Bayesian Learning of Structured Covariances, with Applications to Cancer Data

by

Tsung-Hung Yao

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2023

Doctoral Committee:

Professor Veerabhadran Baladandayuthapani, Co-Chair
Assistant Professor Zhenke Wu, Co-Chair
Professor Karthik Bharath
Professor Jian Kang
Professor XuanLong Nguyen

Tsung-Hung Yao
yaots@umich.edu
ORCID iD: 0000-0001-5342-5430

ACKNOWLEDGEMENTS

First, I would like to convey my greatest gratitude to my advisor Professor Veera Baladandayuthapani. This dissertation will not be possible without Veera's continuous guidance, mentoring, and encouragement during the past five years. Veera is very supportive and inspiring, and I learned a lot from both the way he does research and how he mentors me. For example, instead of calling me his student, he always tells other collaborators that he "works" with me. As I work with Veera longer, Veera shows me not only how to be a great researcher also an awesome mentor.

My gratitude also goes to my co-advisor Professor Zhenke Wu, whose whose insights, feedback, and expertise are instrumental in refining the direction and depth of my research. Zhenke's dedicated guidance and collaborative spirit greatly enrich my academic experience.

My sincere appreciation goes to my amazing collaborators and coauthors: Professor Karthik Bharath, Professor Yang Ni, Professor Anindya Bhadra, Professor Jian Kang, and Dr. Jinju Li. I have enjoyed a lot during our discussions and learned so much from our collaborations. I would like to thank Professor Long Nguyen for serving on the dissertation committee and providing valuable feedback and constructive criticism that greatly improved the quality of this work.

I am thankful to my colleagues, lab mates and friends who provide valuable discussions and brainstorming sessions. Your input and encouragement have been instrumental in my growth as a researcher. I would also like to thank all the faculty members and staff members who provide the resources, facilities, and academic envi-

ronment that facilitated the progress of my research. I can not finish this dissertation without them.

Finally, my family deserve my heartfelt thanks for their unwavering support, understanding, and belief in my abilities. I want to express my gratitude to my partner, Shih-Hao Liu, especially in the most difficult time during Covid-19. Their love and encouragement sustained me through the challenging times and motivated me to strive for excellence.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

The identification of scientifically-driven dependence structures is of interest across many biomedical domains. Examples include tree- and graph-based structures that manifest themselves in precision medicine and genomic contexts. Such dependence structures can be compactly represented as covariance or precision matrices, which are useful for both characterizing and interpreting complex relationships. This dissertation develops a family of Bayesian models for structured covariances to investigate the biological dependencies, motivated by two applications in cancer research. These models are derived to adapt to different aspects of biological dependencies, such as the tree structure for assessing treatment similarity in pre-clinical cancer models and robust network structures for proteogenomics data incorporating tumor heterogeneity

In Chapter II, a novel Bayesian probabilistic tree-based framework is proposed for patient-derived xenografts data to investigate the hierarchical relationships between treatments by inferring treatment cluster trees, referred to as treatment trees (R_x-tree). This framework motivates a new measure of mechanistic similarity between two or more treatments accounting for inherent uncertainty in tree estimation; treatments with a high estimated similarity have potentially high mechanistic synergy. Building upon Dirichlet Diffusion Trees, I derive a closed-form marginal likelihood encoding the tree structure, which facilitates computationally efficient posterior inference via a new two-stage algorithm. Simulation studies demonstrate superior performance of the proposed method in recovering the tree structure and treatment similarities. My analyses of a recently collated PDX dataset produce treatment similarity estimates

that show a high degree of concordance with known biological mechanisms across treatments in five different cancers. More importantly, I uncover new and potentially effective combination therapies that uncover synergistic regulation of specific downstream biological pathways for future clinical investigations.

In Chapter III, I extend the work of the tree structure and the corresponding ultrametric matrices in Chapter II. Tree-structured covariances, or the equivalent ultrametric matrices, are an important class of matrices in statistics and machine learning with numerous applications. Ultrametric matrices are positive definite matrices that satisfy further ultrametric inequalities. Although projection- and relaxation-based estimation methods exist, there is a dearth of inferential techniques that provide appropriate uncertainty quantifications. The primary challenges lie in its non-trivial geometry. In this chapter, I first propose a novel consistent Markovian fragmentation prior over ultrametric matrices, building on Nabben-Varga decomposition in the matrix algebra literature. Importantly, the decomposition admits one-to-one mapping of ultrametric matrices to rooted trees, which I exploit to conduct inference in the surrogate Billera-Holmes-Vogtmann (BHV) space of rooted trees. My approach is novel because the metricized BHV space naturally motivates quick local moves along geodesics between neighboring tree topologies. In addition, because these moves do not rely on projection or relaxation during posterior computation, posterior summaries of central tendency and dispersion are readily available via Fréchet mean and geodesic distance in the BHV space. Simulation studies show that the proposed algorithm accurately recovers the matrix and the tree along with uncertainty quantification. I demonstrate the utility of the proposed method on the pre-clinical dataset by constructing the treatment tree and the mechanism similarity for multiple cancer treatments.

In Chapter IV, I shift the focus to Graphical models and investigate complex dependency structures in high-throughput datasets. Currently, most existing graphical

models make one of two canonical assumptions: (i) a homogeneous graph with a common network for all subjects; or (ii) rely on the normality assumption, especially in the context of Gaussian graphical models. Both assumptions are restrictive and can fail in certain applications, such as the proteomic networks in cancer. I propose robust Bayesian graphical regression (rBGR) to estimate heterogeneous graphs for non-normally distributed data. rBGR allows a flexible framework to estimate graphs by accommodating the non-normality through the random marginal transformations and constructs covariate-dependent graphs through graphical regression techniques. I also formulate a new characterization of edge dependencies in such models called conditional sign independence with covariates. In simulation studies, I demonstrate that rBGR outperforms existing Gaussian graphical regression for data generated under various levels of non-normality in both edge and covariate selection. I use rBGR to assess proteomic networks across two cancers: lung and ovarian, to systematically investigate the effects of immunogenic heterogeneity within tumors. My analyses reveal several important protein-protein interactions that are differentially impacted by the immune cell abundance; some corroborate existing biological knowledge but also discover several novel associations for future investigations.

CHAPTER I

Introduction

As the high-throughput screening techniques advance, modern data collection methods have allowed systematic assessment of multiple high-dimensional biomedical datasets simultaneously on the same or different tumor samples (Akbani et al., 2014; Baladandayuthapani et al., 2014). Subsequently, these high-dimensional datasets have enabled biologists to build scientific hypotheses on the relationships among different datasets and to recognize the importance of dependency in many fundamental biological processes (Airoldi, 2007; Sonawane et al., 2019). One famous example is the complex interactions among proteins that play a pivotal role in different molecular processes (Cheng et al., 2020). Statistically, these dependencies and interactions can be formulated using covariance matrices, to describe and conduct inference on the dependencies among different data. However, due to scientific hypotheses assumed about the different dependencies on the biomedical data, various conditions are imposed on the structure of the covariance matrix for different data and domains (e.g. McCullagh, 2006; Zorzi and Ferrante, 2012; Mieldzioc et al., 2021; hrer et al., 2023). In this dissertation, I develop a family of Bayesian models for structured covariances to investigate different biological dependencies that encode the underlying scientific hypotheses in cancer research.

Structured covariance matrices are widely used in statistics (and other fields)

and provide a flexible framework to characterize different aspects of dependencies in biomedical data. Consider a p-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_p)^\mathsf{T}$ representing observed data. In this dissertation, I model the dependency using a normal distribution as

$$\boldsymbol{X} \sim \boldsymbol{N}_p(\mathbf{0}, \boldsymbol{\Sigma}), \tag{1.1}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix describing the dependencies between components of $\boldsymbol{X}$. The models in this dissertation either implicitly or explicitly use the normal distribution for different scenarios. I choose the normal distribution for two reasons: (i) the covariance matrix in the normal distribution is a parameter that captures the second and higher orders of dependency (Casella and Berger, 2001); (ii) the ubiquitous use of the normal distribution in biomedical data. However, the unconstrained covariance does not satisfy the assumptions stemming from scientific hypotheses. I address two different constraints based on hypotheses for two different datasets in cancer research, resulting in tree- and graph-based covariances in this dissertation.

I investigate two types of dependencies that demonstrate different aspects of scientific hypotheses in cancer research. First, I focus on tree-based covariances from a dataset that investigates the treatment effectiveness for different cancer treatments. Specifically, I build a treatment tree that encodes the underlying mechanism similarity among different treatments and infer the treatment effectiveness based on this mechanism similarity. To model the tree structure as a matrix, I impose the ultrametric inequality (Lapointe and Legendre, 1991; Nabben and Varga, 1994; Bravo et al., 2009) on elements of the matrix, which regulates the values of off-diagonal elements. However, modeling the matrix with ultrametric inequality is a non-trivial problem, as the set of these matrices is neither a manifold (McCullagh, 2006) nor convex (Chierchia and Perret, 2020). Second, I shift the focus to graph-based covariance and aim

to construct dependencies that vary on covariate. Specifically, the structure of the covariance depends on the covariate-specific information, resulting in the network requires us to model a covariance that varies based on covariates from different subjects. Currently, existing approaches require at least one of the canonical assumptions of (i) a common network for all subjects or (ii) the normality assumption. However, both assumptions fail to hold in the motivating data. These challenges motivate this dissertation, and I address them in detail in the following chapters. For each chapter, I further elaborate on the key scientific and statistical themes and outline the progression of the dissertation.

1.1 Probabilistic Learning of Treatment Trees in Cancer

Key scientific questions. Accurate identification of synergistic treatment combinations and their underlying biological mechanisms is critical across many disease domains, especially cancer. Due to the impracticality of administering different treatment combinations on the same patient, preclinical systems such as patient-derived xenografts (PDX) have emerged to assess promising treatments and compounds before they are phased into human clinical trials. In translational oncology, PDX is a preclinical system with an experimental design that evaluates multiple treatments administered to samples from the same human tumor implanted into genetically identical mice (Hidalgo et al., 2014; Lai et al., 2017).

In Chapter II, I consider an experimental design of the PDX clinical trial that includes a large number of patients (Abdolahi et al., 2022) and tests a set of common treatments. This experimental design results in a data matrix such that each row in the matrix represents responses for a treatment from different patients, and each column presents responses from multiple mice with tumors implanted from the same patient. Due to this experimental design and the high clinical relevance (Oh and Bang, 2020; Abdolahi et al., 2022), a PDX clinical trial mimics a real human clinical trial

(Clohessy and Pandolfi, 2015), which allows me to answer the key scientific questions of: (a) identification of underlying plausible biological mechanisms, and (b) evaluation of the effectiveness of drug combinations based on mechanistic understanding.

Statistical themes. Ideally, treatments with the same target/mechanism should induce similar responses and engender mechanism-related clustering among treatments. Based on this idea, I use a tree structure to not only recuresively partition treatments into clusters but also quantify the similarity among clusters. In Chapter II, I propose a probabilistic tree-based framework for PDX data to investigate the mechanistic dependencies among treatments by inferring treatment trees. Specifically, I adapt a generative approach of Dirichlet diffusion trees that allows us to model the inherent uncertainty in the tree structure and, therefore, the underlying mechanism similarity. Building upon Dirichlet diffusion trees, I derive a closed-form marginal likelihood with covariance that encodes the tree structure. The likelihood further inspires a parameter decoupling strategy that facilitates an efficient new two-stage algorithm. I also develop posterior summaries that measure mechanistic similarity between two or more treatments, accounting for inherent uncertainty in tree estimation. I demonstrate the superior performance of my method in recovering the tree structure and the treatment similarities through a series of simulation studies under different data generating mechanisms. My analyses corroborate existing synergistic combination therapies while uncovering new ones. Additionally, I discover potentially effective combination therapies that confer synergistic regulation of specific downstream biological pathways for future clinical investigations. This Chapter is based on Yao et al. (2023).

1.2 Bayesian Inference for Ultrametric Covariances

Key scientific questions. In this Chapter, I continue and generalize the work of the tree-structured covariance in Chapter II. Tree-structured covariances, or ultrametric matrices, play an important role in statistics and machine learning with various scientific applications. For instance, in a multivariate Gaussian distribution, the covariance matrix is an ultrametric matrix if and only if the Gaussian density is multivariate totally positive of order two (Karlin and Rinott, 1983; Lauritzen et al., 2019), which implies a conditional positive dependency between two random variables (Fallat et al., 2017). Recently, ultrametric matrices have been applied in various scenarios as covariance matrices in Gaussian distributions, such as graphical models (Fallat et al., 2017) and Brownian motion tree models (e.g. Neal, 2003; Sturmfels et al., 2021), with applications in cancer biology (Yao et al., 2023) and finance (Agrawal et al., 2020). However, the inequalities required on ultrametric matrices also impose difficult constraints, as the constraints are highly non-convex (Chierchia and Perret, 2020), resulting in both computation and inference challenges. Although, many approaches have been proposed, to the best of my knowledge, no existing methods can quantify the uncertainty in ultrametric matrices.

Statistical themes. To address the problem of uncertainty quantification in ultrametric matrices, I propose a consistent Markovian prior for ultrametric matrices and develop a flexible Bayesian framework to obtain posterior samples of ultrametric matrices efficiently, thereby providing uncertainty quantification alongside point estimates. Specifically, I characterize the geometry of the space of ultrametric matrices through its bijection with the well-known BHV phylogenetic tree space, which allows us to conduct inference in the surrogate space of rooted trees. I leverage this characterization to develop an efficient posterior inference of Metropolis-Hastings algorithm. The algorithm makes local moves along geodesics without projection or relaxation.

Since the algorithm moves geodesically without leaving the space, posterior summaries of central tendency and dispersion are readily available via Fréchet mean and geodesic distance in the BHV space. Simulation studies show that the proposed algorithm recovers the matrix and the tree along with uncertainty quantification. I demonstrate utility of the proposed method on a pre-clinical dataset by constructing the treatment tree and the mechanism similarity for multiple cancer treatments.

1.3 Robust Bayesian Graphical Regression Models for Assessing Tumor Heterogeneity in Proteomic Networks

Key scientific questions. In Chapter IV, I shift focus to the dependencies that incorporate covariate-specific information and let the dependency structure (i.e. covariances) vary based on different covariates. Proteins control many fundamental cellular processes through a complex but organized system of interactions, termed protein-protein interactions (PPIs) (Cheng et al., 2020). Moreover, aberrant PPIs are associated with cancer, and investigating PPIs can lead to effective strategies and treatments (Lu et al., 2020). Recently, accumulating evidence suggests that considering tumor heterogeneity at the level of PPIs can enhance our understanding of tumorigenesis and the development of anti-cancer treatments (Cheng et al., 2020). Specifically, tumor heterogeneity differentially impacts the PPIs across different patients and results in varied treatment responses (Cheng et al., 2020). Hence, incorporating covariate-specific information, i.e., accounting for tumor heterogeneity, could provide valuable clues to identify PPIs disrupted during carcinogenesis. Consequently, it is highly desirable to elucidate PPIs in cancer and construct flexible graphical models that can identify multiple types and ranges of dependencies that vary based on different subjects. The key scientific questions I conclude are: (i) identify important PPIs across different cancer types and (ii) discover the effect of tumor

heterogeneity on aberrant PPIs as potential targets for future investigation.

Statistical themes. To construct the PPI network that includes covariate-specific information, I adapt graphical models to investigate complex dependency structures in proteomics (Airoldi, 2007). However, most existing graphical models make one of two canonical assumptions: (i) a homogeneous graph with a common network for all subjects; or (ii) rely on the normality assumption especially in the context of Gaussian graphical models (Ni et al., 2022a). As the tumor heterogeneity described above, presuming a common graph for all subjects is not appropriate. More importantly, the normality assumption does not always hold either for certain biomedical data such as the proteomic networks in cancer. In Chapter IV, I propose robust Bayesian graphical regression (rBGR) to estimate heterogeneous graphs for non-normally distributed data. Specifically, rBGR accommodates non-normality through a random transformation and constructs covariate-dependent graphs using graphical regression techniques. I also formulate a new characterization of edge dependencies in such models called conditional sign independence with covariates. In simulation studies, I demonstrate that rBGR outperforms existing Gaussian graphical regression for data generated under various levels of non-normality in both edge and covariate selection. I use rBGR to assess proteomic networks across two cancers: lung and ovarian, to systematically investigate the effects of immunogenic heterogeneity within tumors. My analyses reveal several important protein-protein interactions that are deferentially impacted by the immune cell abundance; some corroborate existing biological knowledge but also discover several novel associations for future investigations.

1.4 Scientific End-user Resources

I provide multiple general purpose `R` packages to estimate these structured covariances which are available at `https://github.com/bayesrx`. Specifically, for the

tree-structured covariances, the package `RxTree` and `UltrametricMat` can be used to fit the models described from Chapter II and III, respectively. The package `RxTree` models the trees with all leaves in the tree are equidistant to the root based on the Dirichlet diffusion tree model. On the other hand, `UltrametricMat` does not require the equidistant constraints on all leaves. Moreover, `UltrametricMat` defines a general prior for the tree structure and enables user-defined priors for tree structure. For graph-based method in Chapter IV, I provide the package of `rBGR` to construct covariate-specific graph under non-normal data.

CHAPTER II

Probabilistic Learning of Treatment Trees in Cancer

2.1 Introduction

According to the World Health Organization, cancer is one of the leading causes of death globally, with ∼10 million deaths in 2020 (Ferlay et al., 2020). Despite multiple advances over the years, systematic efforts to predict efficacy of cancer treatments have been stymied due to multiple factors, including patient-specific heterogeneity and treatment resistance (Dagogo-Jack and Shaw, 2018; Groisberg and Subbiah, 2021). Given that the evolution of tumors relies on a limited number of biological mechanisms, there has been a recent push towards combining multiple therapeutic agents, referred to as "combination therapy" (Sawyers, 2013; Groisberg and Subbiah, 2021). This is driven by the core hypothesis that combinations of drugs act in synergistic manner, with each drug compensating for the drawbacks of other drugs. However, despite higher response rates and efficacy in certain instances (Bayat Mokhtari et al., 2017), combination therapy can lead to undesired drug-drug interactions, lower efficacy, or severe side effects (Sun et al., 2016). Consequently, it is highly desirable to advance the understanding of underlying mechanisms that confer synergistic drug effects and identify potential favorable drug-drug interaction mechanisms for further

investigations.

Given that not all possible drug combinations can be tested on patients in actual clinical trials, cancer researchers rely on preclinical "model" systems to guide the discovery of the most effective combination therapies (note, models have a different contextual meaning here). In translational oncology, preclinical models assess promising treatments and compounds, before they are phased into human clinical trials. The traditional mainstay of such preclinical models has been cell-lines, wherein cell cultures derived from human tumors are grown in an *in vitro* controlled environment. However, it has been argued that they do not accurately reflect the true behavior of the host tumor and, in the process of adapting to *in vitro* growth, lose the original properties of the host tumor, thus leading to limited clinical relevance and successes (Tentler et al., 2012; Bhimani et al., 2020). To overcome these challenges, there has been a push towards more clinically relevant model systems that maintain a high degree of fidelity to human tumors. One such preclinical model system is Patient-Derived Xenograft (PDX) wherein tumor fragments obtained from cancer patients are directly transplanted into genetically identical mice (Hidalgo et al., 2014; Lai et al., 2017). Compared to traditional oncology models such as cell-lines (Yoshida, 2020), PDX models maintain key cellular and molecular characteristics, and are thus more likely to mimic human tumors and facilitate precision medicine. More importantly, accumulating evidence suggests responses (e.g. drug sensitivity) to standard therapeutic regimens in PDXs closely correlate with patient clinical data, making PDX an effective and predictive experimental model across multiple cancers (Topp et al., 2014; Nunes et al., 2015).

PDX experimental design and key scientific questions. Overall, the PDX experimental design depends on the purpose of the study and we consider a PDX study of the PDX clinical trial that includes a large number of patients (Abdolahi et al., 2022) and tests a set of common treatments. The PDX experiment then

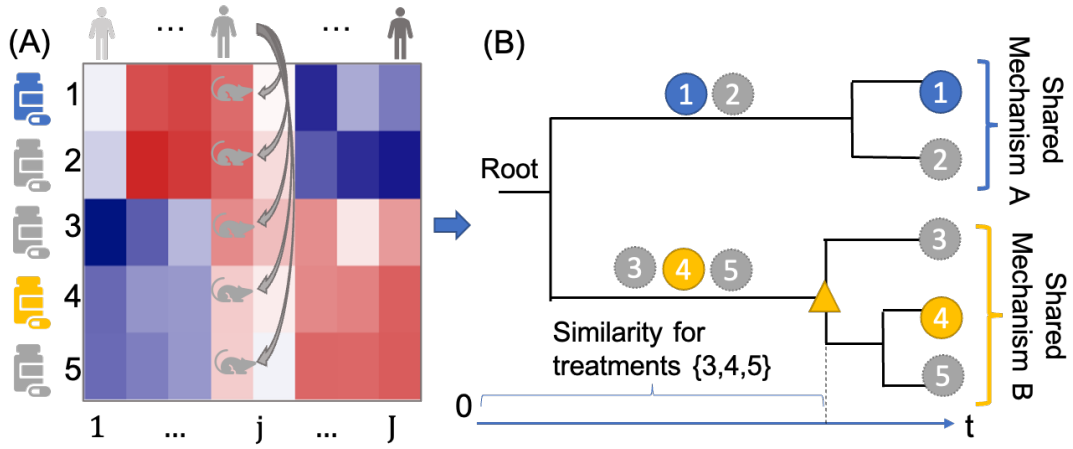


Figure II.1: PDX experimental design and tree-based representation. Panel **A**: an illustrative PDX dataset with five treatments (row) and eight patients (column). Mice in a given column are implanted with tumors from the same patient and receive different treatments (across rows). The level of tumor responses are shown along a color gradient. Panel **B**: a tree structure that clusters the treatments and quantifies the similarity among mechanisms. Two treatments (1 and 4) are assumed to have different but known biological mechanisms (in different colors); the rest three treatments (2,3, and 5) have unknown mechanisms (in gray). The tree suggests two treatment groups are present ($\{1, 2\}$ and $\{3, 4, 5\}$) that may correspond to two different known mechanisms. The horizontal position of "△" represents the divergence time (defined in Section 2.2.1) and the mechanism similarity for treatments $\{3, 4, 5\}$. In a real data analysis, the tree (topology and divergence times) is unknown and is to be inferred from PDX data.

implants the tumor cell to multiple mice and each treatment is given to multiple mice with tumors implanted from the same (matched) patient (see conceptual schema in Figure II.1(A)). Treatment responses (e.g. tumor size) are then evaluated, resulting in a data matrix (treatments × patients) as depicted in the heatmap in Figure II.1(A). The PDX-based clinical trial is a powerful tool for detecting the drug efficacy and drug sensitivity (Abdolahi et al., 2022) and has been adapted in several studies for different cancers (e.g. Zhang et al., 2013 for the breast cancer and Bertotti et al., 2011 for the colorectal cancer). Due to the relatively high fidelity between PDX models and the human tumors (Oh and Bang, 2020; Abdolahi et al., 2022), a PDX-based clinical trial mirrors a real human clinical trial using mouse "avatars" (Clohessy and Pandolfi, 2015). Thus this protocol serves as a scalable platform to: (a) identify underlying plausible biological mechanisms responsible for tumor growth and resistance, and (b)

evaluate the effectiveness of drug combinations based on mechanistic understanding. In this context, the (biological) mechanism refers to the specific mechanism of action of a treatment, which usually represents a specific target, such as an enzyme or a receptor (Grant et al., 2010). From the perspective of treatment responses as data, responses are the consequences of the downstream biological pathways from the corresponding interaction between a treatment and the target/mechanism.

Ideally, treatments with the same target/mechanism should induce similar responses and engender mechanism-related clustering among treatments. Evidently then, a sensible clustering of treatments would not only partition treatments into clusters but also explicate how the clusters relate to one another; in other words, a hierarchy among treatment clusters is more likely to uncover plausible mechanisms for combinations of treatments with "similar" responses when compared to "flat" clusters (e.g., k-means clustering). Such response-based identification of potential synergistic effects from combinations of treatments will augment understanding from known mechanistic synergy. In our application, using tree-based clustering, we assume known entities at the leaves, i.e., the different treatments. The treatments are assumed to act upon potentially distinct biological pathways, resulting in different levels of responses across the treated mice. In this Chapter, we use PDX response data on the leaves to infer a hierarchy over treatments that may empirically characterize the similarity in the targeted mechanistic pathways. The primary statistical goals are to (i) define and estimate a general metric measuring the similarity within any subset comprising two or more treatments, and (ii) facilitate (i) by conceptualizing and inferring an unknown hierarchy among treatments.

Tree-based representations for PDX data. To this end, we consider a tree-based construct to explore the hierarchical relationships between treatments, referred to as *treatment tree (R_x-tree*, in short). We view such a tree structure as a representation of clustering of treatments based on mechanisms that confer synergistic effects, wherein

similarities between mechanisms are captured through branch lengths. Hierarchy among treatments can be interpreted through branch lengths (from the root) that are potentially reflective of different cancer processes; this would then help identify common mechanisms and point towards treatment combinations disrupting oncological processes if administered simultaneously.

We will focus on rooted trees. The principal ingredients of a rooted tree comprise a root node, terminal nodes (or, leaves), internal nodes and branch lengths. In the context of the R_x-tree for PDX data, the leaves are observed treatment responses, whereas internal nodes and branch lengths are unobserved. Internal nodes are clusters of treatments, and lengths of branches between nodes are indicative of strengths of mechanism similarities. The root is a single cluster consisting of all treatments. This leads to the following interpretation: at the root all treatments share a common target or mechanism; length of path from the root to the internal node (sum of branch lengths) at which two treatments split into different clusters measures mechanism similarity between the two treatments. Thus treatments that stay clustered "longer" have higher mechanism similarities.

Throughout, we will use 'tree' when describing methodology for an abstract tree (acyclic graphs with distinguished root node) and 'treatment tree' or 'R_x-tree' when referring to the latent tree within the application context.

An illustrative example. A conceptual R_x-tree and its interpretation is illustrated in Figure II.1 where five treatments (1 to 5) are applied on eight patients' PDXs (Figure II.1(A)) with the corresponding (unknown true) R_x-tree (Figure II.1(B)) based on the PDX data. Assume two treatment groups based on different mechanisms – treatments $\{1, 2\}$ and treatments $\{3, 4, 5\}$; further, suppose treatment 4 is approved by the Food and Drug Administration (FDA). The heatmap in Panel (A) visualizes the distinct levels of response profiles to the five treatments so that treatments closer in the tree are more likely to have similar levels of responses. The R_x-tree captures

the mechanism similarity by arranging treatments $\{1, 2\}$ and $\{3, 4, 5\}$ to stay in their respective subtrees longer and to separate the two sets of treatment early in the tree. Based on the R_x-tree, treatments $\{3, 5\}$ share high mechanism similarity values with treatment 4; treatment 5 is the closest to the treatment 4, suggesting the most similar synergistic mechanism among all the evaluated treatments 1 to 5.

