry*, 12(1).

O'Connell, M. J. and Lock, E. F. (2016). R.jive for exploration of multi-source molecular data. *Bioinformatics*, 32(18):2877–2879.

Pan, S. J., Tsang, I. W., Kwok, J. T., and Yang, Q. (2011). Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210.

Park, D. C. and Bischof, G. N. (2013). The aging mind: Neuroplasticity in response to cognitive training. *Dialogues in Clinical Neuroscience*, 15(1):109–119.

Patil, I., Makowski, D., Ben-Shachar, M. S., Wiernik, B. M., Bacher, E., and Lüdecke, D. (2022). Datawizard: An r package for easy data preparation and statistical transformations. *Journal of Open Source Software*, 7(78):4684.

Pato, M. T., Sobell, J. L., Medeiros, H., Abbott, C., Sklar, B. M., Buckley, P. F., Bromet, E. J., Escamilla, M. A., Fanous, A. H., Lehrer, D. S., and et al. (2013). The genomic psychiatry cohort: Partners in discovery. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 162(4):306–312.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86(3):677–690.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2010). *Semiparametric regression*. Cambridge University Press.

Sachs, G. S., Thase, M. E., Otto, M. W., Bauer, M., Miklowitz, D., Wisniewski, S. R., Lavori, P., Lebowitz, B., Rudorfer, M., Frank, E., and et al. (2003). Rationale, design, and methods of the systematic treatment enhancement program for bipolar disorder (step-bd). *Biological Psychiatry*, 53(11):1028–1042.

Saha, A., Ha, M. J., Acharyya, S., and Baladandayuthapani, V. (2022). A bayesian precision medicine framework for calibrating individualized therapeutic indices in cancer. *The Annals of Applied Statistics*, 16(4).

Sakellaropoulos, T., Vougas, K., Narang, S., Koinis, F., Kotsinas, A., Polyzos, A., Moss, T. J., Piha-Paul, S., Zhou, H., Kardala, E., and et al. (2019). A deep learning framework for predicting response to therapy in cancer. *Cell Reports*, 29(11).

Santolini, M. and Barabási, A.-L. (2018). Predicting perturbation patterns from the topology of biological networks. *Proceedings of the National Academy of Sciences*, 115(27).

Schloesser, R. J., Martinowich, K., and Manji, H. K. (2012). Mood-stabilizing drugs: Mechanisms of action. *Trends in Neurosciences*, 35(1):36–46.

Sharifi-Noghabi, H., Peng, S., Zolotareva, O., Collins, C. C., and Ester, M. (2020). Aitl: Adversarial inductive transfer learning with input and output space adaptation for pharmacogenomics. *Bioinformatics*, 36(1):i380–i388.

Sharma, S. V., Haber, D. A., and Settleman, J. (2010). Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. *Nature Reviews Cancer*, 10(4):241–253.

SHER, L. (2008). Manic-depressive illness: Bipolar disorders and recurrent depression, second edition. *American Journal of Psychiatry*, 165(4):541–542.

Shi, D., Mi, G., Wang, M., and Webster, T. J. (2019). In vitro and ex vivo systems at the forefront of infection modeling and drug discovery. *Biomaterials*, 198:228–249.

Sit, D. (2004). Women and bipolar disorder across the life span. *Journal of the American Medical Women's Association (1972)*, 59(2):91–100.

Spitzer, R. L., Kroenke, K., Williams, J. B., and Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder. *Archives of Internal Medicine*, 166(10):1092.

Spoorthy, M. S., Chakrabarti, S., and Grover, S. (2019). Comorbidity of bipolar and anxiety disorders: An overview of trends in research. *World Journal of Psychiatry*, 9(1):7–29.

Su, R., Jin, C., Jin, C., Kuang, M., and Xiang, J. (2021). Identification of hub genes in key hallmarks of ovarian cancer via bioinformatics analysis. *Translational Cancer Research*, 10(2):827–841.

Subbiah, V. (2023). The next generation of evidence-based medicine. *Nature Medicine*, 29(1):49–58.

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., Gould, J., Davis, J. F., Tubelli, A. A., Asiedu, J. K., and et al. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6).

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

Toscano, F., Parmentier, B., Fajoui, Z. E., Estornes, Y., Chayvialle, J.-A., Saurin, J.-C., and Abello, J. (2007). P53 dependent and independent sensitivity to oxaliplatin of colon cancer cells. *Biochemical Pharmacology*, 74(3):392–406.