Existing methods and modeling background. The Pearson correlation is a popular choice to assess mechanism similarity between treatments (Krumbach et al., 2011), but is inappropriate to examine multi-way similarity. A tree-structured approach based on a (binary) dendrogram obtained from hierarchical clustering of cell-line data using the cophenetic distance (Sokal and Rohlf, 1962) was adopted in Narayan et al. (2020); their approach, however, failed to account for uncertainty in the dendrogram, which is highly sensitive to measurement error in the response variables as well distance metrics (we show this via simulations and in real data analyses). Another example with a binary dendrogram of hierarchical clustering was proposed by Rashid et al. (2020), which also utilizes the same PDX dataset as this Chapter. However, their model uses the tree structure to model the individualized treatment rule for different patients, while our method focuses on the tree structure itself and the corresponding mechanism similarity. In this Chapter, we consider a model for PDX data parameterized by a tree-structured object representing the R_x-tree. The model is derived from the Dirichlet diffusion tree (DDT) (Neal, 2003) generative model for (hierarchically) clustered data. The DDT engenders a data likelihood and a prior distribution on the tree parameter with support in the space of rooted binary trees. We can then use the posterior distribution to quantify uncertainty about the latent R_x-tree.

Summary of novel contributions and organization of the article. Our approach based on the DDT model for PDX data results in three main novel contributions:

(a) *Derivation of a closed-form likelihood that encodes the tree structure.* The DDT specification results in a joint distribution on PDX data, treatment tree parameters and other model parameters. By marginalizing over unobserved data that correspond to internal nodes of the tree, we obtain a new multivariate Gaussian likelihood with a special tree-structured covariance matrix, which completely characterizes the treatment tree (Proposition 1 and Lemma 2.3.1).

(b) *Efficient two-stage algorithm for posterior sampling.* Motivated by the form of marginal data likelihood in (a), we decouple the Euclidean and tree parameters and propose a two-stage algorithm that combines an approximate Bayesian computation (ABC) procedure (for Euclidean parameters) with a Metropolis-Hasting (MH) step (for tree parameters). We demonstrate via multiple simulation studies the superiority of our hybrid approach over approaches based on classical single-stage MH algorithms (Sections 2.4.2 and 2.4.1).

(c) *Corroborating existing, and uncovering new, synergistic combination therapies.* We define and infer a new similarity measure that accounts for inherent uncertainty in estimating a latent hierarchy among treatments. As a result, the *maximum a posteriori* R_x-tree and the related mechanism similarity show high concordance with known existing biological mechanisms for monotherapies and uncover new and potentially useful combination therapies (Sections 2.5.3 and 2.5.4).

Of particular note is contribution (c), where we leverage a recently collated PDX dataset from the Novartis Institutes for BioMedical Research - PDX Encyclopedia [NIBR-PDXE, (Gao et al., 2015)] that interrogated multiple targeted therapies across five different cancers. Our pan-cancer analyses of the NIBR-PDXE dataset show a high degree of concordance with known existing biological mechanisms across different cancers; for example, a high mechanistic similarity is suggested between two agents currently in clinical trials: CGM097 and HDM201 in breast cancer and colorectal

cancer, known to target the same gene MDM2 (Konopleva et al., 2020). In addition, our model uncovers new and potentially effective combination therapies. For example, exploiting knowledge of the combination therapy of a class of agents targeting the PI3K-MAPK-CDK pathway axes – PI3K-CDK for breast cancer, PI3K-ERBB3 for colorectal cancer and BRAF-PI3K for melanoma – confers possible synergistic regulation for prioritization in future clinical studies.

The rest of the Chapter is organized as follows: we first review our probabilistic formulation for PDX data based on the DDT model and present the marginal data likelihood and computational implications in Section 2.2. In Section 2.3, we derive the posterior inference algorithm based on a two-stage algorithm. In Section 2.4, we conduct two sets of simulations to evaluate the operating characteristics of the model and algorithm. A detailed analysis of the NIBR-PDXE dataset, results, biological interpretations and implications are summarized in Section 2.5. This Chapter concludes by discussing implications of the findings, limitations, and future directions.

2.2 Modeling $\mathbf{R_x}$-tree via Dirichlet Diffusion Trees

Given a PDX experiment with I correlated treatments and J independent patients, we focus on the setting with $1 \times 1 \times 1$ design (one animal per PDX model per treatment) with no replicate response for each treatment and patient. A PDX experiment produces an observed data matrix $\mathbf{X}_{I\times J} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_I]^\mathsf{T}$ where $\boldsymbol{X}_i = [X_{i1}, \ldots, X_{iJ}]^\mathsf{T}$ is data under treatment i across J patients; let the observed response column for each patient be $\boldsymbol{X}_{\cdot,j} = [x_{1j}, \ldots, x_{Ij}]^\mathsf{T} \in \mathbb{R}^I, j = 1, \ldots, J$.

In this Chapter, the observed treatment responses are continuous and we model the responses through a generative model that results in a Gaussian likelihood with a structured covariance:

$$\boldsymbol{X}_{\cdot,j} | \boldsymbol{\Sigma}^{\mathcal{T}}, \sigma^2 \overset{iid}{\sim} \mathsf{N}_I(0, \boldsymbol{\Sigma}^{\mathcal{T}}), \quad j = 1, \ldots, J, \tag{2.1}$$

where the $\mathbf{\Sigma}^{\mathcal{T}}$ is a tree-structured covariance matrix that encodes the tree $\mathcal{T}$. In particular, $\mathbf{\Sigma}^{\mathcal{T}} = \left\{\mathbf{\Sigma}^{\mathcal{T}}_{i,i'}, i, i' = 1, \ldots, I\right\}$ encodes the tree $\mathcal{T}$ through two constraints (Lapointe and Legendre, 1991; McCullagh, 2006):

$$\mathbf{\Sigma}^{\mathcal{T}}_{i',i} = \mathbf{\Sigma}^{\mathcal{T}}_{i,i'} \geq 0; \mathbf{\Sigma}^{\mathcal{T}}_{i,i} \geq \mathbf{\Sigma}^{\mathcal{T}}_{i,i'}, \tag{2.2}$$

$$\mathbf{\Sigma}^{\mathcal{T}}_{i,i'} \geq \min\{\mathbf{\Sigma}^{\mathcal{T}}_{i,i''}\ ,\ \mathbf{\Sigma}^{\mathcal{T}}_{i',i''}\} \text{ for all } i \neq i' \neq i''. \tag{2.3}$$

Each element $\mathbf{\Sigma}^{\mathcal{T}}_{i,i}$ is the covariance between treatments i and i' and measures their similarity. The inequality (2.2) imposes the symmetry of covariance matrix and ensures the divergence of all leaves. The tree structure is characterized by the ultrametric inequality (2.3) that ensures $\mathbf{\Sigma}^{\mathcal{T}}$ bijectively maps to a tree $\mathcal{T}$; for more details on the relationship between the covariance $\mathbf{\Sigma}^{\mathcal{T}}$ and the tree $\mathcal{T}$ see McCullagh (2006) and Bravo et al. (2009). Of note, mean parameterized models (e.g. mixed effects models) are inappropriate for uncovering the tree parameter under the given data structure since the latent tree is completely encoded in covariance matrix $\mathbf{\Sigma}^{\mathcal{T}}$.

A Bayesian formulation requires an explicit prior distribution on $\mathbf{\Sigma}^{\mathcal{T}}$ which satisfies constrains (2.2) and (2.3); this requirement is far from straightforward since the set of tree-structured matrices is complicated (e.g., it is not a manifold (McCullagh, 2006)). We instead consider the Dirichlet Diffusion tree (DDT) model (Neal, 2003) for hierarchically clustered data which provides two useful ingredients:

1. a prior is implicitly specified on the latent treatment tree, comprising the root, internal nodes, leaves, and branch lengths;

2. upon integrating out the internal nodes, a tractable Gaussian likelihood on PDX data with tree-structured covariance is specified.

We first provide a brief description of the DDT model proposed by Neal (2003) and its joint density on data and tree (Section 2.2.1). Subsequently, we derive an expression for the likelihood and demonstrate how it can be profitably employed to

develop a generative model for PDX data and carry out R_x-tree estimation (Section 2.2.2 and 2.2.3).

2.2.1 The Generative Process of DDT

The DDT prescribes a fragmentary, top-down mechanism to generate a binary tree (acyclic graph with a preferred node or vertex referred to as the root), starting from a root, containing J-dimensional observed responses $\boldsymbol{X}_i$ at I leaves/terminal nodes; each node in the tree has either 0 or 2 children excepting the root which has a solitary child. This prescription manipulates dynamics of a system of I independent Brownian motions $B_1, \ldots, B_I$ on $\mathbb{R}^J$ in a common time interval $t \in [0, 1]$. As shown in Figure II.2(A), all Brownian motions $B_i(t)$ start at the same point at time $t = 0$, location of which is the root $\mathbf{0} \in \mathbb{R}^J$, and diverge at time points in $[0, 1]$ and locations in $\mathbb{R}^J$ before stopping at the time $t = 1$ at locations $\boldsymbol{X}_i$. The Brownian trajectories and their divergences engender the tree structure as shown in Figure II.2(A).

Specifics on when and how the Brownian motions diverge are as follows: the first Brownian motion $B_1(t)$ starts at $t = 0$ and generates $\boldsymbol{X}_1$ at $t = 1$; a second independent Brownian motion $B_2(t)$ starts at the same point at $t = 0$, branches out from the first Brownian motion at some time t, after which it generates $\boldsymbol{X}_2$ at time 1. The probability of divergence in a small interval $[t, t + dt]$ is given by a *divergence function* $t \mapsto a(t)$, assumed as in Neal (2003) to be of the form $a(t) = c(1 - t)^{-1}$ for some divergence parameter $c > 0$. Inductively then, the vector of observed responses to treatment i, $\boldsymbol{X}_i$, is generated by $B_i(t)$, which follows the path of previous ones. If at time t, $B_i(t)$ has not diverged and meets the previous divergent point, it will follow one of the existing path with the probability proportional to the number of data points that have previously traversed along each path. Eventually, given $B_i(t)$ has not diverged at time t, it will do so in $[t, t + dt]$ with probability $a(t)dt/m$, where m is the number of data points that have previously traversed the current path.

From the illustration in panel (A) of Figure II.2, we note that B_3 diverges from the B_1 and B_2 at time t_1 at location $\boldsymbol{X}_1'$ and at $t = 1$ is at location $\boldsymbol{X}_3$, which is the J-dimensional response vector for treatment 3; this creates a solitary branch of length t_1 from the root and an unobserved internal node at location $\boldsymbol{X}_1'$. Continuing, given three Brownian motions B_1, B_2 and B_3, B_4 does not diverge before t_1 and meet the previous divergent point t_1. B_4 chooses to follow the path of B_3 with probability $1/3$ at t_1 and finally diverges from B_3 at time $t_2 > t_1$ at location $\boldsymbol{X}_2'$; this results in observation $\boldsymbol{X}_4$ for treatment 4 and an unobserved internal node at $\boldsymbol{X}_2'$, and so on. As a consequence, the binary tree that arises from the DDT comprises of:

(i) an unobserved root at the origin in $\mathbb{R}^J$ at time $t = 0$;

(ii) observed data $\mathbf{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_I]^\mathsf{T} \in \mathbb{R}^{I \times J}$ situated at the leaves of the tree;

(iii) unobserved internal nodes $\mathbf{X}^I = [\boldsymbol{X}_1', \ldots, \boldsymbol{X}_{I-1}']^\mathsf{T} \in \mathbb{R}^{(I-1)\times J}$;

(iv) unobserved times $\boldsymbol{t} = (t_1, \ldots, t_{I-1})^\mathsf{T} \in [0, 1]^{I-1}$ that characterize lengths of branches;

(v) unobserved topology $\mathcal{T}$ that links (i)-(iv) into a tree structure, determined by the number of data points $\boldsymbol{X}_i$ that have traversed through each segment or branch.

Conceptually, observed data at the leaves $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_I$ collectively form the observed PDX responses generated through a process involving a few parameters: tree-related parameters $(\mathcal{T}, \boldsymbol{t})$ and the locations of internal nodes $\boldsymbol{X}_i'$. The tree $\mathcal{T}$ clusters I treatments as a hierarchy of $(I-1)$ levels (excluding the last level containing leaves). At level $0 < d \leq I - 1$ of the hierarchy, characterized by the pair $(\boldsymbol{X}_d', t_d)$, the I treatments are clustered into $d + 1$ groups; a measure of similarity (or dissimilarity) between treatment clusters at levels d and $d+1$ is given by the branch length $t_{d+1} - t_d$.

We now give a brief description of how the joint density of $(\mathbf{X}, \mathbf{X}^I, \boldsymbol{t}, \mathcal{T})$ can be derived; for more details we direct the reader to Neal (2003) and Knowles and

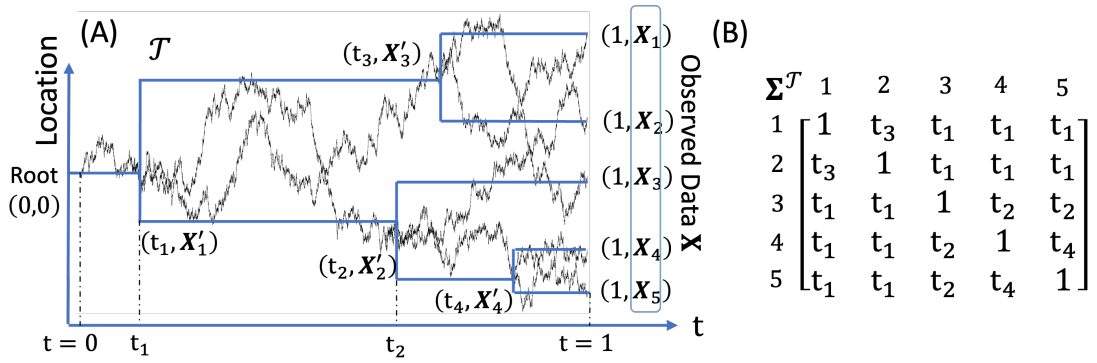


Figure II.2: (A) A binary tree with $I = 5$ leaves underlying the diffusion dynamics. The observed response vector $\boldsymbol{X}_i, i = 1, \ldots, I$ is generated by the Brownian motion up to $t = 1$. The unobserved response vector $\boldsymbol{X}'_d, d = 1, \ldots, (I-1)$ at the divergence is generated by the Brownian motion at time t_d. (B) A tree-structured matrix $\mathbf{\Sigma}^{\mathcal{T}}$ that encapsulates the tree $\mathcal{T}$. See the Proposition 1 for the definition of $\mathbf{\Sigma}^{\mathcal{T}}$.

Ghahramani (2015). For a fixed $c > 0$ that governs the divergence function $a(t) = c(1-t)^{-1}$, probabilities associated with the independent Brownian motions $B_1, \ldots, B_I$ induce a joint (Lebesgue) density on the generated tree. Note that the binary tree arising from the DDT is encoded by the triples $\{(t_d, \boldsymbol{X}'_d, \boldsymbol{X}_i), d = 1, \ldots, I-1; i = 1, \ldots, I\}$. An internal node at $\boldsymbol{X}'_d$ contains l_d and r_d leaves below to its left and right with $m_d = l_d + r_d$. If each of the Brownian motions is scaled by $\sigma^2 > 0$, then given $\mathcal{T}$ and a branch with endpoints $(t_u, \boldsymbol{X}'_u)$ and $(t_v, \boldsymbol{X}'_v)$ with $0 < t_u < t_v < 1$, from properties of a Brownian motion we see that $\boldsymbol{X}'_v \sim N_J(\boldsymbol{X}'_u, \sigma^2(t_v - t_u)\boldsymbol{I}_J)$, and the (Lebesgue) density of $\mathcal{T}$ can be expressed as the product of contributions from its branches. Then the joint density of all nodes, times and the tree topology is given by

$$P(\mathbf{X}, \mathbf{X}^I, \boldsymbol{t}, \mathcal{T} | c, \sigma^2) = \prod_{[u,v] \in \mathcal{S}(\mathcal{T})} \frac{(l_v - 1)!(r_v - 1)!}{(l_v + r_v - 1)!} c(1 - t_v)^{cJ_v - 1} N(\boldsymbol{X}'_u, \sigma^2(t_v - t_u)\boldsymbol{I}_J) \tag{2.4}$$

where $\mathcal{S}(\mathcal{T})$ is the collection of branches and $\mathbf{X}^I_{(I-1)\times J} = [\boldsymbol{X}'_1, \ldots, \boldsymbol{X}'_{(I-1)}]^{\mathsf{T}}$ are unobserved locations of the internal nodes. On each branch $[u, v]$, the first term $\frac{(l_v-1)!(r_v-1)!}{(l_v+r_v-1)!}$ represents the chance the branch containing l_v and r_v leaves to its left and right re-

spectively; $c(1-t_v)^{cJ_v-1}$ represents the probability of diverging at t_v with l_v and r_v leaves, where $J_v = H_{l_v+r_v-1} - H_{l_v-1} - H_{r_v-1}$ with $H_n = \sum_{i=1}^{n} 1/i$ is the nth harmonic number.

The joint density is hence parameterized by (c, σ^2), where c plays a crucial role in determining the topology $\mathcal{T}$: through the divergence function $a(t)$, it determines the propensity of the Brownian motion to diverge from its predecessors; consequently, a small c engenders later divergence and a higher degree of similarity among treatments in PDX. The latent tree has two components: (i) topology $\mathcal{T}$ and (ii) vector of divergence times $\boldsymbol{t}$ determining branch lengths. We refer to (c, σ^2) as the *Euclidean parameters* and $(\mathcal{T}, \boldsymbol{t})$ as *tree parameters*.

2.2.2 Prior on Tree and Closed-form Likelihood

The joint density in (2.4) factors into a prior $P(\boldsymbol{t}, \mathcal{T}|c, \sigma^2)$ on the tree parameter through $(\mathcal{T}, \boldsymbol{t})$ and a density $P(\mathbf{X}, \mathbf{X}^I|\boldsymbol{t}, \mathcal{T}, c, \sigma^2)$ that is a product of J-dimensional Gaussians on the internal nodes and leaves. The prior distribution on the latent tree is thus implicitly defined through the Brownian dynamics and is parameterized by $(\mathcal{T}, \boldsymbol{t})$ with hyperparameters (c, σ^2). In (2.4) the product is over the set of branches $\mathcal{S}(\mathcal{T})$, and the contribution to the prior $P(\mathcal{T}, \boldsymbol{t}|c, \sigma^2)$ from each branch $[u, v]$ is $\frac{(l_v-1)!(r_v-1)!}{(l_v+r_v-1)!} c(1-t_v)^{cJ_{l_v,r_v}-1}$, which is free of σ^2; on the other hand, the contribution to $P(\mathbf{X}, \mathbf{X}^I|\boldsymbol{t}, \mathcal{T}, c, \sigma^2)$ from $[u, v]$ is the J-dimensional $N_J(\boldsymbol{X}'_u, \sigma^2(t_v - t_u)\boldsymbol{I}_J)$, which is independent of c. The likelihood function based on the observed $\mathbf{X}$ is thus obtained by integrating out the unobserved internal nodes $\mathbf{X}^I$ from $P(\mathbf{X}, \mathbf{X}^I|\boldsymbol{t}, \mathcal{T}, \sigma^2)$. Accordingly, our first contribution is to derive a closed-form likelihood function for efficient posterior computations; to our knowledge, this task is currently achieved only through sampling-based or variational methods (Neal, 2003; Knowles and Ghahramani, 2015).

Denote as $\mathsf{MN}_{I\times J}(M, U, V)$ the matrix normal distribution of an $I \times J$ random matrix with mean matrix M, row covariance U, and column covariance V,

and let $\boldsymbol{I}_k$ denote the $k \times k$ identity matrix. Evidently, $\mathbf{X}$ follows a matrix normal distribution since Gaussian laws of the Brownian motions imply that $[\mathbf{X}, \mathbf{X}^I] = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_I, \boldsymbol{X}'_1, \ldots, \boldsymbol{X}'_{(I-1)}]^\mathsf{T}$ follow a matrix normal distribution.

Proposition 1. Under the assumption that the root is located at the origin in $\mathbb{R}^J$, the data likelihood $\mathbf{X}|\sigma^2, \mathcal{T}, \boldsymbol{t} \sim \mathsf{MN}_{I\times J}(\mathbf{0}, \sigma^2\boldsymbol{\Sigma}^{\mathcal{T}}, \boldsymbol{I}_J)$, where $\boldsymbol{\Sigma}^{\mathcal{T}} = \left(\boldsymbol{\Sigma}^{\mathcal{T}}_{i,i'}\right)$ is an $I \times I$ tree-structured covariance matrix satisfying (2.2) and (2.3) with $\boldsymbol{\Sigma}^{\mathcal{T}}_{i,i} = 1$ and $\boldsymbol{\Sigma}^{\mathcal{T}}_{i,i'} = t_d$, for $i \neq i'$ where $i, i' = 1, \ldots, I$ and $d = 1, \ldots, I-1$.

Proposition 1 asserts that use of the DDT model leads to a centered Gaussian likelihood on PDX data $\boldsymbol{X}$ with a tree-structured covariance matrix. Proposition 1 also implies that each patient independently follows the normal distribution of (2.1) with an additional scale parameter (σ^2) from the Brownian motion:

$$\boldsymbol{X}_{\cdot,j}|\boldsymbol{\Sigma}^{\mathcal{T}}, \sigma^2 \overset{iid}{\sim} \mathsf{N}_I(0, \sigma^2\boldsymbol{\Sigma}^{\mathcal{T}}), \quad j = 1, \ldots, J. \tag{2.5}$$

By setting $\boldsymbol{\Sigma}^{\mathcal{T}}_{i,i'} = t_{i,i'}$ as the divergence time of i and i', $\boldsymbol{\Sigma}^{\mathcal{T}}$ satisfies (2.2) and (2.3) and encodes the tree $\mathcal{T}$. For example, consider a three-leaf tree with $\boldsymbol{\Sigma}^{\mathcal{T}}_{i,i'} = t_{i,i'}$, inequality (2.3) implies that for the three leaves, say, i, i' and i'', one of the following conditions must hold: (i) $t_{i',i''} \geq t_{i,i'} = t_{i,i''}$; (ii) $t_{i,i''} \geq t_{i,i'} = t_{i',i''}$; (iii) $t_{i,i'} \geq t_{i,i''} = t_{i',i''}$. We then obtain a tree containing 1) a subtree of two leaves with a higher similarity and 2) a singleton clade with a lower similarity between the singleton leaf and the two leaves in the first subtree. In particular, if $t_{i',i''} \geq t_{i,i'} = t_{i,i''}$ holds, the three-leaf tree has leaf i diverging earlier before the subtree of (i', i'').

2.2.3 Decoupling Tree and Euclidean Parameters for Efficient Sampling.

In the full joint density in (2.4) the Euclidean and tree parameters are confounded across row and column dimensions of $\mathbf{X}$, and this may result in slow mixing of chains using traditional MCMC algorithms (Turner et al., 2013). State-of-the-art posterior

inference on $(c, \sigma^2, \mathcal{T}, \boldsymbol{t})$ can be broadly classified into sampling-based approaches (e.g., Knowles and Ghahramani, 2015) and deterministic approaches based on variational message passing (e.g., Knowles et al., 2011, VMP). Variational algorithms can introduce approximation errors to the joint posterior via factorization assumptions (e.g.,mean-field) and choice of algorithm is typically determined by the speed-accuracy trade-off tailored for particular applications. On the other hand, in classical MCMC-based algorithms for DDT we observed slow convergence in the sampling chains for c and σ^2 with high autocorrelations for the corresponding chains, owing to possibly the high mutual dependence between c in the divergence function and the tree topology $\mathcal{T}$, resulting in slow local movements in the joint parameter space of model and tree parameters (Simulation II in Section 2.4.2).

Notwithstanding absence of the parameter c in the Gaussian likelihood, the dependence, and information about, c is implicit: the distribution of divergence times $\boldsymbol{t}$ that populate $\boldsymbol{\Sigma}^{\mathcal{T}}$ are completely determined by the divergence function $t \mapsto c(1-t)^{-1}$. In other words, c can indeed be estimated from treatment responses $\{\boldsymbol{X}_{\cdot,j}\}$ using the likelihood. From a sampling perspective, however, form of the likelihood obtained by integrating out the internal nodes $\mathbf{X}^I$, suggests an efficient two-stage sampling strategy that resembles the classical collapsed sampling (Liu, 1994) strategy in MCMC literature: first draw posterior samples of (c, σ^2) and then proceed to draw posterior samples of $(\mathcal{T}, \boldsymbol{t})$ conditioned on each sample of (c, σ^2).

2.3 $\mathbf{R_x}$-tree Estimation and Posterior Inference

In line with the preceding discussion, we consider a two-stage sampler for Euclidean and tree parameters. While in principle MCMC techniques could be used in both stages, we propose to use a hybrid ABC-MH algorithm. Specifically, we use an approximate Bayesian computation (ABC) scheme to draw weighted samples of (c, σ^2) in the first stage followed by a Metropolis-Hastings (MH) step that samples

$(\mathcal{T}, \boldsymbol{t})$ given ABC samples of (c, σ^2) in the second stage. Motivation for using ABC in the first stage stems from: (i) availability of informative statistics; (ii) generation of better quality samples of the tree (compared to a single-stage MH); and (iii) better computational efficiency. We refer to Section 2.4.2 for more details.

2.3.1 Hybrid ABC-MH Algorithm

ABC is a family of inference techniques that are designed to estimate the posterior density $\text{pr}(\theta|\mathcal{D})$ of parameters θ given data $\mathcal{D}$ when the corresponding likelihood $\text{pr}(\mathcal{D}|\theta)$ is intractable but fairly simple to sample from. Summarily, ABC approximates $\text{pr}(\theta|\mathcal{D})$ by $\text{pr}(\theta|\boldsymbol{S}_{obs})$ where $\boldsymbol{S}_{obs}$ is a d-dimensional summary statistic that ideally captures most information about θ. In the special case where $\boldsymbol{S}_{obs}$ is a vector of sufficient statistics, it is well known that $\text{pr}(\theta \mid \mathcal{D}) = \text{pr}(\theta \mid \boldsymbol{S}_{obs})$. To generate a sample from the partial posterior distribution $\text{pr}(\theta \mid \boldsymbol{S}_{obs})$, ABC with rejection sampling proceeds by: (i) simulating N^{syn} values $\theta_l, l = 1, \ldots, N^{\text{syn}}$ from the prior distribution $\text{pr}(\theta)$; (ii) simulating datasets D_l from $\text{pr}(\mathcal{D}|\theta_l)$; (iii) computing summary statistics $\boldsymbol{S}_l, l = 1, \ldots, N^{\text{syn}}$ from $\mathcal{D}_l$; (iv) retaining a subset of $\{\theta_{l_s}, s = 1 \ldots, k\}$ of size $k < N^{\text{syn}}$ that corresponds to 'small' $\|\boldsymbol{S}_{l_s} - \boldsymbol{S}_{obs}\|$ values based on some threshold. Given pairs $\{(\theta_{l_s}, \boldsymbol{S}_{l_s})\}$, the task of estimating the partial posterior translates to a problem of conditional density estimation, e.g., based on Nadaraya-Waston type estimators and local regression adjustment variants to correct for the fact that $\boldsymbol{S}_{l_s}$ may not be exactly $\boldsymbol{S}_{obs}$; see Sisson et al. (2019) for a comprehensive review. To implement ABC, the choice of summary statistics is central.

We detail the specialization of ABC to the marginal posterior distributions of c and σ^2 in Section 2.3.1.1. Given any pair of (c, σ^2), we can sample trees from a density function up to an unknown normalizing constant based on an existing MH algorithm (Knowles and Ghahramani, 2015). Our proposal is to condition on the posterior median of (c, σ^2) of ABC-weighted samples from the first stage, when sampling the

trees in the second stage; clearly, other choices are also available. This strategy produced comparable MAP trees and inference of other tree-derived results relative to tree samples based on full ABC samples of c and σ^2.

Pseudo code for the two-stage algorithm is presented in the Supplementary Material Algorithm 2. We briefly describe below its key components.

2.3.1.1 Stage 1: Sampling Euclidean Parameters (c, σ^2) using ABC

Accuracy and efficiency of the ABC procedure is linked to two competing desiderata on the summary statistics: (i) informative, or ideally sufficient, and (ii) low-dimensional.

Summary statistic for σ^2. From the closed-form likelihood in Equation (2.5), a sufficient statistic of $\sigma^2\boldsymbol{\Sigma}^{\mathcal{T}}$ is easily available, using which we construct a summary statistics for σ^2.

Lemma 2.3.1. *With* $\mathbf{X}$ *as the observed data, the statistic* $\boldsymbol{T} := \sum_j \boldsymbol{X}_{\cdot,j}\boldsymbol{X}_{\cdot,j}^{\mathsf{T}}$ *is sufficient for* $\sigma^2\boldsymbol{\Sigma}^{\mathcal{T}}$ *and follows a Wishart distribution* $W_I(J, \sigma^2\boldsymbol{\Sigma}^{\mathcal{T}})$*, where* $\boldsymbol{X}_{\cdot,j} = [x_{1j}, \ldots, x_{Ij}] \in \mathbb{R}^I$*. Then with* $S^{(\sigma^2)} := \frac{tr(\boldsymbol{T})}{IJ}$ *we have* $E[S^{(\sigma^2)}] = \sigma^2$ *and* $Var[S^{(\sigma^2)}] = \frac{2\sigma^4 tr((\boldsymbol{\Sigma}^{\mathcal{T}})^2)}{I^2J}$.

Due to the normality of $\mathbf{X}$ in (2.5), and the Factorization theorem (Casella and Berger, 2001), we see that $\boldsymbol{T}$ is complete and sufficient for $\sigma^2\boldsymbol{\Sigma}^{\mathcal{T}}$ and $\boldsymbol{T} \sim W_I(J, \sigma^2\boldsymbol{\Sigma}^{\mathcal{T}})$. Well-known results about the trace and determinant of $\mathbf{X}$ (see for e.g. Mathai (1980)) provide the stated results on the mean and variance of $tr(\boldsymbol{T})$. Owing to its unbiasedness, we choose $S^{(\sigma^2)} = tr(\boldsymbol{T})/IJ$ as the summary statistic for σ^2 and examine its performance through simulations in Section 2.4; other choices are assessed in the Supplementary Material Section A.4.1.

Summary statistic for c. Based on the matrix normal distribution of Proposition 1, the divergence parameter c does not appear in the observed data likelihood. Any statistic based on the entire observed data set $\mathbf{X}$ is sufficient, but not necessarily

informative about c. In DDT, the prior distribution of the vector of branching times $\boldsymbol{t}$ is governed by divergence parameter c via the divergence function $a(t; c)$. Thus an informative summary statistic for c can be chosen by assessing its information about $\boldsymbol{t}$. For example, tighter observed clusters indicate small c (e.g., $c < 1$), where the level of tightness is indicated by the branch lengths from leaves to their respective parents. We construct summary statistics for c based on a dendrogram estimated via hierarchical clustering of $\mathbf{X}$ based on pairwise distances $\delta_{i,i'} := \|\boldsymbol{X}_i - \boldsymbol{X}_{i'}\|, i \neq i'$. The summary statistics $\boldsymbol{S}^{(c)}$ we choose is a ten-dimensional concatenated vector comprising the 10th, 25th, 50th, 75th and 90th percentiles of empirical distribution of: (i) $\delta_{i,i'}$; (ii) branch lengths associated with leaves of the dendrogram. Other candidate summary statistics for c are examined in Supplementary Material Section A.4.1.

2.3.1.2 Stage 2: Sampling Tree Parameters using Metropolis-Hastings

For the second stage, we proceed by choosing a representative value (c_0, σ_0^2) chosen from the posterior sample of (c, σ^2), which in our case is the posterior median. Then a Metropolis-Hastings (MH) algorithm to sample from $\text{pr}((\mathcal{T}, \boldsymbol{t})|c_0, \sigma_0^2, \mathbf{X})$; recall that the $\text{R}_\text{x}$tree is characterized by both the topology $\mathcal{T}$ and divergence times $\boldsymbol{t}$. In particular, after initialization (e.g., the dendrogram obtained from hierarchical clustering), we first generate a candidate tree $(\mathcal{T}', \boldsymbol{t}')$ from the current tree $(\mathcal{T}, \boldsymbol{t})$ in two steps: (i) detaching a subtree from the original tree; (ii) reattaching the subtree back to the remaining tree. Acceptance probabilities for a candidate tree can be computed exactly and directly using the explicit likelihood in (2.5), without which they would have to be calculated iteratively (Neal, 2003; Knowles and Ghahramani, 2015). See Supplementary Material Section A.2.2 for details of the proposal function and the acceptance probabilities.

Remark 1. In order to use the explicit likelihood in (2.5) from Proposition 1 to generate observed data $\mathbf{X}$, a tree-structured covariance $\boldsymbol{\Sigma}^{\mathcal{T}}$ needs to be specified,

whose entries in-turn depend on the parameter c through the divergence function. It is not straightforward to fix or sample a $\mathbf{\Sigma}^{\mathcal{T}}$ since its entries need to satisfy the inequalities (2.3). It is easier to generate data $\mathbf{X}$ directly using the DDT generative mechanism in the ABC stage, and this is the approach we follow and is described in Supplementary Section A.2.

Summarily, there are three main advantages to using the explicit likelihood from Proposition 1: (i) decoupling of Euclidean and tree parameters to enable an efficient two-stage sampling algorithm; (ii) direct and exact computation of tree acceptance probabilities in MH stage; (iii) determination of informative sufficient statistic for σ^2 (Lemma 2.3.1).

Remark 2. From the computational aspect, the calculation of the explicit Gaussian likelihood of (2.5) in Proposition 1 through the matrix decomposition is slower (e.g. Cholesky decomposition with $\mathcal{O}(I^3)$) than the message passing (e.g. the belief propagation with $\mathcal{O}(I)$ (Mezard and Montanari, 2009)) in terms of the big O notation (Knuth, 1976). However, the computation speed also depends on the implementation. For this Chapter, we implemented our algorithm in `R` and found that the matrix decomposition is faster than the message passing on `R`. We offer more details with a simulation study in Supplementary Material Section A.5.3.

2.3.2 Posterior Summary of $\mathbf{R_x}$-Tree, $(\mathcal{T}, \boldsymbol{t})$

While quantifying uncertainty concerning the tree parameters $(\mathcal{T}, \boldsymbol{t})$ is of main interest, we note that, from definition of the DDT, this is influenced by uncertainty in the model parameters. In particular, the first stage of ABC-MH produces weighted samples and we calculate the posterior median by fitting an intercept-only quantile regression with weights (see details in the Supplementary Material Section A.2.1). For the $\mathrm{R_x}$-tree, we consider global and local tree posterior summaries that capture uncertainty in the latent hierarchy among all and subsets of treatments.

Flexible posterior inference is readily available based on L posterior samples of $(\mathcal{T}, \boldsymbol{t})$ from the MH step. It is possible to construct correspond tree-structured covariance matrices $\boldsymbol{\Sigma}^{\mathcal{T}}$ from sample $(\mathcal{T}, \boldsymbol{t})$. Instead, we compute:

(a) a global *maximum a posteriori* (MAP) estimate of the $\mathrm{R_x}$-tree that represents the overall hierarchy underlying the treatment responses;

(b) local uncertainty estimates of co-clustering probabilities among a subset $\mathcal{A} \subset \{1, \ldots, I\}$ of treatments based on posterior samples of the corresponding subset of divergence times.

Posterior co-clustering probability functions. We elaborate on the local summary (b). Suppose $\mathcal{A} = \{i, i', i''\}$ consists of three treatments. Given a tree topology $\mathcal{T}$, note that at every $t \in [0, 1]$ a clustering of all I treatments is available and the clustering changes only at times $0 < t_1 < \cdots < t_{I-1}$. Consequently, for a given tree topology $\mathcal{T}$ drawn from its posterior, we can compute for every level $t \in [0, 1]$ a posterior probability that i, i' and i'' belong to the same cluster. Such a posterior probability can be approximated using Monte Carlo on the L posterior samples. Accordingly, we define the estimated posterior co-clustering probability (PCP) function associated with $\mathcal{A}$ as,

$$\mathrm{PCP}_{\mathcal{A}}(t) = \frac{\sum_{l=1}^{L} \mathbb{I}_{[0, t^{(l)}_{i,i',i''})}(t)}{L}, \tag{2.6}$$

where $\mathbb{I}_B$ is the indicator function on the set B and $t^{(l)}_{i,i',i''}$ is the divergence time of $\mathcal{A} = \{i, i', i''\}$ in the l-th tree sample. Essentially, the $\mathrm{PCP}_{\mathcal{A}}(t)$ can be viewed as the proportion of tree samples with $\{i, i', i''\}$ having the most recent common ancestor later than t.

For every subset $\mathcal{A}$, the function $[0, 1] \ni t \mapsto \mathrm{PCP}_{\mathcal{A}}(t) \in [0, 1]$ is non-increasing starting at 1 and ending at 0, and reveals propensity among treatments in $\mathcal{A}$ to cluster as one traverses down an (estimate of) $\mathrm{R_x}$-tree starting at the root: a curve

that remains flat and drops quickly near 1 indicates higher relative similarity among the treatments in $\mathcal{A}$ relative to the rest of the treatments. A scalar summary of $\text{PCP}_{\mathcal{A}}(t)$ is the area under its curve known as integrated PCP $\text{iPCP}_{\mathcal{A}}$, which owing to the definition of $\text{PCP}_{\mathcal{A}}(t)$, can be interpreted as the expected (or average) chance of co-clustering for treatments in $\mathcal{A}$.

Figure II.3 illustrates an example of a three-way $\text{iPCP}_{\mathcal{A}}$ with $\mathcal{A} = \{i, i', i''\}$ for a PDX data with I treatments and J patients (Figure II.3(A)). Given $L = 3$ posterior trees samples (Figure II.3(B)) drawn from the PDX data, we first calculate the whole $\text{PCP}_{\mathcal{A}}(t)$ function by moving the time t from 0 to 1. Starting from time $t = 0$, no treatment diverges at time $t = 0$ and the $\text{PCP}_{\mathcal{A}}(t)$ is 1. At time t', treatments diverge in one out of the three posterior trees and $\text{PCP}_{\mathcal{A}}(t)$ therefore drops from 1 to 2/3. Moving the time toward $t = 1$, treatments diverge in all trees and the $\text{PCP}_{\mathcal{A}}(t)$ drops to 0. The $\text{iPCP}_{\mathcal{A}}$ then can be obtained by the area under the $\text{PCP}_{\mathcal{A}}(t)$.

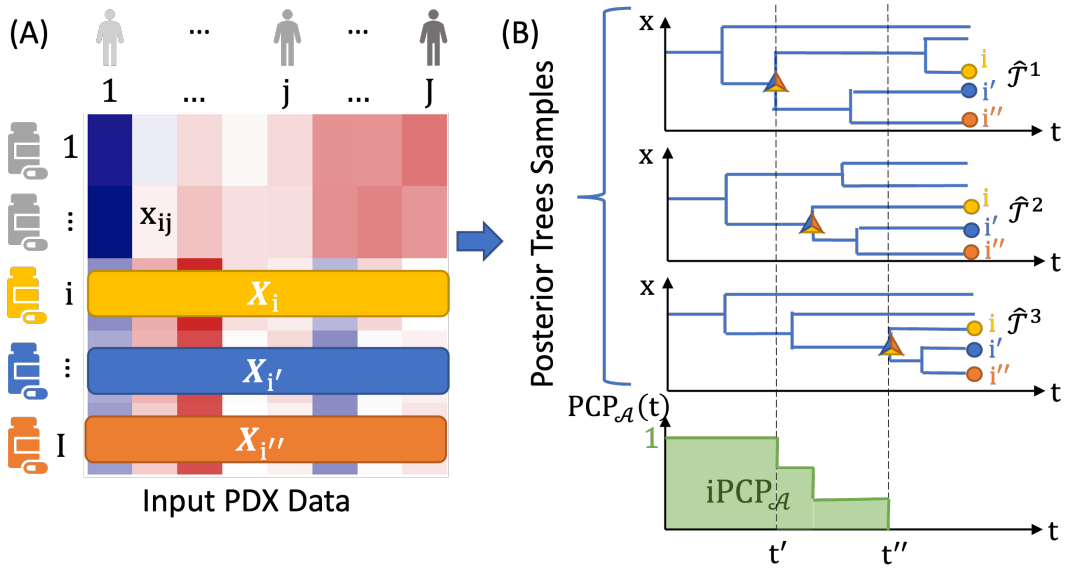


Figure II.3: Posterior tree summaries. (A) The input PDX data with I treatments and J patients, and treatments $\mathcal{A} = \{i, i', i''\}$ are of interest. (B) $\text{PCP}_{\mathcal{A}}(t)$ and $\text{iPCP}_{\mathcal{A}}$ for treatments $\mathcal{A}$ based on $L = 3$ posterior trees. The relevant divergence times are represented by a "△" in each posterior tree sample. For example, at time t', the treatments in $\mathcal{A}$ diverge in one out of the three trees. Because $\text{PCP}_{\mathcal{A}}(t')$ is defined by the proportion of posterior tree samples in which $\mathcal{A}$ has *not* diverged up to and including t', it drops from 1 to 2/3.

Remark 3. In the special case of $\mathcal{A} = \{i, i'\}$ for two treatments, the definition of $\text{iPCP}_{\mathcal{A}}$ can be related to the cophenetic distance (Sokal and Rohlf, 1962; Cardona et al., 2013) and, moreover, extends definition of the cophenetic distance to multiple trees. Given two treatments i and i' in a single tree, let t_d be the time at which their corresponding Brownian paths diverge. Then $\text{PCP}_{\mathcal{A}}(t) = \mathbb{I}_{[0,t_d)}(t)$ and $\text{iPCP}_{\mathcal{A}} = t_d$; this implies that the cophenetic distance is $2(1 - t_d)$ and thus $\text{iPCP}_{\mathcal{A}}$ and the cophenetic distances uniquely determines the same tree structure. For $L > 1$ trees, a Carlo average of divergence times of L trees leads to the corresponding $\text{iPCP}_{\mathcal{A}}$.

Remark 4. Given I treatments, since pairwise cophenetic distances from one tree determines a tree (Lapointe and Legendre, 1991; McCullagh, 2006), one might consider summarizing and represent posterior trees in terms of an $I \times I$ matrix $\mathbf{\Sigma}$ consisting of entries $\text{iPCP}_{\{i,i'\}}$ for every pair of treatments of (i, i'), estimated from the posterior sample of trees. However, $\mathbf{\Sigma}$ need not to be a tree-structured matrix that uniquely encodes a tree. It is possible to project $\mathbf{\Sigma}$ on to the space of tree-structured matrices (see for e.g., Bravo et al. (2009)) but the projection might result in a non-binary tree structure. We discuss this issue and its resolution in Supplementary Material Section A.3.

2.4 Simulations

Accurate characterization of similarities among any subset of treatments is central to our scientific interest in identifying the promising treatment subsets for further investigation. In addition, we have introduced a two-stage algorithm to improve our ability to efficiently draw tree samples from the posterior distribution (similarly for the Euclidean parameters). To demonstrate the modeling and computational advantages, we conduct two sets of simulations. The first simulation shows that the proposed model estimates the similarity (via iPCP) better than alternatives,

even when the true data generating mechanisms deviate from DDT assumptions in terms of the form of divergence function, prior distribution for the unknown tree, and normality of the responses. The second simulation illustrates the computational efficiency of the proposed two-stage algorithm in producing higher quality posterior samples of Euclidean parameters, resulting in more accurate subsequent estimation of an unknown tree and iPCPs, two key quantities to our interpretation of real data results.

2.4.1 Simulation I: Estimating Treatment Similarities

We first show that iPCPs estimated by DDT are closer to the true similarities (operationalized by functions of elements in the true divergence times in $\mathbf{\Sigma}^{\mathcal{T}}$) under different true data generating mechanisms that may follow or deviate from the DDT model assumptions in three distinct aspects (the form of divergence function, the prior distribution over the unknown tree, and normality).

Simulation setup. We simulate data by mimicking the PDX breast cancer data (see Section 2.5) with $I = 20$ treatments and $J = 38$ patients. We set the true scale parameter as the posterior median σ_0^2 and the true tree $\mathcal{T}_0$ as the MAP tree that are estimated from the breast cancer data; We consider four scenarios to represent different levels of deviation from the DDT model assumptions:

(i) No deviation of the true data generating mechanism from the fitted DDT models: given σ_0^2 and $\mathcal{T}_0$, simulate data based on the DDT marginal data distribution (Equation (2.5));

The true data generating mechanism deviates from the fitted DDT in terms of:

(ii) divergence function: same as in (i), but the true tree is a random tree from DDT with misspecified divergence function, $a(t; r) = \frac{r}{(1-t)^2}, r = 0.5$;

(iii) prior for tree topology: same as in (i), but the true tree is a random tree from

the coalescence model (generated by function `rcoal` in R package `ape`), and,

(iv) marginal data distribution: same as in (i), but the marginal likelihood is a centered multivariate t distribution with degree-of-freedom four and scaled by $\sigma_0^2 \mathbf{\Sigma}^{\mathcal{T}_0}$.

For each of four true data generating mechanisms above, we simulate $B = 50$ replicate data sets. In the following, we use the DDT model and the two-stage algorithm for all estimation regardless of the true data generating mechanisms. For DDT, we ran the two-stage algorithm where the second stage is implemented with five parallel chains. For each chain, we ran $10,000$ iterations, discarded first $9,000$ trees and combined five chains with a total of $5,000$ posterior tree samples.

First, we compute the iPCPs for all pairs of treatment combinations following the definition of $\text{iPCP}_{\mathcal{A}}$ where $\mathcal{A} = \{i, i'\}, 1 \leq i < i' \leq I$. Two alternative approaches to defining and estimating similarities between treatments are considered: (i) similarity derived from agglomerative hierarchical clustering, and (ii) empirical Pearson correlation of the two vectors of responses $\boldsymbol{X}_i$ and $\boldsymbol{X}_{i'}$, for $i \neq i'$. In particular, for (i), we considered five different linkage methods (Ward, Ward's D2, single, complete and Mcquitty) with Euclidean distances. Given an estimated dendrogram from hierarchical clustering, the similarity for a pair of treatments is defined by first normalizing the sum of branch lengths from the root to leaf as 1, and then calculating the area under of the co-clustering curve (AUC) obtained by cutting the dendrogram at various levels from 0 to 1. For three- or higher-way comparisons, (i) can still produce an AUC based on a dendrogram obtained from hierarchical clustering, while the empirical Pearson correlation in (ii) is undefined hence not viable as a comparator beyond assessing pairwise treatment similarities.

Performance metrics. For treatment pairs $\mathcal{A} = \{i, i'\}$, to assess the quality of estimated treatment similarities for each of the methods above (DDT-based, hierarchical-

clustering-based, and empirical Pearson correlation), we compare the estimated values against the true branching time $\Sigma^{\mathcal{T}_0}_{i,i'}$; similarly when assessing recovery of three-way treatment similarities, e.g., $\mathcal{A} = \{i, i', i''\}$, $\Sigma^{\mathcal{T}_0}_{i,i',i''}$ is defined as the time when $\{i, i', i''\}$ first branches in the true tree $\mathcal{T}_0$. In particular, for replication data set $b = 1, \ldots, B$, let $\widehat{\Sigma}^{(b)}_{i,i'}$ generically represent the pairwise similarities for treatment subsets (i, i') that can be based on DDT, hierarchical clustering or empirical pairwise Pearson correlation. For three-way comparisons, let $\widehat{\Sigma}^{(b)}_{i,i',i''}$ generically represent the three-way similarities for treatment subset (i, i', i'') that can be based on DDT, or hierarchical clustering.

We assess the goodness of recovery by computing $\sqrt{\sum_{i,i'}(\widehat{\Sigma}^{(b)}_{i,i'} - \Sigma^{\mathcal{T}_0}_{i,i'})^2}$, the Frobenious norm of the matrix in recovering the entire $\Sigma^{\mathcal{T}_0}$. We compute $\max_{i,i',i''} |\widehat{\Sigma}^{(b)}_{i,i',i''} - \Sigma^{\mathcal{T}_0}_{i,i',i''}|$, the max-norm of the matrix in recovering the true three-way similarities. For a given method and treatment subset $\mathcal{A}$, the above procedure results in B values, the distribution of which can be compared across methods; smaller values indicate better recovery of the true similarities.

Alternatively, for each method and each treatment subset, we also compute the Pearson correlation between the estimated similarities and the true branching times across replicates for pairwise or three-way treatment subsets:

$$\widehat{\text{Cor}}\left((\widehat{\Sigma}^{(b)}_{i,i'}, \Sigma^{\mathcal{T}_0}_{i,i'}), b = 1, \ldots, B\right);\ \widehat{\text{Cor}}\left((\widehat{\Sigma}^{(b)}_{i,i',i''}, \Sigma^{\mathcal{T}_0}_{i,i',i''}), b = 1, \ldots, B\right),$$

for $B = 50$ and treatments $i < i' < i''$. We refer to this metric as "Correlation of correlations" (the latter uses the fact that the entries in the true $\Sigma^{\mathcal{T}_0}$ being correlations; see Equation (2.5)); higher values indicate better recovery of the true similarities.

Simulation results. We observe that DDT better estimates the treatment similarities even under misspecified models. In particular, under scenarios where the true data generating mechanisms deviate from the fitted DDT model assumptions (ii-iv),

the DDT captures the true pairwise and three-way treatment similarities the best by higher values in correlation of correlations (left panels, Figure II.4) and lower matrix/array distances (right panels, Figure II.4). In particular, the fitted DDT with divergence function $a(t) = c/(1-t)$ under Scenario i, ii and iii performed similarly well indicating the relative insensitivity to the DDT modeling assumptions with respect to divergence function and the tree generative model. Under Scenario iv where the marginal likelihood assumption deviates from Gaussian with heavier tails, the similarity estimates from all methods deteriorate relative to Scenarios i-iii. Comparing between methods, the similarities derived from hierarchical clustering with single linkage is comparable to DDT model when evaluated by correlation of correlation, but worse than DDT when evaluated by the matrix norm.

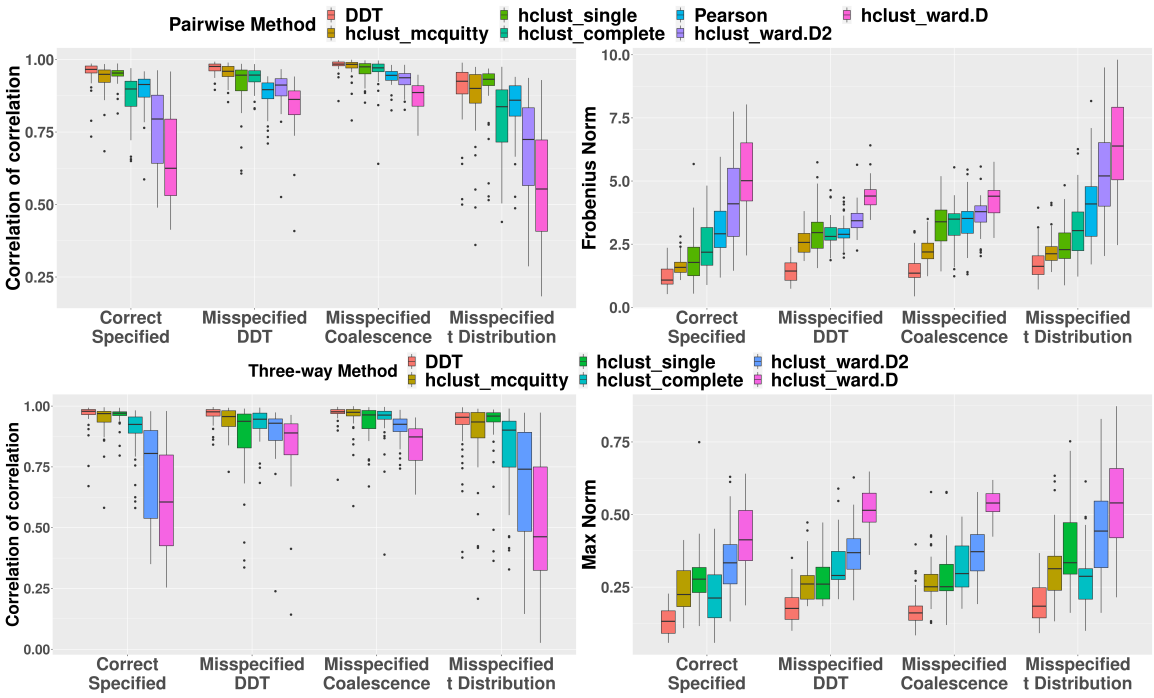


Figure II.4: Simulation studies for comparing the quality of estimated treatment similarities based on DDT, hierarchical clustering, and empirical Pearson correlation. Two performance metrics are used: (Left) Correlation of correlation (higher values are better); (Right) Matrix distances with Frobenius norm for pairwise similarity and max norm for three-way similarity (lower values are better). DDT captures both true pairwise (upper panels) and three-way (lower panels) similarity best under four levels of misspecification scenarios.

Additional simulations. Another alternative to bring the information of the posterior samples of c and σ^2 is to use the whole posterior samples instead of the fixed representative statistics only. Following the same set-up, we offer another simulation result to empirically compare the inference performance from the algorithm with the posterior median only and the the whole posterior samples. See more details in Supplementary Material Section A.5.4.

2.4.2 Simulation II: Comparison with Single-Stage MCMC Algorithms

We have also conducted extensive simulation studies that focus on the computational aspect of the proposed algorithms and demonstrate the advantage of the proposed two-stage algorithm in producing higher quality posterior samples of the unknown tree than classical single-stage MCMC algorithms. In particular, we demonstrate that the proposed algorithm produces (i) MAP trees that are closer to the true tree than alternatives (hierarchical clustering, single-stage MH with default hierarchical clustering or the true tree at initialization) and (ii) more accurate estimation of pairwise treatment similarities compared to single-stage MCMC algorithms. See Supplementary Material Section A.5 for further details.