Tsherniak, A., Vazquez, F., Montgomery, P. G., Weir, B. A., Kryukov, G., Cowley, G. S., Gill, S., Harrington, W. F., Pantel, S., Krill-Burger, J. M., and et al. (2017). Defining a cancer dependency map. *Cell*, 170(3).

Tu, D., Mahony, B., Moore, T. M., Bertolero, M. A., Alexander-Bloch, A. F., Gur, R., Bassett, D. S., Satterthwaite, T. D., Raznahan, A., and Shinohara, R. T. (2022). Cocoa: Conditional correlation models with association size. *Biostatistics*, 25(1):154–170.

Van den Berge, K., Perraudeau, F., Soneson, C., Love, M. I., Risso, D., Vert, J.-P., Robinson, M. D., Dudoit, S., and Clement, L. (2018). Observation weights unlock bulk rna-seq tools for zero inflation and single-cell applications. *Genome Biology*, 19(1).

van Staveren, W., Solís, D. W., Hébrant, A., Detours, V., Dumont, J., and Maenhaut, C. (2009). Human cancer cell lines: Experimental models for cancer cells in situ? for cancer stem cells? *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1795(2):92–103.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120.

Wilding, G. E., Cai, X., Hutson, A., and Yu, Z. (2011). A linear model-based test for the heterogeneity of conditional correlations. *Journal of Applied Statistics*, 38(10):2355–2366.

Williams, S. P. and McDermott, U. (2017). The pursuit of therapeutic biomarkers with high-throughput cancer cell drug screens. *Cell Chemical Biology*, 24(9):1066–1074.

Wu, Q., Ma, T., Liu, Q., Milton, D. K., Zhang, Y., and Chen, S. (2021). Icn: Extracting interconnected communities in gene co-expression networks. *Bioinformatics*, 37(14):1997–2003.

Yang, J., Huang, Y., Song, M., Pan, Q., Zhao, J., He, J., Ouyang, D., Yang, C., Han, Y., Tang, Y., and et al. (2022). Spc25 promotes proliferation and stemness of hepatocellular carcinoma cells via the dna-pk/akt/notch1 signaling pathway. *International Journal of Biological Sciences*, 18(14):5241–5259.

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J. A., Thompson, I. R., and et al. (2013). Genomics of drug sensitivity in cancer (gdsc): A resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, 41(D1).

Yang, Z. and Ho, Y. (2021). Modeling dynamic correlation in zero-inflated bivariate count data with applications to single-cell rna sequencing data. *Biometrics*, 78(2):766–776.

Yocum, A. K., Anderau, S., Bertram, H., Burgess, H. J., Cochran, A. L., Deldin, P. J., Evans, S. J., Han, P., Jenkins, P. M., Kaur, R., and et al. (2023). Cohort profile update: The heinz c. prechter longitudinal study of bipolar disorder. *International Journal of Epidemiology*, 52(6).

Yonesaka, K., Kudo, K., Nishida, S., Takahama, T., Iwasa, T., Yoshida, T., Tanaka, K., Takeda, M., Kaneda, H., Okamoto, I., and et al. (2015). The pan-her family tyrosine kinase inhibitor afatinib overcomes her3 ligand heregulin-mediated resistance to egfr inhibitors in non-small cell lung cancer. *Oncotarget*, 6(32):33602–33611.

Yoon, G., Carroll, R. J., and Gaynanova, I. (2020). Sparse semiparametric canonical correlation analysis for data of mixed types. *Biometrika*, 107(3):609–625.

Zeller, K. I., Jegga, A. G., Aronow, B. J., O'Donnell, K. A., and Dang, C. V. (2003). An integrated database of genes responsive to the myc oncogenic transcription factor: Identification of direct genomic targets. *Genome Biology*, 4(10).

Zhang, J., Baran, J., Cros, A., Guberman, J. M., Haider, S., Hsu, J., Liang, Y., Rivkin, E., Wang, J., Whitty, B., and et al. (2011). International cancer genome consortium data portal–a one-stop shop for cancer genomics data. *Database*, 2011.

Zhang, J. and Li, Y. (2022). High-dimensional gaussian graphical regression models with covariates. *Journal of the American Statistical Association*, 118(543):2088–2100.

Zhang, N., Wang, H., Fang, Y., Wang, J., Zheng, X., and Liu, X. S. (2015). Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model. *PLOS Computational Biology*, 11(9).

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.