Additional simulations and sensitivity analyses. Aside from the simulations above focusing on the tree structure and the divergence time, Supplementary Material A.4 offers additional details for Euclidean parameters including the parameter inference, algorithm diagnostics, and sensitivity analysis for the number of the synthetic data. In particular, we empirically show that current $\boldsymbol{S}^{(c)}$ and $S^{(\sigma^2)}$ outperform other candidate summary statistics in terms of bias in Section A.4.1. In Section A.4.2, we present additional simulation results that demonstrate that the two-stage algorithm (i) enjoys stable effective sample size (ESS) for (c, σ^2); (ii) leads to similar or better inference on (c, σ^2), as ascertained using credible intervals. In Section A.4.3, we check the convergence of MH and the goodness of fit for ABC. A sensitivity analysis for the

number of the synthetic data providing the possible acceleration for ABC is shown in Section A.4.3.3.

2.5 Treatment Trees in Cancer using PDX Data

2.5.1 Dataset Overview and Key Scientific Questions

We leverage a recently collated PDX dataset from the Novartis Institutes for BioMedical Research - PDX Encyclopedia [NIBR-PDXE, (Gao et al., 2015)] that interrogated multiple targeted therapies across different cancers and established that PDX systems provide a more accurate measure of the response of a population of patients than traditional preclinical models. Briefly, the NIBR-PDXE consists of $> 1,000$ PDX lines across a range of human cancers and uses a $1 \times 1 \times 1$ design (one animal per PDX model per treatment); i.e., each PDX line from a given patient was treated simultaneously with multiple treatments allowing for direct assessments of treatment hierarchies and responses. In this Chapter, we focus on our analyses on a subset of PDX lines with complete responses across five common human cancers: Breast cancer (BRCA), Cutaneous Melanoma (CM, skin cancer), Colorectal cancer (CRC), Non-small Cell Lung Carcinoma (NSCLC), and Pancreatic Ductal Adenocarcinoma (PDAC). After re-scaling data and missing data imputation, different numbers of treatments, I, and PDX models, J, presented in the five cancers were, (I, J): BRCA, $(20, 38)$; CRC, $(20, 40)$; CM, $(14, 32)$; NSCLC, $(21, 25)$; and PDAC, $(20, 36)$. (See Supplementary Material Table A.9 for treatment names and Section A.6.1 for details of pre-processing procedures.)

In our analysis, we used the best average response (BAR) as the main response, by taking the untreated group as the reference group and using the tumor size difference before and after administration of the treatment(s) following Rashid et al. (2020). Positive values of BAR indicate the treatment(s) shrunk the tumor more

than the untreated group with higher values indicative of (higher) treatment efficacy. To apply the Proposition 1, we also checked the distributional assumption for each cancer (see Supplementary Material Section A.6.2). The treatments included both drugs administered individually with established mechanisms (referred to as "monotherapy") and multiple drugs combined with potentially unknown synergistic effects (referred to as "combination therapy"). Our key scientific questions were as follows: (a) identify plausible biological mechanisms that characterize treatment responses for monotherapies within and between cancers; (b) evaluate the effectiveness of combination therapies based on biological mechanisms. Due to a potentially better outcome and lower resistance, combination therapy with synergistic mechanism is highly desirable (Bayat Mokhtari et al., 2017).

DDT model setup. For all analyses we followed the setup in the Section 2.4.1 and obtained $N^{\text{syn}} = 600,000$ synthetic datasets from the ABC algorithm (Section 2.3.1.1) with prior $c \sim \mathsf{Gamma}(2,2)$ and $1/\sigma^2 \sim \mathsf{Gamma}(1,1)$ and took the first 0.5% ($d = 0.5\%$) closest data in terms of $\boldsymbol{S}^{(c)}$ and $S^{(\sigma^2)}$. We calculated the posterior median of (c, σ^2) as described in Section 2.3.2. For the second-stage MH, we ran five chains of the two-stage algorithm with (c, σ^2) fixed at the posterior median by 10,000 iterations and discarded the first 9,000 trees, which resulted in 5,000 posterior trees in total. Finally, we calculated the R_x-tree (MAP) and iPCP based on 5,000 posterior trees for all subsequent analyses and interpretations. All computations were divided on multiple different CPUs (see the Supplementary Table A.7 for the full list of CPUs). For the BRCA data with $I = 20$ and $J = 38$, we divided the ABC stage into 34 compute cores with a total of 141 CPU hours and maximum 4.7 hours in real time. For the MH stage and the single-stage MCMC, we split the computation on 5 compute cores with a total of 8.6 and 12 CPU hours, and a maximum 1.7 and 2.5 hours in real time, respectively.

Our results are organized as follows: we provide a summary of the R_x-tree es-

timation and treatment clusters in Section 2.5.2 followed by specific biological and translational interpretations in Sections 2.5.3 and 2.5.4 for monotherapy and combination therapy, respectively. Additional results can be accessed and visualized using our companion `R`-shiny application (see Supplementary Material Section A.6.6 for details).

2.5.2 R_x-Tree Estimation and Treatment Clusters

We focus our discussion on three cancers: BRCA, CRC and CM here – see Supplementary Materials Section A.6.5 for NSCLC and PDAC. In Figure II.5, R_x-tree, pairwise iPCP and (scaled) Pearson correlation are shown in the left, middle and right panels, respectively. Focusing on the left two panels, we observe that the R_x-tree and the pairwise iPCP matrix show the similar clustering patterns. For example, three combination therapies in CM form a tight subtree and are labeled by a box in the R_x-tree of Figure II.5 and a block with higher values of iPCP among three combination therapies also shows up in the corresponding iPCP matrix with a box labeled. In our analysis, the treatments predominantly target six oncogenic pathways that are closely related to the cell proliferation and cell cycle: (i) phosphoinositide 3-kinases, PI3K; (ii) mitogen-activated protein kinases, MAPK; (iii) cyclin-dependent kinases, CDK; (iv) murine double minute 2, MDM2; (v) janus kinase, JAK; (vi) serine/threonine-protein kinase B-Raf, BRAF. We label targeting pathways above for monotherapies with solid dots and further group PI3K, MAPK and CDK due to the common downstream mechanisms (e.g., Repetto et al., 2018; Kurtzeborn et al., 2019). Roughly, the R_x-tree from our model clusters monotherapies targeting oncogenic processes above and largely agrees with common and established biology mechanisms. For example, all PI3K-MAPK-CDK inhibitors (solid square) belong to a tighter subtree in three cancers; two MDM2 monotherapies (solid triangle) are closest in both BRCA and CRC. While visual inspection of the MAP R_x-tree agrees with known biology, iPCP

further quantifies the similarity by assimilating the information across multiple trees from our MCMC samples. For the ensuing interpretations in Sections 2.5.3 and 2.5.4, we focus on iPCP and verify our model through monotherapies with known biology, since our a priori hypothesis is that monotherapies that share the same downstream pathways should exhibit higher iPCP values. Furthermore, we extend our work to identify combination therapies with synergy and discover several combination therapies for each cancer.

2.5.3 Biological Mechanisms in Monotherapy

Our estimation procedure exhibits a high level of concordance between known biological mechanisms and established monotherapies for multiple key signalling pathways. From the R_x-tree in Figure II.5, aside from the oncogenic process (solid dots) introduced above, monotherapies also target receptors (hollow circles) or other non-kinase targets (e.g. tubulin; crosses). We summarize our key findings and interpretations along with their implications in monotherapy across different cancers for PI3K-MAPK-CDK in this section and list the rest signaling pathways and their regulatory axes, namely, MDM2 from cell cycle regulatory pathways, human epidermal growth factor receptor 3 (ERBB3) from receptor pathways, and tubulin from non-kinase pathways in Supplementary Material Section A.6.4. For the following sections, because we wish to conduct fully-exploratory analyses where we do not assume prior knowledge about treatment mechanism, we set the threshold of the co-clustering at the 75-th percentile of all pairwise iPCPs. Specifically, we set the cut-off at 0.753, 0.687 and 0.801 for BRCA, CRC and CM, respectively. See Supplementary Material Section A.6.3 for more details about cut-off choices under full and partially exploratory settings related to prior knowledge about monotherapies.

PI3K-MAPK-CDK inhibitors. For treatments targeting PI3K, MAPK and CDK, treatments have the same target share high iPCP. In the NIBR-PDXE dataset, three

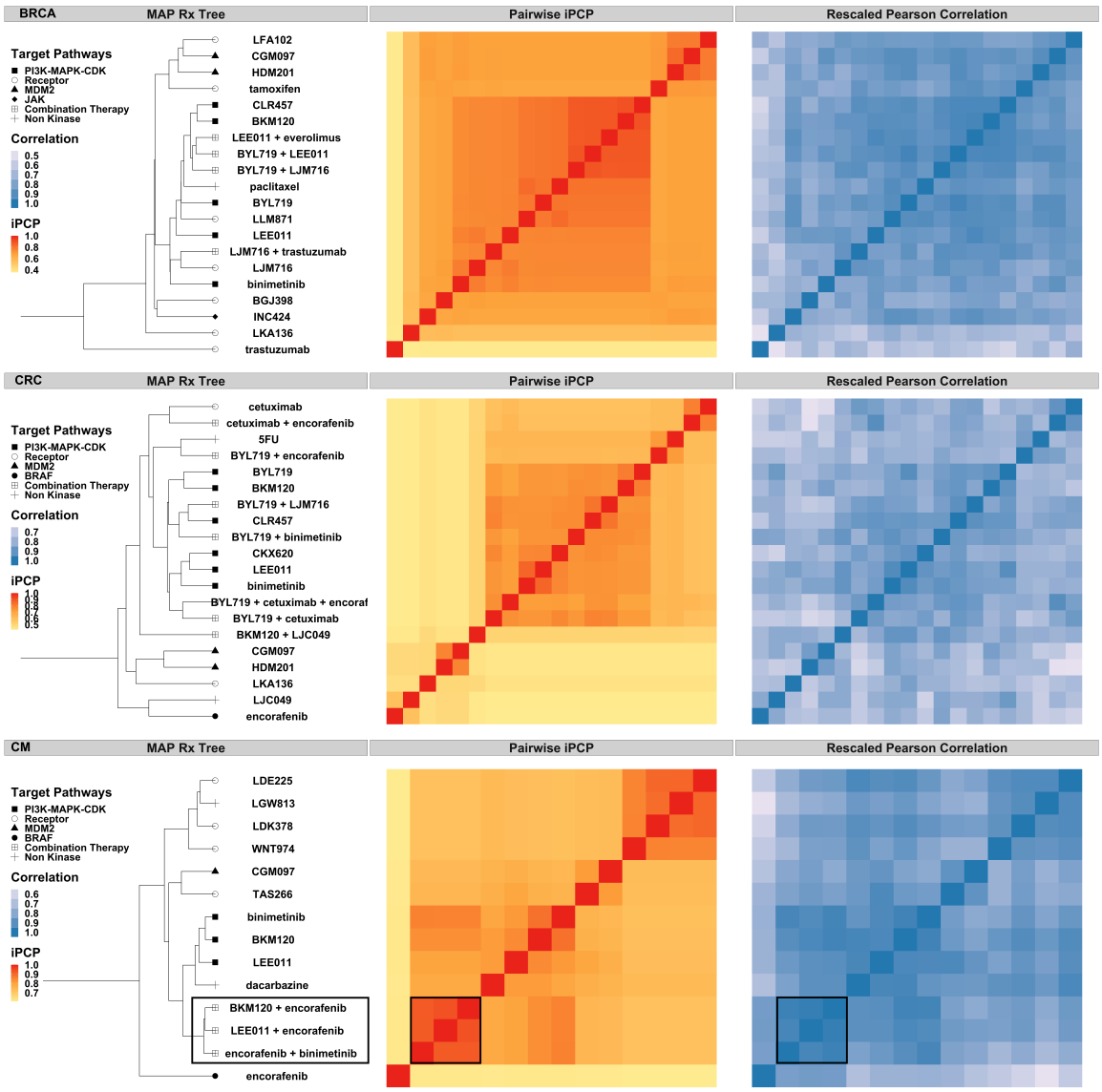


Figure II.5: The R_x-tree and iPCP for breast cancer (BRCA, top row), colorectal cancer (CRC, middle row) and melanoma (CM, lower row). Three panels in each row represent: (left) estimated R_x-tree (MAP); distinct external target pathway information is shown in distinct shapes for groups of treatments on the leaves; (middle) estimated pairwise iPCP, i.e., the posterior mean divergence time for pairs of entities on the leaves (see the result paragraph for definition for any subset of entities); (right) scaled Pearson correlation for each pair of treatments. Note that the MAP visualizes the hierarchy among treatments; the iPCP is not calculated based on the MAP, but based on posterior tree samples (see definition in Section 2.3.2)

PI3K inhibitors (BKM120, BYL719 and CLR457), two MAPK inhibitors (binimetinib and CKX620) and one CDK inhibitor (LEE011) were tested, but different cancers contain different numbers of treatments. Specifically, all three PI3K inhibitors present

in BRCA and CRC, but only BKM120 is tested in CM; CRC contains two MAPK inhibitors while BRCA and CM only have binimetinib; LEE011 is tested in all three cancers. In Figure II.6, BKM120, BYL719 and CLR457 share high pairwise iPCPs (box (1)) and all target PI3K for BRCA and CRC (BRCA, (BKM120, CLR457): 0.8986, (BKM120, BYL719): 0.8002, (BYL719, CLR457): 0.8002; CRC, (BKM120, CLR457): 0.7555, (BKM120, BYL719): 0.8041, (BYL719, CLR457): 0.7597); MAPK (box (2)) inhibitors, binimetinib and CKX620, show a high pairwise iPCP in CRC (0.7792). Asides from the pairwise iPCPs, our model also suggests high multi-way iPCPs among PI3K inhibitors in BRCA (0.8002) and CRC (0.7513). Among these inhibitors, PI3K inhibitor of BYL719 was approved by FDA for breast cancer; MAPK inhibitor of binimetinib was approved by FDA for BRAF mutant melanoma in combination with encorafenib; and CDK inhibitor of LEE011 was approved for breast cancer.

Our model suggests treatments targeting different pathways also share high iPCP values across different cancers. Monotherapies targeting different cell cycle regulatory pathways (PI3K, MAPK and CDK) exhibit high iPCPs. CDK inhibitor, LEE011, and MAPK inhibitors share high pairwise iPCP values in BRCA ((LEE011, binimetinib): 0.7709), CRC ((LEE011, binimetinib): 0.8617, (LEE011, CKX620): 0.7820) and CM ((LEE011, binimetinib): 0.8210) in the Figure II.6 with box (3). High iPCP among MAPK and CDK inhibitors agree with biology, since it is known that CDK and MAPK collaboratively regulate downstream pathways such as Ste5 (Repetto et al., 2018). High pairwise iPCP values between PI3K and MAPK inhibitors were observed in box (3) in the Figure II.6. Specifically, our model suggests high pairwise iPCPs as follows: (i) BRCA, (binimetinib, BKM120): 0.7427, (binimetinib, BYL719): 0.7441, (binimetinib, CLR457): 0.7427)); (ii) CRC, (binimetinib, BKM120): 0.7374, (binimetinib, BYL719): 0.7388, (binimetinib, CLR457): 0.7541, (CKX620, BKM120): 0.7366, (CKX620, BYL719): 0.7357, (CKX620, CLR457):

0.7676)); (iii) CM, (binimetinib, BKM120): 0.8882. Aside from the pairwise iPCPs above, high multi-way iPCPs in BRCA (0.7422), CRC (0.7300) and CM (0.8882) also show the similar information. From the existing literature, both PI3K and MAPK can be induced by ERBB3 phosphorylation (Balko et al., 2012) and it is not surprising to see high iPCPs between PI3K and MAPK inhibitors.

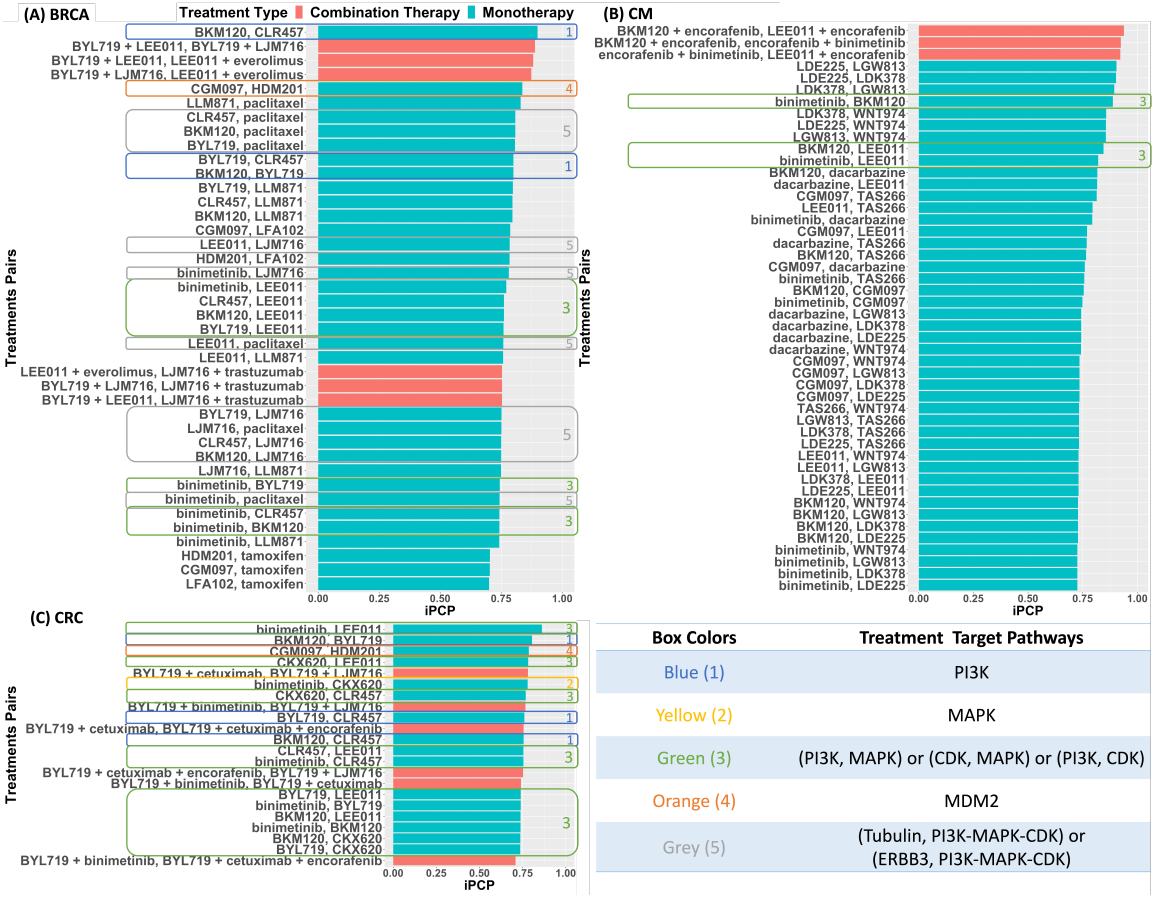


Figure II.6: Bar plot of iPCPs for pairs of combination therapies (red bars) and pairs of monotherapies (green bars): (A) breast cancer, (B) colorectal cancer and (C) melanoma. The bar plots are sorted by the iPCP values (high to low); pairs of treatments are shown only if the estimated iPCP is greater than 0.7. Monotherapies have different known targets which are listed in the bottom-right table (see Section 2.5.3 for more details and discussion on monotherapies).

2.5.4 Implications in Combination Therapy

Based on the concordance between the monotherapy and biology mechanism, we further investigate combination therapies to identify mechanisms with synergistic effect. In NIBR-PDXE, 21 combination therapies were tested and only one of them includes three monotherapies (BYL719 + cetuximab + encorafenib in CRC) and the rest contain two monotherapies. Out of 21 combination therapies, only three do not target any cell cycle (PI3K, MAPK, CDK, MDM2, JAK and BRAF) pathways (see Supplementary Material Table A.10 for the full list of combination therapies). From the R_x-tree in Figure II.5, combination therapies tend to form a tighter subtree and are closer to monotherapies targeting PI3K-MAPK-CDK, which implies that the mechanisms under combination therapies are similar to each other and are closer to the PI3K-MAPK-CDK pathways. We identified several combination therapies with known synergistic effects and provide a brief description for each of the cancers in the following paragraphs.

Breast cancer. Four combination therapies were tested in BRCA and three therapies targeting PI3K-MAPK-CDK (BYL719 + LJM716, BYL719 + LEE011 and LEE011 + everolimus) form a subtree in R_x-tree with a high three-way iPCP (0.8719). Among these combination therapies, PI3K-CDK inhibitor, BYL719 + LEE011, suggests a possible synergistic regulation (Vora et al., 2014; Bonelli et al., 2017; Yuan et al., 2019). Based on the high iPCP between BYL719 + LEE011 and the rest two therapies, we suggest synergistic effect for combination therapies targeting PI3K-ERBB3 (BYL719 + LJM716), and CDK-MTOR (LEE011 + everolimus) for future investigation.

Colorectal cancer. Our model suggests a high three-way iPCP (0.7437) among PI3K-EGFR (BYL719 + cetuximab), PI3K-EGFR-BRAF (BYL719 + cetuximab + encorafenib) and PI3K-ERBB3 (BYL719 + LJM716) inhibitors. Since the triple therapy (BYL719 + cetuximab + encorafenib) enters the phase I clinical trial with synergy

(Geel et al., 2014), our model proposes the potential synergistic effect for PI3K-ERBB3 based on iPCP for future investigation. Of note, we found a modest iPCP (0.6280) between the FDA-approved combination therapy EGFR-BRAF (cetuximab + encorafenib) and PI3K-EGFR-BRAF (BYL719 + cetuximab + encorafenib) and the modest iPCP can be explained by an additional drug-drug interaction between BYL719 and encorafenib in triple-combined therapy (van Geel et al., 2017).

Melanoma. In NIBR-PDXE, three combination therapies were tested in CM, and all of them consist one monotherapy targeting PI3K-MAPK-CDK and the other one targeting BRAF. A tight subtree is observed in the R_x-tree and our model also suggests a high iPCP (0.9222) among three combination therapies. Since PI3K, MAPK and CDK work closely and share a high iPCP (0.8204) among monotherapies in CM, a high iPCP (0.9222) among three combination therapies is not surprising. Since two combination therapies of BRAF-MAPK (dabrafenib + trametinib and encorafenib + binimetinib) are approved by FDA for BRAF-mutant metastatic melanoma (Dummer et al., 2018a,b; Robert et al., 2019), we recommend the synergy for BRAF-PI3K (encorafenib + BKM120) and BRAF-CDK (encorafenib + LEE011) inhibitors.

Comparison to alternative approaches. Unlike the probabilistic generative modeling approach proposed in this Chapter, standard distance-based agglomerative hierarchical clustering and Pearson correlation can also be applied to the PDX data to estimate the similarity. However, simple pairwise similarities can be potentially noisy and the uncertainty in the estimation is not fully incorporated due to the absence of a generative model. As we showed in the Section 2.4.1 (Simulation I) that agglomerative hierarchical clustering and the Pearson correlation leads to inferior recovery of the true branching times and the true tree structure under different data generating mechanisms mimicking the real data. As further evidence, we compute pairwise similarities based on Pearson correlation (other distance metrics show similar patterns) in the right panel of Figure II.5. By mapping the original Pearson correlation $\rho \in [-1, 1]$

through a linear function $\frac{\rho+1}{2}$, we make the range of iPCP and Pearson correlation comparable. We observe that pairwise iPCP estimated through the DDT model is less noisy than Pearson correlation. For example, both iPCP and Pearson correlation in CM show higher similarities among combination therapy framed by a box, but iPCP exhibits a clearer pattern than Pearson correlation.

2.6 Summary and Discussion

In translational oncology research, PDX studies have emerged as a unique study design that evaluates multiple treatments when applied to samples from the same human tumor implanted into genetically identical mice. PDX systems are promising tools for large-scale screening to evaluate a large number of FDA-approved and novel cancer therapies. However, there remain scientific questions concerning how distinct treatments may be synergistic in inducing similar efficacious responses, and how to identify promising subsets of treatments for further clinical evaluation. To this end, in this Chapter, we propose a probabilistic framework to learn treatment trees (R_x-trees) from PDX data to identify promising treatment combinations and plausible biological mechanisms that confer synergistic effect(s). In particular, in a Bayesian framework based on the Dirichlet Diffusion Tree, we estimate a *maximum a posteriori* rooted binary tree with the treatments on the leaves and propose a posterior uncertainty-aware similarity measure (iPCP) for any subset of treatments. The divergence times of the DDT encode the tree topology and are profitably interpreted within the context of an underlying plausible biological mechanism of treatment actions.

From the class of probabilistic models with an unknown tree structure component, we have chosen the DDT mainly owing to the availability of a closed-form marginal likelihood that directly links the tree topological structure to the covariance structure of the observed PDX data, which additionally decouples the Euclidean and tree parameters; to the best of our knowledge this method has not been proposed or ex-

plored hitherto for the DDT. The decoupling leads to efficient posterior inference via a two-stage algorithm that confers several advantages. The algorithm generates posterior samples of Euclidean parameters through approximate Bayesian computation and passes the posterior medians to a second stage classical Metropolis-Hastings algorithm for sampling from the conditional posterior distribution of the tree given all other quantities. Through simulation studies, we show that the proposed two-stage algorithm generates better posterior tree samples and captures the true similarity among treatments better than alternatives such as single-stage MCMC and naive Pearson correlations. The posterior samples of trees are summarized by iPCP, which we propose to measure the empirical mechanistic similarity for multiple treatments incorporating uncertainty.

Using the proposed methodology on NIBR-PDXE data, we estimate R_x-trees and iPCPs for five cancers. Among the monotherapies, iPCP is highly concordant with known biology across different cancers. For example, BKM120 and BYL719 show a high iPCP value among treatments in breast and colorectal cancer, which corroborates known mechanisms, since both monotherapies target the same biological pathway, PI3K, and BYL719 was approved by FDA for breast cancer. The proposed iPCP can also suggest improvements upon an existing combination therapy. We first identify a combination therapy with known synergy (not based on the our data) and then determine which additional therapies (monotherapies or combination therapies) have high iPCPs when considered together with the existing combination therapy. Based on the NIBR-PDXE data, for each cancer, we suggest potential synergies between PI3K-ERBB3 and CDK-MTOR for breast cancer, PI3K-ERBB3 for colorectal cancer, and BRAF-PI3K and BRAF-CDK for melanoma that could be potentially explored in future translational studies.

Our current analysis infers treatment trees based on the drug responses from the NIBR-PDXE dataset which provides treatment similarity information that may be

used to guide potential treatment strategies. However, there are a few limitations. First, the PDX experiments may fail to capture the difference in the microenvironment between the human and the immunodeficient mouse (Dobrolecki et al., 2016), which must be considered in disease contexts when findings are generalized to human. As PDX technology matures, this can be compensated by better PDX experiments that capture the tumor microenvironment more precisely. For example, one can use the genetically engineered mice to reconstruct the human immune system (Abdolahi et al., 2022), and some studies have started to adapt this method in the context of immunotherapies (Zhao et al., 2018). Second, on experimental design, current literature points to the potential advantage of designs with multiple animals per treatment and patient (Abdolahi et al., 2022). We can incorporate the random effects in the current model of (2.4) for the multiple-animal-per-patient design and we refer the reader to the Supplementary Material Section A.7 for more details. Also, to evaluate PDX designs with fewer treatments and patients that is common in co-clinical trials (e.g., Koga and Ochiai, 2019), we conducted a simulation for two datasets with a smaller dimension ($(I, J) = (5, 5)$ and $(10, 15)$) which confirmed the advantage of the proposed method in terms of recovering treatment similarities (see Supplementary Material Section A.5.5). Finally, from a statistical perspective, we have assumed independent patients without using the underlying patient-specific genomic information that is also available in the NIBR-PDXE. By including patient-specific genomic information, we may further improve our ability to identify synergistic treatments that may be specific to a subset of patients. One approach to utilizing genomic information could be to extend the DDT model to incorporate patient-specific genomic information in the mean structure or the column covariance of the marginal likelihood of Equation (2.4). In addition, models with non-Gaussian marginal likelihood and non-binary treatment tree in principle can be defined by considering generative tree models based on general diffusion processes (Heaukulani et al., 2014; Knowles and Ghahramani, 2015).

Both extensions raise significant, non-trivial methodological and computation issues (e.g., deriving tractable likelihoods; finding low-dimensional summary statistics for new parameters) and constitute the foundation for future work.

Code and data availability We also provide a general purpose code in `R` that accompanies this manuscript along with all the necessary documentation and datasets required to replicate our results (see `https://github.com/bayesrx/RxTree`). Furthermore, to aid access and visualization of the results, we have also developed an `R`-shiny application (see Supplementary Material Section A.6.6).

CHAPTER III

Geometry-driven Bayesian Inference for Ultrametric Covariances

3.1 Introduction

Ultrametric matrices are central to a multitude of machine learning and scientific applications. For instance, in a multivariate Gaussian distribution, the covariance matrix is an ultrametric matrix if and only if the Gaussian density is multivariate totally positive of order two (Karlin and Rinott, 1983; Lauritzen et al., 2019), which implies a conditional positive dependency between two random variables (Fallat et al., 2017). Recently, ultrametric matrices have been applied in various scenarios as covariance matrices in Gaussian distributions, such as graphical models (Fallat et al., 2017) and Brownian motion tree models (e.g. Neal, 2003; Sturmfels et al., 2021), with applications in cancer biology (Yao et al., 2023) and finance (Agrawal et al., 2020). However, due to the inequalities required on ultrametric matrices, the geometry of the space of ultrametric matrices is non-trivial, as it is neither a manifold (McCullagh, 2006) nor a convex set (Chierchia and Perret, 2020). As a result, challenges lie in both inference and computation, leading existing methods to primarily focus on point estimation without uncertainty quantification. In this paper, we characterize the geometry of the set of ultrametric matrices and develop a flexible Bayesian frame-

work to obtain posterior samples of ultrametric matrices efficiently, thereby providing uncertainty quantification alongside point estimates.

An ultrametric matrix is a square matrix with non-negative elements that satisfies the ultrametric inequality (Dellacherie et al., 2014). When the diagonal elements are all positive, the matrix is called a strictly ultrametric matrix and guarantees positive definiteness (Nabben and Varga, 1994). In the context of covariance, strictly ultrametric matrices are of interest as they ensure positive definiteness. However, the ultrametric inequality imposes a special structure on the matrix elements and entails challenging constraints on the space of positive definite matrices. Specifically, consider off-diagonal elements in a covariance matrix of dimension three by three. The ultrametric inequality requires that at least two elements be the same with the third element being equal or bigger. Consequently, the space of ultrametric matrices is embedded in a higher-dimensional space of positive definite matrices, represented as a simplicial cone contained in the spectrahedron (Sturmfels et al., 2021). Moreover, to address the inequality and the resulting geometry, only projection- (e.g. Bravo et al., 2009) and relaxation-based (e.g. Lauritzen et al., 2019) estimation methods exist.

While directly tackling the inequality and the geometry of the space of ultrametric matrices is difficult, the same set of inequalities determines a bijection between a (strictly) ultrametric matrix and a rooted tree structure and (Dellacherie et al., 2014; Steel, 2016). This bijection allows us to characterize the structure of the space by leveraging the geometry and the coordinate system of tree space introduced by Billera-Holmes-Vogtmann (BHV) (Billera et al., 2001). Specifically, our proposed algorithm makes efficient local moves on the BHV space along geodesics between neighboring tree topologies, resulting in efficient sampling and posterior matrices that automatically satisfy the ultrametric inequalities. These moves do not rely on projection or relaxation during posterior computation. Therefore, our algorithm allows for straightforward posterior summaries of both central tendency and dispersion

using the Fréchet mean (Miller et al., 2015) and geodesic distance (Owen and Provan, 2011) in the BHV space.

Most existing approaches for ultrametric matrix estimation treat the problem as an optimization task constrained by a set of ultrametric inequalities. However, the constrained objective function is highly non-convex (Chierchia and Perret, 2020) for standard optimization algorithms. To satisfy the ultrametric inequality, various optimization techniques are employed. For example, Bravo et al. (2009) uses a mixed-integer programming formulation and projects the sample covariance onto the space of ultrametric matrices. Similarly, Lauritzen et al. (2019) and Agrawal et al. (2020) relax the constraints and address the optimization problem through a dual problem. Additionally, Chierchia and Perret (2020) circumvent the constraint by using subdominant ultrametricity and the min-max operator. However, without additional projection or relaxation, all these methods can not satisfy the ultrametric inequalities. Moreover, they fail to estimate the matrices geodesically, which is essential for uncertainty quantification.

By leveraging a bijection between the labelled, rooted tree structure and ultrametric matrices, our proposed makes main three contributions. First, we *define a geometry for the space of ultrametric matrices* by relating an existing decomposition on an ultrametric matrix to coordinate of a point in the BHV tree space. Second, we define *a general consistent Markovian prior on the set of ultrametric matrices*, which includes several existing priors on the tree structure as special cases. Third, we devise *an efficient algorithm to draw posterior samples* that makes local moves geodesically on the BHV tree space.

The rest of the Chapter is organized as follows: we introduce the characterization of the ultrametric matrix space and decomposition of the ultrametric matrix in the tree space in Section 3.2. Section 3.3 and Section 3.4 delineates a general prior on the ultrametric matrix and the posterior inference via Metropolis-Hasting algorithm,

respectively. In Section 3.5, we conduct a series of simulations to evaluate our algorithm in terms of the matrix recovery with the uncertainty quantification. Section 3.6 demonstrate the utility of the proposed method with an pre-clinical data analysis for potential cancer treatment. The paper concludes by discussing implications of the findings, limitations, and future directions in Section 3.7. A general purpose code in `R` with packages and datasets for the proposed method is also provided on `https://github.com/bayesrx/ultrametricMat`.

3.2 Ultrametric Matrices and their Geometry

3.2.1 Bijection of the Ultrametric Matrix and the Tree Structures

Consider p-dimensional continuous random vectors $X_i = (X_{i1}, \ldots, X_{ip})^\mathsf{T}$ with the covariance matrix $\Sigma^T = \{\sigma_{j,k}\}, j, k = 1, \ldots, p$ for all $i = 1, \ldots, n$. We call Σ^T a strictly ultrametric matrix if Σ^T has positive diagonal elements $\sigma_{j,j} > 0$ and satisfies the following conditions:

$$\sigma_{j,j} \geq \sigma_{j,k} \geq 0, \text{ and } \sigma_{j,k} \geq \min\{\sigma_{j,h}, \sigma_{k,h}\}, \text{ for all } j \neq k \neq h. \tag{3.1}$$

We consider only strictly ultrametric matrices and simply refer to such matrices as ultrametric. The first condition ensures that the variable j is more similar to itself than any other variables. The second condition is referred to as the ultrametric inequality, which guarantees the bijection between the ultrametric matrix and the underlying rooted tree structure if all diagonal elements are positive (McCullagh, 2006; Bravo et al., 2009). Specifically, Σ^T uniquely identifies a weighted tree T with p leaves if the element of $\sigma_{j,k}$ measures the sum of branch lengths from the root to the most recent common ancestor of leaves j and k (Bravo et al., 2009; Dellacherie et al., 2014). Conversely, a tree T also uniquely determines an ultrametric matrix Σ^T by the same construction above. Equation (3.1) also ensures the positive definiteness

of the ultrametric matrix if the matrix has positive diagonal elements (Dellacherie et al., 2014).

To this point, though the bijection of the ultrametric matrices and the tree structure is established, the information from the geometry of the BHV space is still not fully utilized to characterized the space of ultrametric matrices, resulting in inefficient inference. For example, existing methods decompose (Nabben and Varga, 1994; Bravo et al., 2009) the ultrametric matrix as follows:

$$\Sigma^T = \sum_{j=1}^{2p-1} d_j v_j v_j^{\mathsf{T}}, \tag{3.2}$$

where $\{v_j\}$ are p-dimensional binary vectors with values in $\{0, 1\}$ and at least one non-zero element, and d_j is a positive branch length on each branch. The vector set $V = \{v_1, \ldots, v_{2p-1}\}$ collectively represents a nested partition corresponding to the binary tree topology. Specifically, for every vector $v_j \in V$ with more than one non-zero elements, we can find the other two vectors that partition the vector v_j such that $v_k + v_h = v_j$ and $v_k, v_h \in V$, referred to as the *partition property* of V (Bravo et al., 2009). However, this partition property also imposes a difficult condition when updating a certain vector element in the set V. If one aims to replace a vector in the set, the new vector must satisfy two conditions simultaneously: (i) it should be decomposed as the sum of two existing vector elements in the original set, and (ii) it must identify another vector such that the sum of these two vectors forms another vector already present in the set. These two conditions pose challenges for the inference algorithm to move locally in an efficient manner.

3.2.2 Geometry of the Set of Ultrametric Matrices

Denote by $\mathcal{U}_p$ the set of $p \times p$ ultrametric matrices. Bayesian inference on $\mathcal{U}_p$ requires a geometry that enables local moves for any sampling algorithm that seeks

to explore the parameter space efficiently. McCullagh (2006) notes that the set $\mathcal{U}_p$ is neither convex nor a manifold, but does not prescribe a geometry and proposes an algorithm that approximately projects an arbitrary covariance onto $\mathcal{U}_p$. One of our main contributions is to equip $\mathcal{U}_p$ with a CAT(0) geometry through its links with the BHV space (Billera et al., 2001).

Consider the set of acyclic graphs T known as *trees* with a unique vertex known as the *root*. Nodes with degree one are referred to as *leaves*, including the root, and all other nodes have degree greater than two and are known as *internal nodes*. We consider trees T on p leaves labelled $L = \{0, 1, \ldots, p\}$ with the root labelled as leaf 0. Vertices are connected by edges from the set $\mathcal{E}_T$, which is the union of the set $\mathcal{E}_T^I$ of edges connecting internal vertices with the set $\mathcal{E}_T^L$ of edges connecting internal vertices to the p leaves and the root. *Resolved* trees T are those with internal vertices of degree three and $p-2$ internal edges in $\mathcal{E}_T^I$, while *unresolved* trees are trees T with fewer than $p-2$ internal edges and containing internal vertices of degree four or higher.

The topology of a tree T is characterized in the connectivity between its internal edges in $\mathcal{E}_T^I$, encoded in the set of partitions into two of $L = \{0, 1, \ldots, p\}$ called *splits* pertaining to each edge in $\mathcal{E}_T^I$. Precisely, each edge $e \in \mathcal{E}_T^I$ uniquely determines a split $L = A \cup A^c$ upon its removal from a tree T, where A contains leaves on the descendant subtree of e and its complement $A^c = L - A$ contains the rest of the leaves. Denote by e_A the corresponding edge. The set $A \subset L$ identifies a split $L = A \cup A^c$, and we use split to refer to A or the edge e_A interchangeably; context will disambiguate the two.

Arbitrary collections of splits do not characterize a valid tree topology, but only a collection of *compatible* ones do: two edges e_{A_1} and e_{A_2} are compatible if one of $A_1 \cap A_2$, $A_1 \cap A_2^c$ and $A_1^c \cap A_2$ from the associated splits is empty. Again, we interchangeably refer to compatibility of splits A_1 and A_2 to sometimes mean compatibility of the edges

e_{A_1} and e_{A_2}, and this extends to a collection $\{A_1, \ldots, A_k\}$ of subsets of L.

Leaf edges $e_A \in \mathcal{E}_T^L$ associated with singleton splits $A \subset L$ are compatible with all internal edges in $\mathcal{E}_T^I$, and thus do not contribute to the topology of T. A compatible edge set $\mathcal{E}_T$ thus fully characterizes the topology of a tree T. There are $(2p-3)!!$ distinct topologies on fully resolved trees on p leaves.

The BHV space $\mathcal{T}_p^I$ parameterizes the space of labelled, resolved and unresolved trees T on p leaves and prescribes a continuous geometry based on the lengths $|e_A|$ of internal edges $e_A \in \mathcal{E}_T^I$, where A is associated with a split of L. A fully resolved topology is parameterized by $\mathbb{R}_{>0}^{p-2}$, where each axis corresponds to one of the $p-2$ internal splits that characterize the topology and the coordinates encode the corresponding lengths of the internal edges. The boundary of $\mathbb{R}_{>0}^{p-2}$ consists of unresolved trees with internal nodes of degree greater than 3, obtained by shrinking the internal edges to zero. Each of the $(2p-3)!!$ topologies is identified with a copy of $\mathbb{R}_{\geq 0}^{p-2}$, known as an *orthant*, and the BHV space $\mathcal{T}_p^I$ is defined by the $(2p-3)!!$ orthants glued isometrically along their common boundaries comprising unresolved trees. Panel (A) of Figure III.1 illustrates that two neighbouring orthants share a common edge in $\mathcal{T}_4^I$. By accounting for lengths of p leaf edges and the root edge, space $\mathcal{T}_p$ of rooted, labelled trees on p leave then becomes

$$\mathcal{T}_p = \mathcal{T}_p^I \times \mathbb{R}_{>0}^{p+1},$$

where we do not allow for zero-length leaf edges nor zero-length root edge. The distance $d_{\text{BHV}}(T_1, T_2)$ between two trees T_1 and T_2 on p leaves is defined to be the infimum of lengths of paths between T_1 and T_2 in $\mathcal{T}_p^I$, which are straight lines within each orthant. A natural distance on $\mathcal{T}_p$ then is

$$d_{\text{tree}}(T_1, T_2) := d_{\text{BHV}}(T_1, T_2) + \|x - y\|_2,$$

where $\|x - y\|_2$ is the L_2 norm, and $x, y \in \mathbb{R}^{p+1}_{>0}$ are the vectors of leaf edge lengths, including the root edge, in T_1 and T_2, respectively,

Theorem 3.2.1. *The map $\Phi : \mathcal{U}_p \to \mathcal{T}_p$ is a bijection. Equipped with the distance*

$$d(\Sigma_1^T, \Sigma_2^T) := d_{tree}(\Phi(\Sigma_1^T), \Phi(\Sigma_2^T))$$

the space $\mathcal{U}_p$ is CAT(0).

Proof. The BHV space $\mathcal{T}_p^I$ with distance d_{BHV} is known to be CAT(0) (Billera et al., 2001, Lemma 4.1). The space $(\mathbb{R}^{p+1}_{>0}, \|\cdot\|_2)$ is Euclidean and hence CAT(0), and $(\mathcal{T}_p, d_{\text{tree}})$ as a product of two CAT(0) spaces is thus CAT(0) (Bridson and Haefilger, 1999). The distance d on $\mathcal{U}_p$ is the pullback of d_{tree} from $\mathcal{T}_p$, and the proof follows if it is established that the map Φ is injective. We prove this for the case of fully resolved trees in the interior of each orthant; an identical argument holds for unresolved ones on the boundaries.

Recall the decomposition of an ultrametric matrix $\Sigma^T = VDV^\top$ in (3.2) with a binary matrix $V \in \{0,1\}^{p\times(2p-1)}$ and a diagonal matrix $D \in \mathbb{R}^{(2p-1)\times(2p-1)}_{>0}$ with positive entries corresponding to edge lengths of $(p-2)$ internal edges, p leaf edges, and one root edge. Every column vector v of V maps uniquely to an edge of a tree T the matrix Σ^T uniquely determines (Dellacherie et al., 2014), and thus to an edge e_A associated with a split $L = A \cap A^c$ on the leaves indexed by $L = \{0, 1 \ldots, p\}$ of T. It suffices to establish a bijective relationship between the matrix V that encodes topology of a tree T with the edge set $\mathcal{E}_T$ of compatible splits that identify an orthant in $\mathcal{T}_p$, since the edge lengths in D evidently map to the coordinates of a point within that orthant. From the partition property of V the proof is complete if it is established that any triplet (v_i, v_j, v_k) of columns vectors in V satisfy $v_i = v_j + v_k$ if and only if there exists a triplet $(e_{A_i}, e_{A_j}, e_{A_k})$ of edges/splits in T that are mutually pairwise compatible.

For each edge e_A associated with a split A define the unique p-dimensional binary vector b_A with ones at indices that are in A and zero for indices in A^c. Arrange the vectors into a $p \times (2p-1)$ binary matrix $B = (b_{A_1}, \ldots, b_{A_{2p-1}})^\top$ corresponding to the $(2p-1)$ edges in a fully resolved T.

In order to relate the columns of B to those of V possessing the partition property, we use the logical `and` operator $\wedge$ on columns of B. In other words, the compatibility criterion that one of $A_1 \cap A_2, A_1 \cap A_2^c$ and $A_1^c \cap A_2$ associated with two splits e_{A_1} and e_{A_2} be empty translates to one of $b_{A_1} \wedge b_{A_2}, \bar{b}_{A_1} \wedge b_{A_2}$ and $b_{A_1} \wedge \bar{b}_{A_2}$ equalling the zero vector $\mathbf{0}$, where $\bar{b}$ denotes the negation of b. Accordingly, suppose first that $v_i = v_j + v_k$ for a triplet (v_i, v_j, v_k) of columns in V with a corresponding triplet $(b_{A_i}, b_{A_j}, b_{A_k})$ of columns in B satisfying $b_{A_i} = b_{A_j} \vee b_{A_k}$, where $\vee$ is the logical `or` operator. It is then easily verified that $b_{A_j} \wedge b_{A_k} = \mathbf{0}$ while $\bar{b}_{A_j} \wedge b_{A_k} \neq \mathbf{0}$ and $b_{A_j} \wedge \bar{b}_{A_k} \neq \mathbf{0}$, rendering the splits corresponding to the pair (b_{A_j}, b_{A_k}) compatible. Similarly, $b_{A_i} \wedge b_{A_j} \neq \mathbf{0}$, and either $\bar{b}_{A_i} \wedge b_{A_j} = \mathbf{0}$ or $b_{A_i} \wedge \bar{b}_{A_j} = \mathbf{0}$ but not both, since otherwise $b_{A_i} \neq b_{A_j} \vee b_{A_k}$. The splits corresponding to pair (b_{A_i}, b_{A_j}) are hence compatible; a similar argument applies to the pair (b_{A_i}, b_{A_k}) rendering them compatible. In fact, we observe that the preceding arguments are biconditional, and proof of the reverse assertion follows. $\square$

The bijection Φ engenders a new decomposition of Σ^T that decouples the geometric and topological content of a tree T, and is equivalent to that in (3.2), proof of which follows from that of Theorem 3.2.1.

Corollary 3.2.2. Each edge e_A, for $A \subset L$, in a collection $\mathcal{E}_T$ of compatible edges/splits is associated with a binary matrix E_A, with $E_A(j,k) = 1$ if $j, k \in A$ and 0 otherwise, such that the ultrametric matrix can be decomposed as

$$\Sigma^T = \sum_{e_A \in \mathcal{E}_T} |e_A| E_A \, . \tag{3.3}$$

Every collection $\mathcal{E}_T$ of $(2p-1)$ compatible splits determines a unique set of $\{E_A\}$

of binary matrices that completely characterizes topology of the tree T; there are $(2p-3)!!$ such compatible splits. More precisely, the subset of $(p-2)$ internal splits within each $\mathcal{E}_T$ that determine $(p-2)$ binary matrices within $\{E_A\}$, corresponding to the internal edges, identifies the orthant in the BHV space $\mathcal{T}_p^I$ pertaining to the topology of T. The remaining p binary matrices within $\{E_A\}$ contain a single non-zero entry on the diagonal and represent the $(p+1)$-axes in $\mathbb{R}_{>0}^{p+1}$ that identify splits associated with the leaf edges and the root edge in $\mathcal{T}_p$. The coefficients $|e_A|$ as e_A varies in $\mathcal{E}_T$ represents the edge lengths of T and ascribe coordinates to the $(2p-1)-$dimensional point within $\mathcal{T}_p$. Figure III.1 illustrates the decomposition of an ultrametric matrix Σ^T in $\mathcal{U}_4$ with corresponding tree $T \in \mathcal{T}_4$ and a compatible edge set $\mathcal{E}_T = \{e_{1234}, e_{123}, e_{23}, e_1, e_2, e_3, e_4\}$.

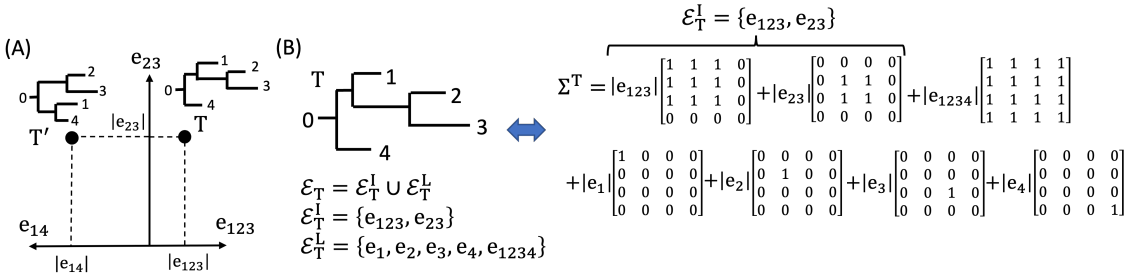


Figure III.1: The decomposition of Σ^T and the corresponding tree T in the tree space. Panel (A) shows the tree space for tree wit 4 leaves. Panel (B) demonstrate the decomposition of the Σ^T by the edge set shown in the tree space.

From the bijection in Theorem 3.2.1, upon discounting the leaf edges, the *decomposition in Corollary 3.2.2 thus provides a novel representation of a tree in the BHV tree space.*

3.3 General Priors for Ultrametric Matrix Parameters

The bijection of the ultrametric matrix and the tree with Theorem 3.2.1 and Corollary 3.2.2 guides prior specification for ultrametric-matrix-valued parameters. Let $p(\Sigma^T)$ be the prior on the ultrametric matrix Σ^T. Corollary 3.2.2 enables the

factorization of the matrix into the topology $\mathcal{E}_T$ and the branch lengths $\mathcal{L}_T$, given by:

$$p(\Sigma^T) = p(\mathcal{E}_T, \mathcal{L}_T) = \pi_{\mathcal{E}}(\mathcal{E}_T)\pi_{\mathcal{L}}(\mathcal{L}_T \mid \mathcal{E}_T), \mathcal{L}_T = \{|e_A| : e_A \in \mathcal{E}_T\}, \tag{3.4}$$

where $\pi_{\mathcal{E}}(\mathcal{E}_T)$ is the prior on the topology and $\pi_{\mathcal{L}}(\mathcal{L}_T \mid \mathcal{E}_T)$ is the prior on the branch lengths conditioning on the topology. Here $p(\Sigma^T)$ is the density function over $\mathbb{R}^{(p+1)p/2}$ with respect to Lebesgue measure; $p(\mathcal{E}_T, \mathcal{L}_T)$ is the density function over the product space of rooted trees and $\mathbb{R}^{2p-1}$ with respect to the product measure of counting and Lebesgue measures on respective spaces for $\mathcal{E}_T$ and $\mathcal{L}_T$.

For the tree topology, we focus on resolved trees with a smaller number of possible tree topologies. Specifically, we consider the binary fragmentation, which describes the topology as a recursive splitting rule of dividing a block into two sub-blocks. The splitting rule is formulated as a distribution:

$$\pi_{\mathcal{E}}(\mathcal{E}_T) = \prod_{e_A, e_B \in \mathcal{E}_T} \pi(e_A, e_B \mid e_{A \cup B}), \tag{3.5}$$

where $\pi_{\mathcal{E}}(e_A, e_B \mid e_{A \cup B})$ is the probability of a block $A \cup B$ splitting into two sub-blocks of A and B.

Currently, multiple models describe different splitting rules with various distributions of (3.5) and properties on the topology. For example, Berestycki et al. (2007) introduces a time-irreversible Markovian fragmentation process that forbids the reversed process as a coagulation process. One popular choice for the splitting rule is of Gibbs type, which assigns the probability of (3.5) as a product of weights depending only on the size of sub-blocks (Pitman, 2006). In this Chapter, we focus on the Gibbs-types splitting rule that results in a consistent Markovian binary fragmentation process. Specifically, the resulting consistent fragmentation engenders a Kolmogorov consist distribution on the tree topology and guarantees the infinite exchangeability with the existence of the fragmentation process. It is well-known that beta-splitting

from Aldous (1996) is the only consistent Markovian binary fragmentation (McCullagh et al., 2008) and assigns the probability as follows:

$$\pi_{\mathcal{E}}(e_A, e_B \mid e_{A\cup B}) \propto \frac{\Gamma(n_A+\beta+1)\Gamma(n_B+\beta+1)}{\Gamma(n_A+n_B+2\beta+2)}, \tag{3.6}$$

where n_A is the cardinality of the set A and $\beta \in (-2, \infty]$ is the hyper-parameter that controls the distribution. For example, $\beta = -1.5$ corresponds to a uniform prior on topology, while $\beta = 0$ corresponds to the Yule model. Further details can be found in McCullagh et al. (2008).

Regarding the prior on the branch lengths $\pi_{\mathcal{L}}(\mathcal{L}_T \mid \mathcal{E}_T)$, we adopt a flexible approach by assign different dependencies between the branch lengths and the topology. For example, diffusion models (e.g. Neal, 2003; Knowles and Ghahramani, 2015) assign a prior through a pre-specified hazard function with the number of leaves on the current branch as a parameter for the hazard function, resulting in branch lengths that always sum to 1 from the root to the leaf nodes. Alternatively, we can assume the independence between the branch lengths and topology and let the branch length follows a distribution with positive support, in which case the lengths of the paths from the root to the leaves may differ and not equal to one.

3.4 Posterior Inference

3.4.1 Metropolis-Hastings Algorithm

The decomposition of the ultrametric matrix in the tree space allows us to leverage the geometry with the coordinate system in the space of rooted trees with p leaves, motivating our proposal of an efficient Metropolis-Hastings algorithm (MH) that moves geodesically on the BHV space. For ease of presentation, we focus on the multivariate normal distribution with the ultrametric matrix as the covariance

matrix and formulate the model as:

$$\boldsymbol{X}_i \overset{i.i.d.}{\sim} N_p(0, \Sigma^T), i = 1, \ldots, n. \tag{3.7}$$

Extensions to other models, e.g., elliptical distributions including the multivariate t-distribution, can be done by changing the likelihood. In addition, in the presence of other parameters, sampling steps in addition to our proposed MH algorithm are needed.

We detail the rationale of an MH iteration in the space of four-leaf rooted trees by proposing a geodesic move to propose a candidate, as highlighted by a green arrow path in Panel (F) of Figure III.2. Specifically, given a tree $T^{(m)}$ with four leaves at l-th iteration, we propose a candidate via a geodesic move from $T^{(m)}$ to $T_1^{\mathsf{cand},(m+1)}$ with all matrices on the path being ultrametric matrices. For example, five matrices on the geodesic path from $T^{(m)}$ to $T_1^{\mathsf{cand},(m+1)}$ are shown in Panel (A) to (E) and the ultrametric inequalities hold for all matrices on the paths. One possible algorithm to achieve this is from Nye (2020) that gives a random walk on the BHV space. However, we observe that Nye's algorithm results in a slower mixing (see Supplementary Material Section 1.2). We make a more efficient move geodesically on the space because our algorithm always proposes a new topology while Nye's algorithm updates the edge set more conservatively with a higher probability of staying in the same orthant.

Based on by Nye's algorithm, we improve the algorithm by decoupling the updates for the edge set and the branch lengths. We force the algorithm to always propose a new topology (though may be rejected). At each iteration, our algorithm updates the edge set by proposing a new edge set that lands on the nearby orthant and adjusts the branch lengths by moving the tree locally within the same orthant. An important result from the CAT(0) geometry of the BHV space is to allow us to find compatible

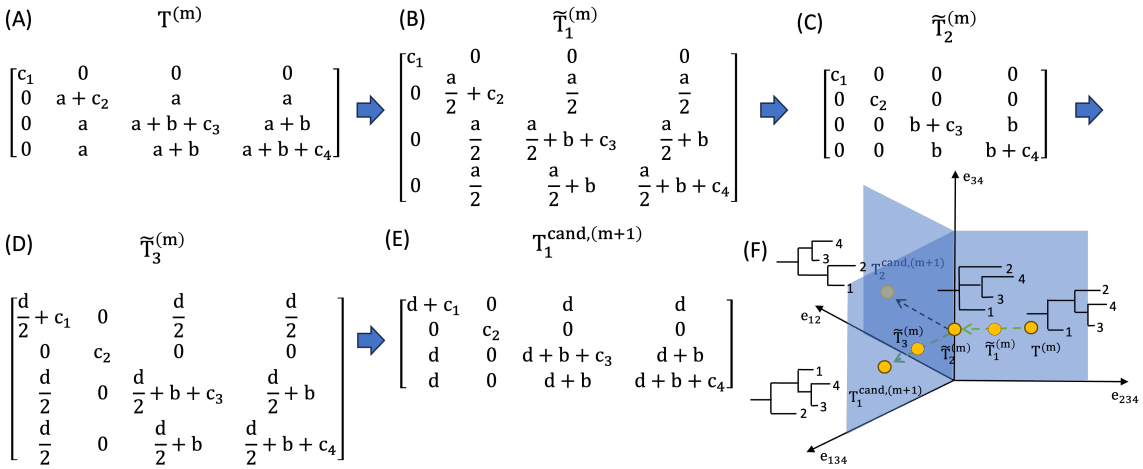


Figure III.2: An illustration of proposing a new edge set for a tree with 4 leaves. Given a tree $T^{(m)}$, the proposal function randomly shrinks a edge and moves to a intermediate tree ($\widetilde{T}^{(m)}$) on the boundary. Two candidate trees ($T_1^{(m+1)}$ and $T_2^{(m+1)}$) that locate in the nearby orthant of tree $T^{(m)}$ can be proposed by our algorithm. The root edge is ignored in matrices in Panel (A) to (E).

splits easily (see Supplementary Material Section B.1). Specifically, if we focus on the resolved trees with $p-2$ internal edges, the BHV space permits only three candidate splits (Nye, 2020) that are compatible with the edge set $\mathcal{E}_T^I \setminus \{e_A\}$, which includes the original split e_A. We ensure that the algorithm proposes a new tree topology by excluding the original split e_A and randomly pick a new split from the remaining two candidates with equal probability. The original length is assigned to the newly proposed split, resulting in a new edge set $\mathcal{E}_{T'}$. We then calculate the acceptance probability with the normal likelihood and the prior $\pi_{\mathcal{E}}$ described in Section 3.3 as follows:

$$\alpha = \max\left\{1, \frac{\pi_{\mathcal{E}}(\mathcal{E}_{T'})N_p(0,\Sigma^{T'})q(\mathcal{E}_T \mid \mathcal{E}_{T'})}{\pi_{\mathcal{E}}(\mathcal{E}_T)N_p(0,\Sigma^{T})q(\mathcal{E}_{T'} \mid \mathcal{E}_T)}\right\}, \tag{3.8}$$

where $N_p(0,\Sigma^T)$ is the normal likelihood with mean zero and the ultrametric matrix Σ^T as the covariance, and $q(\mathcal{E}_{T'} \mid \mathcal{E}_T)$ is the jumping probability from the edge set $\mathcal{E}_T$ to the new edge set $\mathcal{E}_{T'}$. Assuming that we delete the split uniformly and select the new candidate split with equal probability, we obtain two equivalent jumping

probabilities with $q(\mathcal{E}_T \mid \mathcal{E}_{T'}) = q(\mathcal{E}_{T'} \mid \mathcal{E}_T)$. We then update the edge set based on the acceptance rate. After updating the topology, we adjust all branch lengths by a regular MH update with the edge set fixed, representing our algorithm locally moves within the same orthant. Consequently, the characterization with the coordinate system allows us to discover nearby orthants of different topology and locally move our algorithm along geodesics between nearby orthants on the BHV space. Our rationale is to make many computationally cheap local moves over long iterations rather than a few computationally expensive moves, resulting in better exploration of the space of rooted trees. Our MH algorithm is summarized in Algorithm 1.

Returning to the example for a tree of 4 leaves in Figure III.2. Given a tree $T^{(m)}$ at l-th iteration with split set $\{e_{234}, e_{34}\}$, the proposal function randomly shrinks an internal split of $|e_{234}| = 0$ and results in a intermediate multifurcating tree $\widetilde{T}_2^{(m)}$ with the split set containing only one element of $\{e_{34}\}$ on the boundary of three nearby orthants. For the intermediate tree, three splits that correspond to nearby orthants are compatible (e_{234}, e_{12} and e_{134}). After excluding the original split of e_{234}, we choose a new split randomly from the remaining candidates representing the underlying trees of $T_1^{(\mathsf{cand},m+1)}$ and $T_2^{(\mathsf{cand},m+1)}$. If our algorithm choose a new split of e_{12}, we assign the same branch length to the new split as $|e_{12}| = |e_{234}|$. Once a new edge set is proposed, we then calculate the acceptance rate of (3.8) and update the topology based on the acceptance rate. Last, we update the branch lengths given the current edge set.

3.4.2 Posterior Summaries

Once we obtain the posterior samples of edge sets and the branch lengths, we map each edge set with branch lengths onto corresponding ultrametric matrix and tree structure. We then summarize posterior trees and matrices in two ways: (i) point estimation with the representative trees, specifically, the *maximum a posterior* (MAP) tree and the Fréchet mean tree (Miller et al., 2015); and (ii) uncertainty

Algorithm 1 MH algorithm using the BHV space characterization

Input:

(a) The edge set $\mathcal{E}_T = \mathcal{E}_T^I \cup \mathcal{E}_T^L$, where $\mathcal{E}_T^I$ is the internal edge set with each internal split representing a axis in BHV space and $\mathcal{E}_T^L$ is the leaf edge set;

(b) The branch lengths of $\mathcal{L}_T = \{|e| : e \in \mathcal{E}_T\}$;

(c) Priors on edge set $\pi_{\mathcal{E}}$ and branch lengths $\pi_{\mathcal{L}}$;

(d) Number of iterations M and a standard deviation $\sigma_{\mathcal{L}}$ for updating the branch lengths.

Output:

- Posterior samples of edge sets $\mathcal{E}_T$ and branch lengths $\mathcal{L}_T$ of size M.

for $m = 1, \ldots, M$ **do**
  procedure UPDATE THE EDGE SET($\mathcal{E}_T$)
    Randomly remove a split from the internal edge set $e_A \in \mathcal{E}^I_{T^{(m)}}$;
    Three candidate splits ($e_A, e_{A'}$ and $e_{A''}$) are compatible with the remainging edge set $\mathcal{E}^I_{T^{(m)}} \setminus \{e_A\}$;
    Exclude the original split e_A and propose the new split e_B from the rest two candidates ($e_B \in \{e_{A'}, e_{A''}\}$);
    Assign branch lengths $|e_B| = |e_A|$;
    Calculate the acceptance rate α from (3.8) and generate $u \sim \text{Unif}(0, 1)$;
    if $u \leq \alpha$ **then**
      Return the edge set $\mathcal{E}_{T^{(m+1)}} = \{e_B\} \cup \mathcal{E}_{T^{(m)}} \setminus \{e_A\}$;
    else
      Return the edge set $\mathcal{E}_{T^{(m+1)}} = \mathcal{E}_{T^{(m)}}$;
  procedure UPDATE THE BRANCH LENGTHS($\mathcal{L}_T$)
    for $e \in \mathcal{E}_{T^{(m+1)}}$ **do**
      Generate the new branch length with truncated normal distribution $\text{TruncN}_{(0,\infty)}(e_A, \sigma_{\mathcal{L}})$;
      Calculate the acceptance rate.

quantification via the frequency of true subtrees visited by the posterior samples and the 95% credible intervals for each element in the ultrametric matrix.

3.5 Simulation Studies

We empirically demonstrate the utility of the proposed method through a series of simulation studies and show that the proposed method can restore the underlying ultrametric matrix under different true data generating mechanisms. Without loss

of generality, we present bijection and simulation results without root edge. Given the true ultrametric matrix Σ^{T^0} with $p = 10$ leaves, we consider three data generating mechanisms of (i) correct specified normal distribution $X_i \overset{i.i.d.}{\sim} N(0, \Sigma^{T^0})$ and (ii) mis-specified t distribution $X_i \overset{i.i.d.}{\sim} t_\nu(0, \Sigma^{T^0})$ with degrees of freedom four and three ($\nu = 3$ and 4). We generate the data with five different sample sizes of $n \in \{3p, 5p, 10p, 25p, 50p\}$ and 50 independent independent replicates.

We summarize the posterior samples by using the statistics in Section 3.4.2. We calculate the MAP tree and the mean tree (Miller et al., 2015) as representative trees and measure the matrix norm and the BHV distance (Owen and Provan, 2011) between the true underlying tree and the representative tree. For each split in the true tree, we measure the split-wise recovery by computing the frequency of the posterior samples that contains the true splits. Lastly, we also investigate the coverage for each element in the matrix for the element-wise 95% credible interval. For point estimation, we compare the representative tree from our method to Bravo et al. (2009), which formulates the matrix estimation as a mixed-integer programming (MIP) problem. Under the matrix norm, we also consider the sample covariance. For the uncertainty quantification, no existing method can directly quantify the uncertainty to our best knowledge. We assign priors of $\beta = -1.5$ as the uniform prior on all topology and exp(1) on the branch lengths. We run the MCMC for $5,000$ iterations and discard the first $4,000$ iterations.

We show the distance from the representative tree to the true tree in Figure III.3. Obviously, the mean and MAP trees from our method are comparable to the estimated matrix from Bravo et al. (2009) and sample covariance in terms of BHV distance and matrix norm across different data generating mechanisms and sample sizes. When the model is correctly specified, all methods benefit from the increase of the sample size with a smaller distance to the true tree. For the mis-specified scenario, the advantage from the larger sample size is moderate.

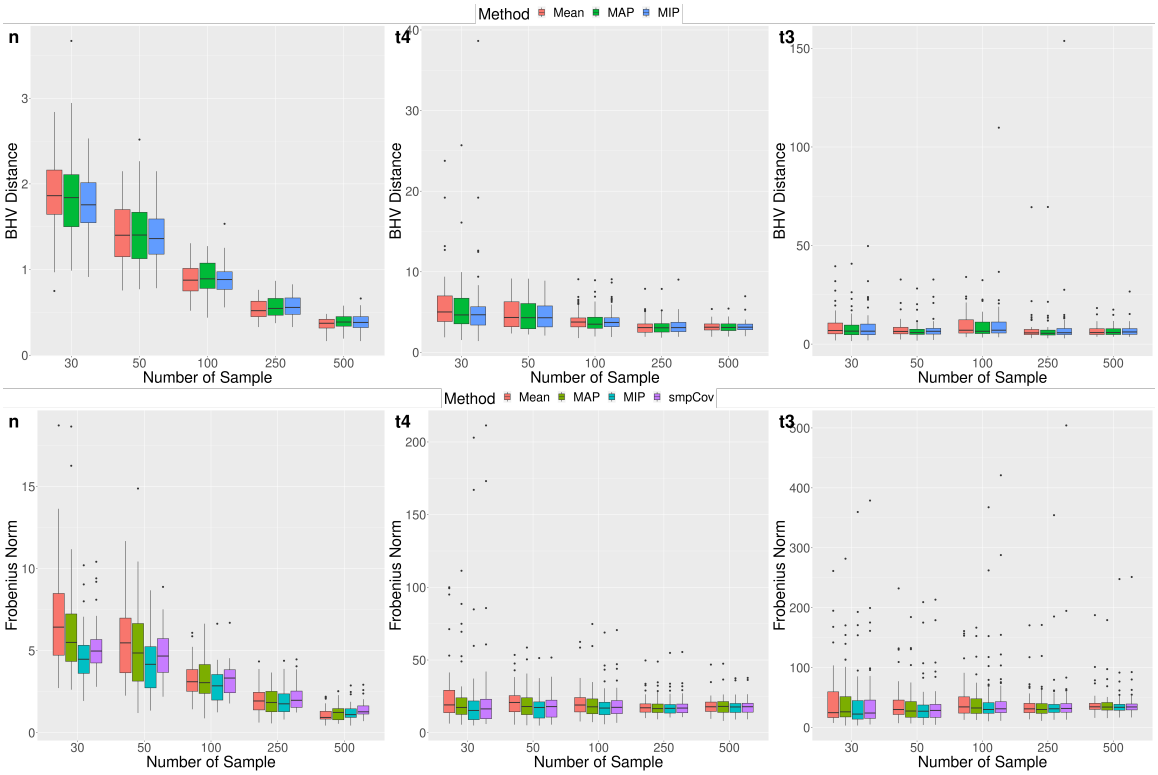


Figure III.3: Distances between the estimated matrix and the true matrix under different data generating mechanism and sample sizes. The mean (red) and MAP tree (green) from our method is comparable to competing methods (blue for MIP and purple for sample covariance) in terms of the BHV distance (top row) and matrix norm (bottom row).

We quantify the uncertainty for ultrametric matrices through the split-wise recovery in Table 3.1 and element-wise coverage in Figure III.4. In Table 3.1, we calculate the proportion of the posterior splits that contain each split in the true tree. The split-wise recovery performs better when the sample size increases for all data generating mechanisms. For different data generating mechanisms, the correct specified model performs the best with sample size of $n = 50$ to ensure around 90% of the posterior samples having correct splits, while for the similar level of recovery, the mis-specified t-distribution requires sample size over 100 and 250 for t_4 and t_3, respectively. Among all splits, we also observe that the split with a smaller length ($|e_{5,6}| = 0.231$) has the worst recovery. We present the results of element-wise coverage of the 95% credible interval for the normal distribution in Figure III.4. The results for t-distribution are available in Supplementary Material Section B.2. Elements in the last row and col-

umn correspond to zero elements in the true covariance and result in an estimated coverage of one. For non-zero elements in true covariance, the estimated coverage are high but slightly lower than the nominal coverage (around 0.75 to 0.94). The estimated coverage is higher when the sample size increase.

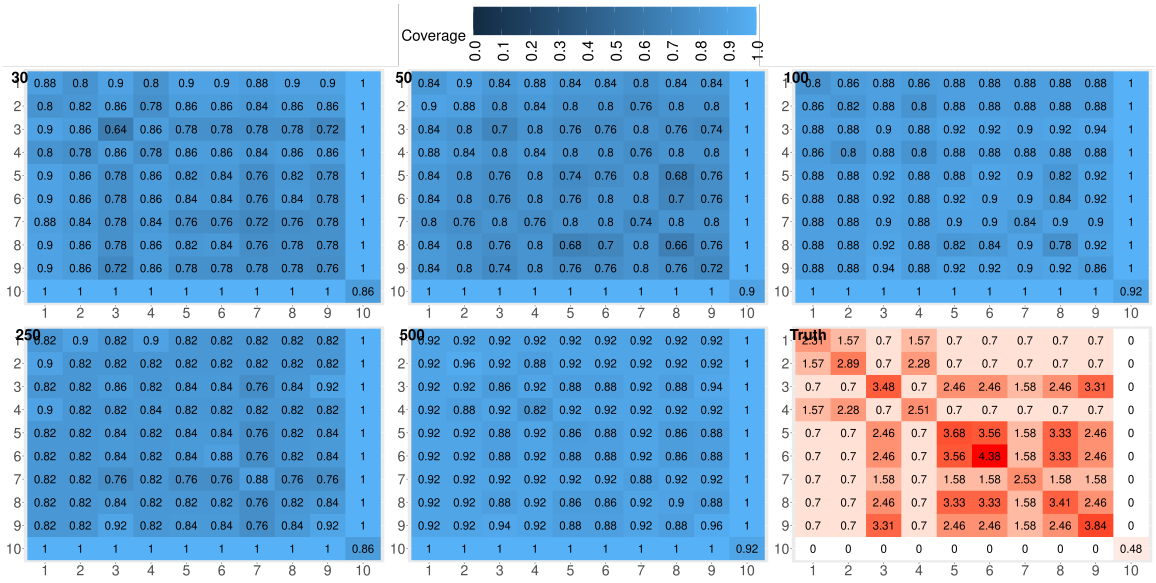


Figure III.4: Element-wise coverage from the 95% credit interval for the correct specified normal distribution with fiver different sample sizes with the true underlying covariance in the lower right panel.

We also provide additional results for simulation including (i) convergence diagnostics, (ii) element-wise coverage for mis-specified t-distribution, (iii) sensitivity analysis for the hyper-parameter of the prior on the branch lengths, (iv) the topology trajectory for the proposed method, and (v) the simulation results for the data generated from a underlying tree with unit-lengths.

3.6 Analysis of Treatment Tree in Cancer

We exemplar the proposed method on a pre-clinical dataset, such as patient-derived xenograft (PDX) data, to discover potential cancer treatments. Due to the impracticality of testing multiple treatments on the same patient, PDX is a experiment design that evaluates multiple treatments administered to samples from the

same human tumor implanted into genetically identical mice. The mice are treated as the "avatars" to mimic responses to different treatments. In this analysis, we leverage a PDX dataset of Novartis Institutes for BioMedical Research - PDX Encyclopedia [NIBR-PDXE, (Gao et al., 2015)] that collects over $1,000$ PDX lines across multiple cancers with a $1 \times 1 \times 1$ design (one animal per PDX model per treatment).

For our analysis, we focus on cutaneous melanoma, which consists of 14 treatments and 32 PDX lines. The main response is the tumor size difference before and after treatment administration, following the approach by Rashid et al. (2020), with the untreated group as the reference group. Positive responses indicate that the treatment shrunk the tumor more than the untreated group with a higher value representing a better efficacy. We assume that treatments with similar mechanism should induce similar levels of responses, and we aim to construct a tree structure to reveal the mechanism similarity based on the main responses. We ran our method with $10,000$ iterations and discard the first $9,000$ iterations. We summarize the results with the MAP and mean trees and highlight subtrees with the frequency over 90%.

Figure III.5 shows the MAP (Panel (A)) and mean trees (Panel (B)) with subtrees that consistently appear in the posterior samples. Three subtrees with frequencies higher than 90% are emphasized by boxes: blue (91%), and yellow (98%). We observe that the MAP and mean trees share many subtrees with the same topology. For example, the subtrees in the boxes are identical in both the mean and MAP trees. Additionally, two combination treatments highlighted by blue box form a tight subtree that appears over 90% of posterior samples, indicating a high level of mechanism similarity of two combination therapies. Two combination therapies consist of two agents, with one agent being encorafenib and the other targeting one of the following pathways: phosphoinositide 3-kinases (BKM120), and cyclin-dependent kinases (LEE011). As these pathways are closely related and share common downstream mechanisms (e.g., Repetto et al., 2018; Kurtzeborn et al., 2019), it is not surprising

Figure III.5: The MAP (Panel (A)) and mean trees (Panel (B)) for the melanoma. Two boxes emphasize the subtrees with high frequencies ($> 90\%$) in the posterior samples: blue: 91%, and yellow: 98%.

to see that all combination therapies form a tight subtree in the tree structure.

3.7 Discussion

In this Chapter, we develop a novel Bayesian framework that conducts the inference on ultrametric matrices by leveraging the bijection of the ultrametric matrix and the tree structure. Based on the decomposition of (3.3), we characterize the space of the ultrametric matrices via the geometry and the coordinate system in BHV space. The same decomposition also enables a general prior for ultrametric matrices that include many existing priors on the tree structure as special cases. By utilizing the geometry of the BHV, we propose an efficient algorithm that moves locally on BHV space along geodesics between nearby orthants. Moreover, the ultrametric inequalities still hold for all matrices on the geodesic path, and allows us to summarize the posterior samples through existing tree modeling tools. Specifically, we use the MAP and Fréchet mean trees as point estimator and measure the performance of point estimator by BHV geodesic distances and matrix norm. We further quantify the uncertainty via the split-wise recovery and element-wise 95% credible interval. In simulation studies, our proposed method generates point estimators that are comparable with existing projection-based method in terms of the BHV distance and matrix norm. Our proposed algorithm also draws posterior samples that result in high split

recovery and element-wise nominal coverage. We exemplify our method in a preclinical dataset and discover that treatments sharing high mechanism similarities align with existing literature.

Currently, priors of ultrametric matrices are built on the decomposition of Corollary 3.2.2 via the bijection Φ of the Theorem 3.2.1. One may consider to describe the prior directly on the set of ultrametric matrices without the decomposition and bijection. However, the direct construction of the prior on ultrametric matrices is difficult due to the non-trivial geometry for the set of ultrametric matrices. Aside from the characterization in Theorem 3.2.1, Brandts and Cihangir (2016) recently showed that the set of ultrametric matrices is a simplex with only nonobtuse triangular facets. Another two extensions may further improve the utility of the proposed prior and sampling algorithm. In the current implementation, we focus on the binary fragmentation due to the prevalence of the binary trees. It is possible to relax the constraints of only allowing for two sub-blocks in the fragmentation process by considering the multifurcating fragmentation process and assigning a consistent Markovian prior on it such as the two-parameter Poisson-Dirichlet model (McCullagh et al., 2008). However, considering the multifurcating fragmentation will increase the computational burden due to the additional hyper-parameters and a larger number of possible tree topologies. Another possible extension is to include the covariates in the prior, resulting in subject-specific matrices. Specifically, we can model the hyper-parameter β as a linear function of covariates and assign priors on the coefficients. By doing so, we allow the prior to assign weights on different topologies based on the covariates and enables different subjects to borrow information from the prior on the fragmentation process. However, modeling the β as a linear function of covariates requires more details in both theoretical foundation and computation techniques. We leave these topics for future work.

Table 3.1: Split-wise recovery for proposed MCMC with exp(1) prior on the branch lengths under different sample size, data generating distribution. Each column shows proportion of posterior splits that contains specific split in the true underlying tree. The average and the standard deviation of the proportion are obtained from $5,000$ iterations over 50 independent replicates.

Sample Size	Generating Distribution	1,2,3,4,5, 6,7,8,9	3,5,6,7,8,9	3,5,6,8,9	3,9	5,6,8	5,6	1,2,4	2,4
True Branch Lengths		0.701	0.88	0.878	0.854	0.869	0.231	0.872	0.712
30	Normal	0.85(0.21)	0.80(0.23)	0.84(0.21)	0.81(0.23)	0.77(0.28)	0.43(0.28)	0.74(0.27)	0.84(0.17)
	t_4	0.71(0.29)	0.59(0.37)	0.69(0.38)	0.70(0.31)	0.72(0.34)	0.37(0.26)	0.77(0.30)	0.73(0.29)
	t_3	0.67(0.36)	0.69(0.37)	0.70(0.36)	0.71(0.34)	0.69(0.36)	0.43(0.35)	0.66(0.36)	0.73(0.30)
50	Normal	0.91(0.14)	0.91(0.17)	0.91(0.17)	0.91(0.21)	0.94(0.12)	0.65(0.26)	0.91(0.16)	0.94(0.13)
	t_4	0.80(0.30)	0.81(0.31)	0.77(0.33)	0.87(0.22)	0.88(0.24)	0.67(0.31)	0.85(0.25)	0.83(0.25)
	t_3	0.84(0.27)	0.70(0.39)	0.74(0.39)	0.79(0.32)	0.77(0.33)	0.48(0.35)	0.78(0.33)	0.89(0.20)
100	Normal	0.98(0.07)	0.97(0.08)	0.99(0.03)	0.99(0.06)	0.99(0.04)	0.86(0.18)	1.00(0)	0.99(0.04)
	t_4	0.90(0.23)	0.90(0.23)	0.90(0.27)	0.97(0.09)	0.93(0.24)	0.75(0.35)	0.91(0.21)	0.96(0.15)
	t_3	0.81(0.31)	0.75(0.41)	0.78(0.37)	0.88(0.30)	0.82(0.36)	0.62(0.40)	0.81(0.38)	0.83(0.33)
250	Normal	1.00(0)	1.00(0)	1.00(0)	1.00(0)	1.00(0)	0.99(0.04)	1.00(0)	1.00(0)
	t_4	0.96(0.17)	0.99(0.07)	0.99(0.01)	0.98(0.14)	1.00(0)	0.90(0.21)	1.00(0)	0.99(0.03)
	t_3	0.82(0.37)	0.95(0.20)	0.92(0.26)	0.91(0.27)	0.95(0.20)	0.86(0.31)	0.94(0.23)	0.90(0.27)
500	Normal	1.00(0)	1.00(0)	1.00(0)	1.00(0)	1.00(0)	1.00(0)	1.00(0)	1.00(0)
	t_4	1.00(0)	1.00(0)	1.00(0)	1.00(0)	1.00(0)	0.96(0.15)	0.98(0.14)	1.00(0)
	t_3	0.96(0.20)	0.92(0.27)	0.95(0.20)	0.98(0.14)	0.96(0.20)	0.86(0.34)	0.95(0.17)	0.96(0.18)

CHAPTER IV

Robust Bayesian Graphical Regression Models for Assessing Tumor Heterogeneity in Proteomic Networks

4.1 Introduction

Graphical models are ubiquitous and powerful tools to investigate complex dependency structures in high-throughput biomedical datasets such as genomics and proteomics (Airoldi, 2007). They allow for holistic exploration of biologically-relevant patterns that can be used for deciphering cellular processes and formulate new testable hypotheses. However, most existing graphical models make one of two canonical assumptions: (i) a homogeneous graph with a common network for all subjects; or (ii) rely on the normality assumption especially in the context of Gaussian graphical models (Ni et al., 2022a). However, in some biomedical applications both assumptions are violated such as proteomic networks in cancer, as we illustrate next.

Proteomic networks and tumor heterogeneity. Proteins control many fundamental cellular processes through a complex but organized system of interactions, termed protein-protein interactions (PPIs) (Cheng et al., 2020). Moreover, aberrant PPIs are associated with various diseases including cancer and investigating PPI can lead to effective strategies and treatments, including immunotherapies, tailored to different

individuals (Cheng et al., 2020; Lu et al., 2020). Consequently, it is highly desirable to elucidate PPIs in cancer and construct flexible graphical models that can identify multiple types and ranges of dependencies. Modern data collection methods have allowed systematic assessment of multiple proteins simultaneously on the same tumor samples, often referred to as high-throughput proteomics (Baladandayuthapani et al., 2014). However, the resulting data are typically not normally distributed even after extensive preprocessing and data transformations (e.g. logarithmic). As an illustration, Figure IV.1 shows the level of non-normality in protein expression data for two cancers: lung adenocarcinoma (LUAD) and ovarian cancer (OV) samples from The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013) that are used case-studies in this Chapter. Panels (A) and (B) display the empirical density and the quantile-quantile (q-q) plots of four exemplar proteins: Akt and PTEN for LUAD, and E-Cadherin and Rb for OV. Both the empirical distributions and q-q plots demonstrate deviations from normal distribution with heavier tails as shown in Panels (A) and (B). The level of non-normality is quantified using the H-score, defined as $H(\boldsymbol{Y}) = 2\Phi(\log(1 - \text{pval}(\boldsymbol{Y})))$, where Φ is the cumulative distribution function of the standard normal distribution, and $\text{pval}(\boldsymbol{Y})$ is the p-value of the Kolmogorov-Smirnov test for the normality of $\boldsymbol{Y}$ (Chakraborty et al., 2021). The H-score is bounded between zero and one, and a higher H-score implies increased departure from normality. The H-scores for all four proteins are > 0.999, consistent with the conclusions from the empirical and q-q plots. Panel (C) shows the H-score across all the proteins in our datasets, indicating a high degree of non-normality across both cancers.

Another axes of complexity that arises in cancer research is *tumor heterogeneity*. It is now well-established that tumors are heterogeneous with distinct proteomic aberrations even for the same type of cancer across different patients (Janku, 2014). Accumulating evidence suggests that considering tumor heterogeneity, in general, and specifically at the level of PPI can enhance our understanding of tumorigenesis and

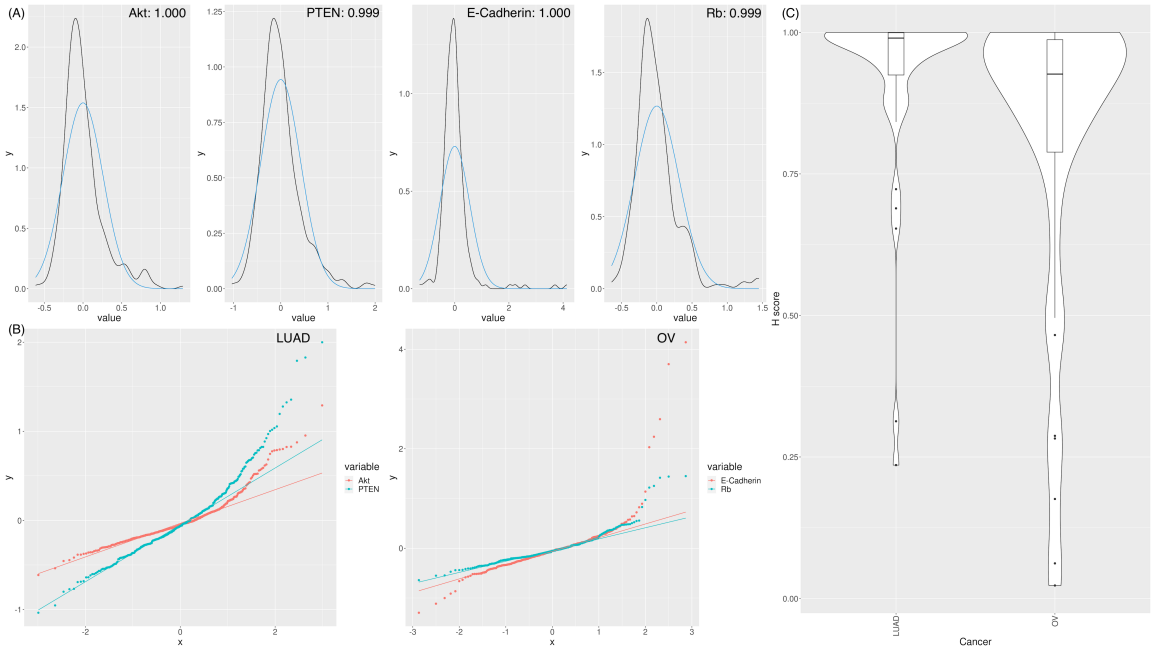


Figure IV.1: Non-normality levels of protein expression in lung adenocarcinoma (LUAD) and ovarian cancer (OV) from TCGA. The empirical density plots from real data (black) and the normal distribution (blue) for the expression of four proteins with the H-score are shown in Panel (A). Panel (B) illustrates the expression of four proteins in LUAD (Akt and PTEN) and OV (E-Cadherin and Rb) with the qq-plots. Panel (C) demonstrates the H-score of LUAD and OV. The H-score is bounded between zero and one, and a higher H-score implies a higher level of non-normality.

the development of anti-cancer treatments (Cheng et al., 2020). Specifically, tumor heterogeneity differentially impacts the PPIs across different patients and results in varied treatment responses (Cheng et al., 2020). Hence, incorporating patient-specific information i.e. accounting for tumor heterogeneity could provide valuable clues to identify PPIs disrupted during carcinogenesis.

In summary, constructing PPI networks poses two main statistical challenges simultaneously: (i) coherently accounting for non-normality in proteomic networks, and (ii) incorporating heterogeneous patient-specific information in graphical modeling.

Existing methods and modeling background. Most existing methods address the aforementioned challenges separately; i.e. either accommodating non-normality without accounting for the sample-specific information (e.g. Pitt et al., 2006; Dobra and Lenkoski, 2011) or requiring normality when incorporating patient-specific informa-

tion (Ni et al., 2022a). To accommodate the non-normality, existing approaches transform the original variables into normal variables either via deterministic functions (e.g. Dobra and Lenkoski, 2011; Liu et al., 2012; Chung et al., 2022) or via random transformations (e.g. Finegold and Drton, 2011, 2014). For instance, Bhadra et al. (2018) generalized the t-distribution to Gaussian scale mixtures and introduced a new graph characterization for undirected graphs. Chakraborty et al. (2021) further generalize concept to characterize chain graphs with both directed and undirected edges. However, all existing models mentioned above assume a common graph across all patients and fail to incorporate the subject-specific information.

More recently, several studies incorporate the subject-specific information under explicit Gaussian assumptions. Multiple Gaussian graphical models were first proposed to estimate graphs that vary across heterogeneous sub-populations (e.g. Peng et al., 2009; Danaher et al., 2014; Peterson et al., 2015). Ni et al. (2019) introduced a more general framework called "Graphical Regression" that construct covariate-dependent graphs through regression model and incorporates both continuous and discrete covariates, in directed as well as undirected settings (Ni et al., 2022b). Similarly, Zhang and Li (2022) provided a penalized procedure to estimate undirected graph by Gaussian graphical regression and introduced continuous covariates in both the mean and the covariance structure. However, all these models are developed under the normality assumption for inferential and computational reasons. To our best knowledge, no existing method incorporates subject-specific information under non-Gaussian settings and motivates development of new methodology. We summarize six important and relevant models mentioned above in Table 4.1 and compare these models in four different aspects.

To address these challenges simultaneously, we develop a unified and flexible modeling strategy called robust Bayesian graphical regression (rBGR), which allows construction of subject-specific graphical models for non-normally distributed continuous data.

Table 4.1: Comparison of existing and proposed methods for different properties.

Method	Uncertainty Quantification	Undirected	Sample-Specific	Non-Normality
GGMx (Ni et al., 2022b)	✓	✓	✓	✗
RegGMM (Zhang and Li, 2022)	✗	✓	✓	✗
GSM (Bhadra et al., 2018)	✓	✓	✗	✓
BGR (Ni et al., 2019)	✓	✗	✓	✗
RCGM (Chakraborty et al., 2021)	✓	✓	✗	✓
rBGR (the proposed)	✓	✓	✓	✓

rBGR makes three main contributions:

(a) *Robust framework to build subject-specific graphs for non-normal data.* rBGR robustifies the normal assumption via random transformation and incorporates covariates employing graphical regression strategies. By accommodating the non-normality via random transformation, we obtain a Gaussian scale mixture, which presumes an underlying latent Gaussian variable and allows explicit incorporation of covariates in the precision matrix (Section 4.2.2) and admits efficient posterior sampling procedures (Section 3).

(b) *New characterization of dependency structures for non-normal graphical models.* The introduction of the random marginal transformations engenders a new type of edge characterization of the conditional dependence for non-normal data, called conditional sign independence with covariates (CSIx, Section 4.2.3 Proposition 4.3.1). CSIx is a generalization of the conditional sign independence (CSI) introduced by Bhadra et al. (2018) which explicitly characterizes the sign dependence between two nodes/variables. We demonstrate via multiple simulations that rBGR can accurately recover dependency structures under different levels of non-normality and against competing graphical regression approaches that assume normality (Section 4.5).

(c) *Deciphering impact of immunogenic heterogeneity in proteomic networks.* We use rBGR to assess proteomic networks across two cancers, lung and ovarian, to systematically investigate the effects of the inherent immunogenic heterogeneity

within tumors. Specifically, we quantify immune cell abundance across tumors and build PPI networks that varies across different immune cell abundance. Our analyses reveal several important hub proteins and PPIs that are deferentially impacted by the immune cell abundance; some corroborate existing biological knowledge but also discover novel associations for future investigations (Section 4.6).

The rest of the Chapter is organized as follows: we introduce rBGR models and characterization in Section 4.2. Section 4.3 focuses on priors and estimation and Section 4.4 delineates the posterior inference via Gibbs sampling. In Section 4.5, we conduct a series of simulations to evaluate the operating characteristics of rBGR and against competing approaches. Section 4.6 provides a detailed analysis of the TCGA dataset, results, biological interpretations and implications. The Chapter concludes by discussing implications of the findings, limitations, and future directions in Section 4.7. A general purpose `R` package and datasets used for constructing PPI networks is also provided on `https://github.com/bayesrx/rBGR`.

4.2 Robust Bayesian Graphical Regression (rBGR)

We start with the Gaussian graphical regression (Section 4.2.1) as a special case of rBGR under the normality assumption and generalize it to the robust case through random transformations (Section 4.2.2). Subsequently, the introduction of the random transformation changes the interpretation of the graph and motivates a new edge characterization (Section 4.2.3).

4.2.1 Gaussian Graphical Regression

Consider p-dimensional random vectors $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ip})^\mathsf{T} \in \mathbb{R}^p$ as (continuous) responses with q-dimensional random vectors of $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iq})^\mathsf{T} \in \mathbb{R}^q$ as covariates for subject $i = 1, \ldots, n$. A subject-specific PPI network from proteomics data

$\boldsymbol{Y}_i$ is constructed to vary based on the immune cell abundance $\boldsymbol{X}_i$ (Section 4.6).

Let $G_i = (V, E_i)$ be an undirected graph over p nodes, where $V = \{1, \ldots, p\}$ is the set of nodes representing $\boldsymbol{Y}_i$ and $E_i \subset V \times V$ is the set of undirected edges in the network for subject i. An undirected edge exists between nodes j and k if $\{j, k\} \in E_i$. Under Gaussian assumption, given the covariates $\boldsymbol{X}_i$, suppose $\boldsymbol{Y}_i$ follows a multivariate normal distribution,

$$\boldsymbol{Y}_i \mid \boldsymbol{X}_i \sim \boldsymbol{N}_p(\boldsymbol{0}, \widetilde{\boldsymbol{\Omega}}^{-1}(\boldsymbol{X}_i)), \text{ for } i = 1, \ldots, n, \tag{4.1}$$

where $\widetilde{\boldsymbol{\Omega}}(\boldsymbol{X}_i) = \{\widetilde{\omega}^{j,k}(\boldsymbol{X}_i)\}_{p \times p}, j, k \in V$ is a functional precision matrix (of covariates) with each element $\widetilde{\omega}^{j,k}(\boldsymbol{X}_i)$ as a function that depends on $\boldsymbol{X}_i$. The functional precision matrix characterizes the graph G_i through zero precision elements. Specifically, zero precision represents a missing edge in the graph e.g. for the case of scalar precision, $\widetilde{\omega}^{j,k}(\boldsymbol{X}_i) = \widetilde{\omega}^{j,k}$, zero precision implies conditional independence (CI) and an missing edge in the graph of CI under Gaussianity (Lauritzen, 1996). For the functional precision matrix, Ni et al. (2022b) introduced covariate-dependent graphs in G and generalized the concept of CI to CI with covariates (CIx, henceforth). In essence, given a covariate $\boldsymbol{X}_i$, the zero precision of $\widetilde{\omega}^{j,k}(\boldsymbol{X}_i) = 0$ implies an missing edge of CIx between nodes j and k. Contrarily, when the functional precision is non-zero $\widetilde{\omega}^{j,k}(\boldsymbol{X}_i) \neq 0$, Y_j and Y_k are conditional dependent with covariates (CDx, henceforth) and an edge exists between nodes j and k given the covariate $\boldsymbol{X}_i$. By modeling the functional precision matrix, CIx defines covariate-specific graphs that vary based on different covariates.

4.2.2 Robust Graphical Regression via Random Transformation

In practice, normal assumption does not always hold (as shown in Figure IV.1). Violation of the normal assumption results in the failure of modeling graphs through

normal precision matrices and motivates new modeling strategies (Finegold and Drton, 2011; Bhadra et al., 2018). In this Chapter, we adapt the random transformation approach (Bhadra et al., 2018) that allows for various non-normal distributions with different tail behaviors. We focus on continuous distributions with heavy tails as observed in our motivating data. To this end, let $0 < d_j < \infty$ for $j = 1, \ldots, p$ be independent positive random scales and have distribution as $d_j \sim p_j$ with $\int dp(d_j) < \infty$ almost surely. Let $\mathbf{D}_i = \text{diag}(1/d_{i1}, \ldots, 1/d_{ip})$ be a diagonal matrix for subject i. Given random scales $d_{ij}, j = 1, \ldots, p$ and the covariates $\boldsymbol{X}_i$, we assume the distribution of $\mathbf{D}_i \boldsymbol{Y}_i$ conditional on $\mathbf{D}_i$ and $\boldsymbol{X}_i$ follows a multivariate distribution,

$$\mathbf{D}_i \boldsymbol{Y}_i = \left[\frac{Y_{i1}}{d_{i1}}, \ldots, \frac{Y_{ip}}{d_{ip}}\right]^{\mathsf{T}} \sim \boldsymbol{N}_p(\mathbf{0}, \boldsymbol{\Omega}^{-1}(\boldsymbol{X}_i)), \text{ for } i = 1, \ldots, n, \tag{4.2}$$

where $\boldsymbol{\Omega}(\boldsymbol{X}_i) = \{\omega^{j,k}(\boldsymbol{X}_i)\}_{p \times p}, j, k \in V$ is the functional precision matrix that characterizes the graph with the covariates $\boldsymbol{X}_i$.

The model in (4.2) generalizes several existing approaches: (i) Equation (4.1) is a special case of Equation (4.2) with d_{ij} as a degenerated distribution of a point mass at one; (ii) when $d_1 = \ldots = d_p = \tau$ with τ^2 following an inverse gamma distribution, Equation (4.2) reduced to a multivariate t-distribution on $\boldsymbol{Y}$ as used by Finegold and Drton (2014), and (iii) for general d_{ij}, (4.2) establishes a rich family of Gaussian scale mixtures for the marginal distribution of Y_j with the density $p(Y_j) = \int (2\pi d_j)^{-1/2} \exp\{-y_j^2/(2d_j)\} \mathrm{d}p(d_j)$.

The introduction of random scales in Equation (4.2) allows us to construct various marginal distribution of Y_j with different tail behaviors. Specifically, by matching tail behaviors of random scales and the target distribution, random scales allow us to construct different marginal distributions. For example, letting Y_j decay polynomially, the Y_j/d_j follows a normal distribution if the random scale d_j also has a polynomial tail (Bhadra et al., 2018). Similar idea can be used for target distribution with

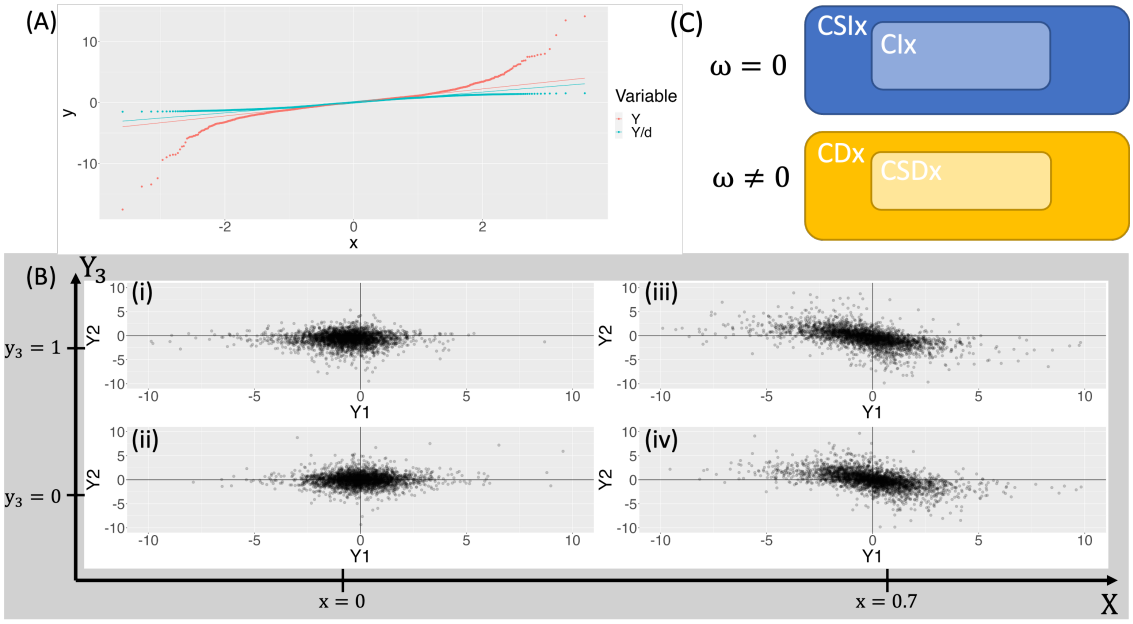


Figure IV.2: The robustification of non-normal distribution with random scales and the visualization of CSIx and CSDx. Panel (A) is the qq-plot to illustrate that random scale d accommodates the non-normal distribution Y with Y/d following the normal distribution. Panel (B) demonstrates CSIx (Case (i) and (ii)) and CSDx (Case (iii) and (iv)) of Y_1 and Y_2 with the partial correlation $\omega^{1,2}(X_i) = X_i$ conditioning on Y_3. Cases (i) and (ii) represent two examples of CSIx with zero precision of $X_i = 0$ given $Y_3 = 1$ and 0. Cases (iii) and (iv) demonstrate the cases of CSDx with non-zero precision of $X_i = 0.7$ given $Y_3 = 1$ and 0. Panel (B) is centered on the values between $[-10, 10]$. Panel (C) shows the nested relationship between CSIx and CIx (top) and CSDx and CDx (bottom). See more details in Section 4.2.3.

exponential tail. In Figure IV.2, Panel (A) shows that the target distribution Y with a polynomial tail deviates from the normal distribution and with the introduction of random scales the distribution of Y/d is normally distributed. While the random scales robustify the model to accommodate non-normality, the resulting functional precision matrix $\boldsymbol{\Omega}(\boldsymbol{X}_i)$ requires careful characterization and interpretation.

4.2.3 Characterization of Functional Precision Matrix

The functional precision matrix in (4.2) determines the graphical dependence as a function of covariates, but the random (marginal) scales changes the standard conditional independence interpretations in the resulting precision matrix which requires a

new characterization. Bhadra et al. (2018) introduced the concept of conditional sign independence (CSI) in non-normal graphs that is defined as follows. Consider random variables Y_1, Y_2, and $\boldsymbol{Y}_3$. Given $\boldsymbol{Y}_3$, Y_1 and Y_2 are conditional sign independence (CSI) if $\mathcal{P}(Y_2 > 0 \mid Y_1, \boldsymbol{Y}_3) = \mathcal{P}(Y_2 > 0 \mid \boldsymbol{Y}_3)$ and $\mathcal{P}(Y_1 > 0 \mid Y_2, \boldsymbol{Y}_3) = \mathcal{P}(Y_1 > 0 \mid \boldsymbol{Y}_3)$. Otherwise, Y_1 and Y_2 are conditional sign dependent (CSD) given $\boldsymbol{Y}_3$. The CSI of Y_1 and Y_2 implies that the information of Y_1 does not affect the sign of Y_2 given $\boldsymbol{Y}_3$. That is, conditioning on $\boldsymbol{Y}_3$, the distribution of the sign of Y_2 is independent of the value of Y_1. Under the multivariate distribution of (4.2) with a constant precision matrix $\boldsymbol{\Omega}(\boldsymbol{X}_i) = \boldsymbol{\Omega}$, zero precision of $\omega^{j,k} = 0$ and the CSI of Y_j and Y_k given the rest are equivalent, which can be represented by a missing edge between nodes j and k in an undirected graph (Bhadra et al., 2018; Chakraborty et al., 2021).

In this Chapter, we generalize the concept of CSI to incorporate covariates and consider subject-specific CSI of two random variables given the rest random variables and a realization of covariates $\boldsymbol{X}_i$; as formalized in the following proposition:

Proposition 4.2.1 (Conditional Sign Independence with Covariate (CSIx))**.** *Given random scales* $\mathbf{D}_i = diag(1/d_{i1}, \ldots, 1/d_{ip})$ *and the covariates* $\boldsymbol{X}_i$*, consider the conditional distribution of* $\mathbf{D}_i\boldsymbol{Y}_i$ *as Equation* (4.2) *with functional precision matrix* $\boldsymbol{\Omega}(\boldsymbol{X}_i)$*. If* $\omega^{j,k}(\boldsymbol{X}_i) = 0$*, then* Y_j *and* Y_k *are CSI. Otherwise, when* $\omega^{j,k}(\boldsymbol{X}_i) \neq 0$*, then* Y_j *and* Y_k *are CSD.*

The proof of Proposition 4.2.1 follows the fact that $\omega^{j,k}(\boldsymbol{X}_i) = 0$ implies the CSI of Y_j and Y_k given $\boldsymbol{X}_i$, and we call Y_j and Y_k are conditional sign independence with covariates $\boldsymbol{X}_i$ to highlight the role of the covariates in the graph. Otherwise, Y_j and Y_k are called CSDx.

Illustrative example. We use a simple low-dimensional example to visually demonstrate and interpret CSIx and CSDx. Following Proposition 4.2.1, we show two examples with a general functional precision matrix $\boldsymbol{\Omega}(\boldsymbol{X}_i)$. Consider a trivariate distribution of (4.2) with unit diagonal elements and $\omega^{1,2}(X_i) = X_i$. We illustrate

two scenarios shown in Panel (B) of Figure IV.2:

- When $X_i = 0$, we obtain the CSIx of Y_1 and Y_2 given two different values of $Y_3 = 0$ (Case (i)) and 1 (Case (ii)).
- When $X_i = 0.7$, Y_1 and Y_2 are CSDx and we observe that the distribution of the sign of Y_2 varies based on the value of Y_1 (see Case (iii) and (iv)). Specifically, as Y_1 increases, Y_2 tends to be negative.

By modeling the functional precision matrix, we can build covariate-specific precision matrix that depends on the different realization of the covariates $\boldsymbol{X}_i$. Consequently, we can construct a graph of CSI corresponding to the precision matrix and the covariates.

We can now conceptually compare models (4.1) and (4.2). Both models incorporate the covariates in the functional precision matrix, which characterizes the covariate-specific graph. However, the interpretation of the graph differs.

The graph from model (4.2) encodes CSIx whereas the graph from model (4.1) encodes CIx. We further visualize the relationship between CSIx and CIx in Panel (C) of Figure IV.2 and summarize as follows:

- For $\omega = 0$, CSIx is a weaker condition than CIx since CSIx only considers the sign while CIx depends on both the sign and the magnitude.
- When $\omega \neq 0$, CSDx is more stringent than CDx for CDx allows either magnitude or the sign to be dependent while CSDx only focuses on the sign.

In summary, the random scales robustify the normality assumption that is violated in our motivation data (Figure IV.1). The introduction of the random scales also causes the loss of CIx in graphical regression with the replacement of a weaker independence condition of the CSIx. Specifically, the CSIx manifests the independence on the sign in the probabilistic manner.

4.3 Priors and Estimation

The functional precision matrix $\boldsymbol{\Omega}(\boldsymbol{X}_i)$ lives in a high-dimensional space. For example, the PPI network for ovarian cancer from our application considers $n \times p \times (p-1)/2 = 197,620$ possible edges. Hence, we use a neighborhood selection procedure (Meinshausen and Bühlmann, 2006) to estimate the functional precision matrix that has been used in several graphical modeling approaches (e.g. Ni et al., 2019; Zhang and Li, 2022). This procedure offers three main benefits: (i) tractable estimation, (ii) reduced computation burden, and (iii) flexible prior elicitation. Specifically, we regresses one node Y_j on the rest nodes $Y_k, k \neq j$ and build the graph based on zero coefficients (Section 4.3.1 and 4.3.2). By employing neighborhood selection, we reduce the number of edges to $q \times p \times (p-1)/2 = 3,280$. Additionally, the number of edges can be further reduced by different model specification like thresholding mechanism (Section 4.3.3) and different priors such as spike-and-slab (Section 4.3.4).

4.3.1 Regression-based Approach for Functional Precision Matrix Estimation

The rBGR model leverages a regression-based framework on model (4.2) to relate the regression coefficients and precision matrix. Given random scales $\boldsymbol{D}_i$, we regress one variable on all other variables and relates the partial correlation with regression coefficients. Zero coefficients is then equivalent to zero partial correlations (Meinshausen and Bühlmann, 2006). Specifically, we define the rBGR as:

$$\frac{Y_{ij}}{d_{ij}} = \sum_{k \neq j}^{p} \beta_{j,k}(\mathbf{X_i}) \frac{Y_{ik}}{d_{ik}} + \epsilon_{ij}, \tag{4.3}$$

where $\epsilon_{ij} \sim N(0, 1/\omega^{j,j}(\boldsymbol{X}_i))$ and the functional coefficient $\beta_{j,k}(\boldsymbol{X}_i) = -\frac{\omega^{j,k}(\boldsymbol{X}_i)}{\omega^{j,j}(\boldsymbol{X}_i)}$. Under this specification, $\beta_{j,k}(\boldsymbol{X}_i) = 0$ if and only if $\omega^{j,k}(\boldsymbol{X}_i) = 0$, which enables

the functional coefficients to characterizes the covariate-specific graphs. However, the interpretation of the coefficients changed from the standard Gaussian graphical models (Meinshausen and Bühlmann, 2006) due to the introduction of the random scales, which will be detailed in the next subsection.

4.3.2 Graph Construction through Regression Coefficients

We build graphs with a missing edge between node j and k when Y_j and Y_k are CSIx given the remaining variables and the covariates $\boldsymbol{X}_i$. Consider $\boldsymbol{Y}_i$ and $\boldsymbol{X}_i$ with the regression (4.3). We call $\beta_{j,k}(\boldsymbol{X}_i)$ the conditional sign independence function (CSIF) because zero CSIF $\beta_{j,k}(\mathbf{X}_i) = 0$ implies that Y_j and Y_k are CSIx given all the other nodes $\boldsymbol{Y}_{-\{j,k\}}$ and covariates $\boldsymbol{X}_i$, as formally characterized in the following proposition.

Proposition 4.3.1. *Consider the data* $\boldsymbol{Y}$ *and* $\boldsymbol{X}$ *with model* (4.3). *If* $\beta_{j,k}(\mathbf{X}_i) = 0$, *then* $\mathcal{P}(Y_j > 0 \mid Y_k, \boldsymbol{Y}_{-\{j,k\}}, \mathbf{X}_i) = \mathcal{P}(Y_j > 0 \mid \boldsymbol{Y}_{-\{j,k\}}, \mathbf{X}_i)$ *and* $\mathcal{P}(Y_k > 0 \mid Y_j, \boldsymbol{Y}_{-\{j,k\}}, \mathbf{X}_i) = \mathcal{P}(Y_k > 0 \mid \boldsymbol{Y}_{-\{j,k\}}, \mathbf{X}_i)$.

We sketch the proof and leave the details in Supplementary Section C.1. The proof follows from the fact that the CSIF $\beta_{j,k}(\boldsymbol{X}_i) = -\frac{\omega^{j,k}(\boldsymbol{X}_i)}{\omega^{j,j}(\boldsymbol{X}_i)}$ is related to the partial correlation, and a zero partial correlation is equivalent to a zero precision of $\omega^{j,k}(\boldsymbol{X}) = 0$, which ensures the CSIx between Y_j and Y_k (see the example in Section 4.2.3). Therefore, zero CSIF indicates the CSIx between Y_j and Y_k given the remaining response variables $\boldsymbol{Y}_{-\{j,k\}}$ and covariates $\boldsymbol{X}_i$. In this Chapter, we further assume the scalar diagonal precision of $\omega^{j,j}$ in CSIF as $\beta_{j,k}(\boldsymbol{X}_i) = -\frac{\omega^{j,k}(\boldsymbol{X}_i)}{\omega^{j,j}}$ to improve the computation. The CSIF is zero if and only $\omega^{j,k}(\boldsymbol{X}_i) = 0$, which is unrelated to the diagonal elements and our main interest of edge selection.

4.3.3 Modeling the Conditional Sign Independence Function

Proposition 4.3.1 transforms the problem of robust graph construction to a more tractable regression coefficient selection (i.e., selecting which part of CSIF is exactly zero). Therefore, modeling the CSIF is crucial to the graph estimation. To this end, we parameterize the CSIF as a product of two components:

$$\beta_{j,k}(\boldsymbol{X}_i) = \underbrace{\theta_{j,k}(\boldsymbol{X}_i)}_{\text{Covariate function}} \quad \underbrace{\mathbb{I}(\mid \theta_{j,k}(\boldsymbol{X}_i) \mid > t_j)}_{\text{Thresholding function}}. \tag{4.4}$$

We elaborate the role and justification of each component below.

Covariate functions $[\theta_\bullet(\bullet)]$. For exposition, we consider only the linear effects of covariates $\boldsymbol{X}_i$, $\theta_{j,k}(\boldsymbol{X}_i) = \sum_{h=1}^{q} \alpha_{j,k,h} X_{ih}$, where $\alpha_{j,k,h}$ represents the coefficients for the h-th covariate. The covariate function allows similar edge sets for individuals with a similar level of $\boldsymbol{X}_i$ and varies the graph thus borrowing strength. If desired, it is relatively straightforward to extend it to nonlinear effects with e.g., using basis expansion techniques such as splines.

Thresholding functions $[\mathbb{I}(\mid \theta_\bullet(\boldsymbol{X}) \mid > t_\bullet)]$. The edge thresholding mechanism is desired to achieve sparse graph in rBGR due to the large number of parameters. For example, the ovarian PPI network in our application requires $qp(p-1)/2 = 3,280$ parameters and results in a dense graph with inefficient inference. To solve the problem, we truncate edges with small magnitudes with an indicator function $\mathbb{I}(\mid \theta_{j,k}(\boldsymbol{X}) \mid > t_j)$, where t_j is the threshold parameter specific to the node j. An edge is shrunk to zero and removed when the magnitude is smaller than the threshold parameter, resulting in a sparse graph. One might consider threshold parameter as $t_{j,k}$. However, $t_{j,k}$ is not fully identifiable when $\alpha_{j,k,h} = 0$ for all $h = 1, \ldots, q$ since when $\theta_{j,k}(\boldsymbol{X}_i) = 0$, the value of $t_{j,k}$ can be arbitrary. To alleviate the problem, we assume $t_{j,k} = t_j$ to improve the identifiability as long as one of of $\theta_{j,k} \neq 0$.

4.3.4 Prior Specification

To complete the model specification, rBGR contains three parameters: (a) random scales d_j, (b) covariate coefficients $\alpha_{j,k,h}$, and (c) threshold parameter t_j. Specifically, we assign priors as follows:

$$d_j \sim (1-\pi_j)\delta_1(d_j) + \pi_j p_j(d_j);\ \alpha_{j,k,h} \sim \text{Spike-and-slab};\ t_j \sim \text{Unif}(0, t_{\max}), \quad (4.5)$$

where $t_{\max}$ is a pre-specified hyper-parameter, π_j models the degree of non-normality with beta prior as $\pi_j \sim Beta(a_\pi, b_\pi)$, and p_j is a function to accommodate the non-normality. Specifically, when $d_j = 1$, Y_j is normally distributed. When $d_j \sim p_j$, Y_j follows a non-normal distribution. We match tail behavior of p_j and the marginal distribution of Y_j and allow each marginal distribution Y_j to have different level of non-normality by specific π_j. For the current model, we focus on the Y_j with polynomial decay as illustrated by the motivating data in Figure IV.1 and assign a inverse gamma distribution on $p_j(d_j^2) \sim InvGa(a_d, b_d)$. For covariate coefficients $\alpha_{j,k,h}$, we assign a spike-and-slab prior to achieve the covariate sparsity because not all covariates necessarily contribute to the varying structure of our graph. For threshold parameter t_j, we assign a uniform prior on t_j to model the thresholding mechanism and control the graph sparsity. Intuitively, when $t_j \to 0$, no edge is truncated and results in a fully connected graph. When $t_j \to \infty$, all edges are shrunk to zero with all nodes disconnected.

4.4 Posterior Inference

Gibbs sampler. In this Section, we introduce an efficient Gibbs sampler for the proposed rBGR model. Instead of the Metropolis-Hastings algorithm, we implement a Gibbs sampler except for the random scales and largely improve the computation and

convergence compared to Ni et al. (2019). Recently, Li et al. (2023+) derived a closed-form of the conditional distributions for Gibbs sampler by formulating the thresholded coefficients as mixture distributions. Specifically, if we view the distribution with one component as a special case of mixture distribution, the mixture distribution from the thresholded coefficient then can achieve conjugacy. We derive the full condition distribution for parameters for covariate coefficients $\alpha_{j,k,h}$ and the threshold parameter t_j, and the full conditions of covariate coefficients and threshold parameter belong to the mixture of truncated normal and the mixture of uniform distribution, respectively. By assigning normal priors on covariate coefficients and a uniform prior on threshold parameter, we obtain conjugacy on all thresholded parameters. We further use the parameter expansion technique (Geyer, 2011) on covariate coefficients to improve the mixing of MCMC. We implement the Metropolis-Hasting algorithm for the random scales.

Covariate and edge selection. The estimated coefficients from rBGR of (4.3) do not guarantee the symmetry required in the undirected graph. Also, due to the introduction of random scales with the CSIx characterization, we focus on the sign of the edge. In this Section, we describe algorithms to symmetrize the estimated covariate coefficients $\widehat{\alpha}_{j,k,h}$ and the sign of graph edges of $\widehat{\beta}_{j,k}(\boldsymbol{X}_i)$. For covariate coefficients, we compare the posterior inclusion probability (PIP) of directed coefficients from two directions ($\widehat{\alpha}_{j,k,h}$ and $\widehat{\alpha}_{k,j,h}$) and assign the undirected coefficients as the directed coefficient with a smaller PIP. Given a cutoff c_0, the rule above requires both directions of coefficients to have PIPs bigger than c_0 implying a network with less edges. For the edge, we symmetrize the edge based on the edge posterior probability (ePP). Specifically, we symmetrize the undirected ePP by taking the maximum of the two directed ePP. Given a cutoff c_1, we call an undirected edge if at least one of the directed ePPs is bigger than c_1. We then decide the sign of the edge by comparing the posterior probability of positive and negative for the chosen direction.

We offer more details of the posterior inference in Supplementary Material Section C.2 including (i) Gibbs sampler derivation with the Algorithm (Section C.2.2) and (ii) symmetrization rule for both covariate coefficients and edges (Section C.2.3).

4.5 Simulation Studies

We empirically demonstrate the performance of rBGR under a variety of non-normal settings and against other competing models in terms of edge and covariate selection. To the best of our knowledge, no other existing method estimates covariate-specific graphs for non-normal data. Therefore, we compare rBGR to two models that estimate the covariate-specific graph without addressing the violation of normality assumption. Specifically, we consider Bayesian graphical regression (BGR) (Ni et al., 2019) and the Gaussian graphical model regression (RegGMM) (Zhang and Li, 2022) representative of a fully Bayesian and a frequentist penalization-based models for the covariate-specific graph under normal assumption, respectively. For RegGMM, we run the algorithm with various tuning parameters to obtain the probability of the signs of edges and covariate coefficients and select the optimal tuning parameter by cross validation using their default algorithm. For rBGR and BGR, we symmetrize the graph mentioned in Section 4.4 and set $c_0 = c_1 = 0.5$. We run $10,000$ and $30,000$ iterations and discard the first 90% iterations for rBGR and BGR, respectively.

Data generating mechanism. We generate the observed non-normal data by multiplying the random scale to the latent normal data $\boldsymbol{Y}_i^* = \left[Y_{i1}^*, \ldots, Y_{ip}^*\right]^\mathsf{T}$ that follows an multivariate normal distribution with a functional precision matrix that represents the undirected graph. Specifically, we generate the covariates $\boldsymbol{X}_i \overset{iid}{\sim} U(-1, 1)$ and latent data $\boldsymbol{Y}_i^* \overset{iid}{\sim} \boldsymbol{N}_p(\boldsymbol{0}, \boldsymbol{\Omega}^{-1}(\boldsymbol{X}_i))$ with the true precision matrix $\boldsymbol{\Omega}(\boldsymbol{X}_i)$. For $\boldsymbol{\Omega}(\boldsymbol{X}_i)$, we assign unit diagonal elements and randomly pick 2% of the off-diagonal to be non-zero. We let the non-zero precision depend on the covariates linearly and truncate the precision with a magnitude smaller than 0.15. We obtain the random scales from

a mixture distribution of the point mass at one and a inverse gamma distribution and assign three different levels non-normal contamination: $\pi \in \{0, 0.5, 0.8\}$. We multiply the random scales to generate the observed data of $[Y_{i1}, \ldots, Y_{ip}] = [Y^*_{i1}d_{i1}, \ldots, Y^*_{ip}d_{ip}]$. For all simulations, we set the sample size and the dimensions of $\boldsymbol{Y}_i$ and $\boldsymbol{X}_i$ as $(n, p, q) = (250, 50, 3)$. We show the results for 50 independent replicates.

Performance metrics. We evaluate the graph recovery through the edge and covariate selection. For covariate selection, we report the true positive rate (TPR), true false rate (TFR), and Matthew's correlation coefficient (MCC) with the cut-off for PIP at $c_0 = 0.5$. We also report the area under the ROC curve (AUC) and partial area under ROC curve (pAUC) between specificity ranging from 0.8 to 1. For edge selection, we show AUC and three metrics of TPR, TNR and MCC with the cut-off for ePP at $c_1 = 0.5$. We further show the sign consistency by examining the agreement between the posterior probability for the signs of CSIF $\text{sgn}(\widehat{\beta}_{j,k}(\boldsymbol{X}_i))$ and the true signs of $\text{sgn}(\beta_{j,k}(\boldsymbol{X}_i))$. Specifically, we exclude the zero CISF and focus on the subset of the data with both true and estimated non-zero CSIF to restrict the problem as two-class classification (positive versus negative). We assess the sign consistency by MCC (referred to as sign-MCC).

Simulation results. Panel (A) of Figure IV.3 shows the simulation results for covariate selection. We observe that rBGR outperforms BGR and RegGMM across all non-normality levels, as indicated by higher MCC and AUC. The difference of MCC and AUC between rBGR and the rest competing methods increases when the non-normality level increase. For TNR, rBGR performs slightly worse than BGR but better than RegGMM across all non-normality levels. However, all three methods select correct covariates ($> 93\%$) with small difference ($< 5\%$) in terms of TNR. For TPR, rBGR outperforms BGR under all levels of non-normality and the advantage of rBGR becomes more prominent as the non-normality increases. Compared to RegGMM, rBGR's performance is comparable under normal distribution in TPR,

but rBGR is preferred when the level of non-normality increases. Overall, modeling the non-normality from random scales in rBGR is favored compared to models without random scales in terms of covariate selection.

We show the graph recovery for the edge selection in Panel (B) of Figure IV.3. For edge selection, rBGR outperforms BGR and RegGMM in AUC, and the advantage of rBGR increases with a larger discrepancy between rBGR and the competing methods when the non-normality level increases. For MCC, rBGR outperforms RegGMM under all levels of non-normality, but is slightly inferior than BGR under the normal distribution. However, rBGR is favored when the non-normality level increases. For TPR, rBGR is better than BGR under all levels of non-normality, and slightly worse than RegGMM under normal assumption. However, when non-normality increases, rBGR starts to surpass the RegGMM. Both TNR and sign-MCC show excellent selection performance ($> 95\%$) for all three methods, with minimal differences ($< 5\%$) across the three non-normality levels. In summary, modeling the non-normality through random scales in rBGR result in equivalent (under normal distribution) or better performances in all metric for edge selection compared to the other methods.

Additional simulations and model evaluations. We provide additional simulation of data generating mechanism and model evaluation results for (i) convergence of the algorithm, and (ii) different cut-off of c_0 and c_1 controlling for false discovery rates – which are summarized in Supplementary Material Section C.3.

4.6 Analyses of Proteomic Networks under Immunogenic Heterogeneity

Key scientific questions and dataset overview. Aberrant protein-protein interactions (PPIs) are associated with various diseases including cancer (Lu et al., 2020), and immune cells around the tumor can modulate malfunctioning PPIs to influence

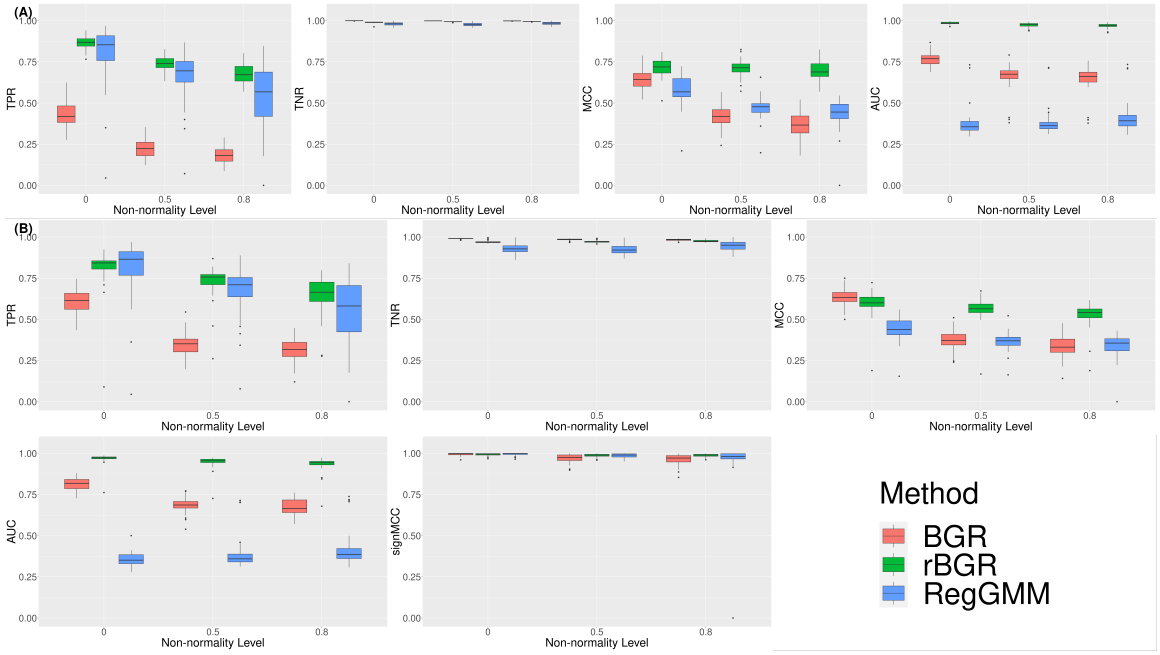


Figure IV.3: Graph recovery for BGR (red), rBGR (green) and RegGMM (blue) under different levels of non-normality in terms of (A) covariates selection (top row) and (B) edge selection (bottom two rows). Panel (A) measures the covariate selection through four metrics (from left to right: TPR, TNR, MCC and AUC) are measured under three different levels of non-normality. Panel (B) demonstrates the edge selection by four criteria (from upper left to lower right: TPR, TNR, MCC, AUC) and the sign consistency by sign-MCC (lower left) for non-zero edges. All values for TPR, TNR and MCC are measured at a cut-off at $c_0 = c_1 = 0.5$.

tumor growth and progression (Joyce and Fearon, 2015). In cancer, cells around the tumor form the tumor microenvironment (TME) that closely interacts with the tumor (Whiteside, 2008). For example, the dysregulated PPIs in tumor suppress multiple immune cells in TME to escape the detection from immune system (Whiteside, 2008) while immune cells in TME can alter the aberrant PPIs to eliminate cancerous cells (Joyce and Fearon, 2015).

This demonstrates the connection between the dysregulated PPIs and the TME and shows the importance of immunogenic heterogeneity in tumor behavior. A better understanding of the impact of the immune cells on aberrant PPIs offers a foundational paradigm for potential targeted therapies in cancer (Cheng et al., 2020). To this end, our key scientific questions were as follows: (i) identify important PPIs across different cancer types and (ii) discover the effect of immunogenic heterogeneity on aberrant PPIs as potential targets for future investigation.

We exemplify the utility of rBGR, using data from The Cancer Genome Atlas (TCGA) to build patient-specific PPI networks and investigate the impact of immunogenic heterogeneity across two different cancers. Specifically, we used reverse-phase protein array for proteomic data ($\boldsymbol{Y}$) to build the PPI network of CSIx graph and incorporated the immune cell transcriptome signatures as covariates ($\boldsymbol{X}$) as marker of immunogenic heterogeneity. Our analysis focuses on ovarian cancer (OV) and lung adenocarcinoma (LUAD) as representative examples of two different types of cancers that elicit distinct immune responses. OV represents a immunologically "cold" tumor with a weaker immune response, while LUAD is considered a immunologically "hot" tumor with a stronger immune response (Galon and Bruni, 2019).

We focus on proteins in 12 important cancer-related pathways (Ha et al., 2018) and obtained $p = 41$ proteins with $n = 241$ and $n = 360$ patients for OV and LUAD, respectively. For covariates, we included mRNA-derived immune cell gene signatures and quantified the immune cell abundance corresponding to T cells and two crucial

members of myeloid-derived suppressor cells, monocytes and neutrophils, for both OV and LUAD. Both T cells and myeloid-derived suppressor cells are essential in both OV and LUAD since T cells is the main immune component that kills cancer cells while myeloid-derived suppressor cells regulates T cells (Whiteside, 2008). We ran rBGR on OV and LUAD with $20,000$ iterations and discarded the first $19,000$ iterations. The convergence diagnostics and the details of data preprocessing procedures are provided in Supplementary Material Section C.4.1.

4.6.1 Population-Level Proteomic Networks

We first focus on the covariate dependent population-level networks for OV and LUAD that are estimated by $\widehat{\alpha}_{j,k,h}$. The corresponding networks are shown in in Figure IV.4 (Panels (A) for LUAD and (B) for OV). We observed that the number of edges is much less in OV compared to LUAD for all immune components (T cells: (7, 15), monocytes: (5, 82) and neutrophils: (7, 260) for (OV, LUAD)). This is further evidenced in Panel (C) that shows the distribution of PIPs for OV and LUAD. Interestingly, we observe that the PIPs for LUAD are higher than those for OV for all immune components (median of (OV, LUAD) for T cells: (0.123, 0.271), monocytes: (0.131, 0.307), and neutrophils: (0.127, 0.380)). The higher PIPs in LUAD imply that immune components have a greater impact on PPIs in LUAD compared to OV. This finding is consistent with the existing biology, as LUAD belongs to the immune hot tumors with a stronger immune response (Galon and Bruni, 2019). Furthermore, we identify HER2, Rb and Bax as the top three hub proteins with the highest degree in LUAD. In LUAD, HER2 mutation is associated inferior survival (Pillai et al., 2017), Bcl-2 family protein including Bax is a prognostic biomarker (Sun et al., 2017), and Rb mutation predicts poor clinical outcomes (Bhateja et al., 2019). For OV, AR is identified as a hub protein with the highest degree (AR: 13 with the rest protein ≤ 10). Recent evidence supports the critical role of AR for the progression of OV

(Zhu et al., 2017).

Population graphs also confer specific information about the interaction between proteins. For example, we observe an edge between Akt and PTEN with the highest PIP regulated by T cell for LUAD (Panel (A)) suggesting the impact from T cell on the PPI between Akt and PTEN. It is well-known that PTEN down-regulates Akt and the loss of tumor suppressor PTEN often leads to dysregulated PI3K pathway including Akt and the following tumor growth for LUAD (Conciatori et al., 2020). For OV, despite the smaller number of PPIs, we still identify PPIs that are consistent with existing literature. For example, rBGR suggests a PPI regulated by T cells between Caveolin-1 and PR. In OV, Caveolin-1 is regulated by progesterone, which is mediated by PR, and suggests a consitent result with the estimated PPI between Caveolin-1 and PR (Syed et al., 2005). Overall, our analyses capture important hub proteins and characterize the cancer PPIs, and the results are highly concordant with the existing cancerliterature.

4.6.2 Patient-Specific Networks

We next focus on patient-specific PPI networks to examine the effect of immune component abundance ($\boldsymbol{X}_i$) on PPIs. Specifically, we vary one immune component with the rest components fixed at their mean and generate networks of CSIx for different individuals at five percentiles (5th, 25th, 50th, 75th and 95th percentiles) of the varying immune component. We set the cut-off for the ePP at $c_1 = 0.5$ and show the networks for LUAD in Figure IV.5 with the networks for OV in Supplementary Material C.4.

For specific immune components, we focus on PPIs of CSDx showing that the PPIs are dependent on the abundance of specific immune component. We present PPIs that change the signs in the 5th and 95th percentiles indicating specific PPIs that are impacted by the immune components, such as Akt-PTEN for T cells, Bid-PCNA for

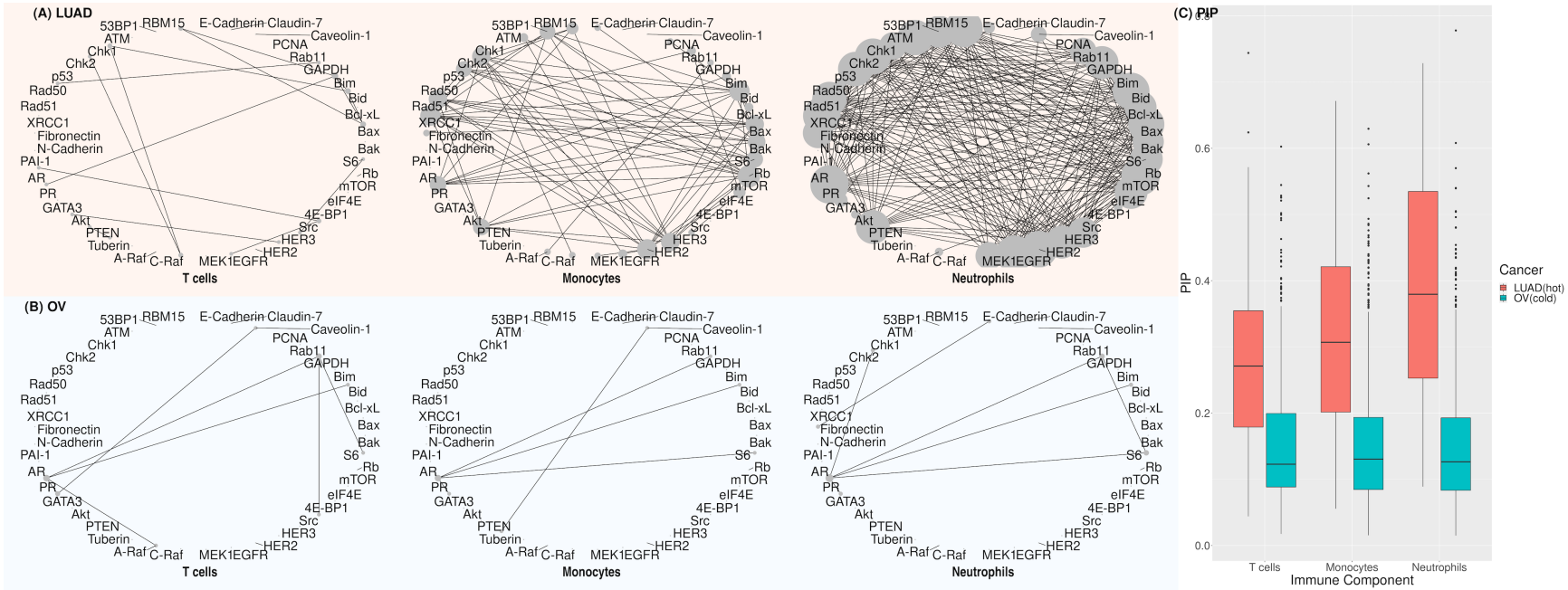


Figure IV.4: The posterior inclusion probability (PIP) (Panel (C)) and the population networks of PPIs with the cut-off at $c_0 = 0.5$ for LUAD (Panel (A)) and OV (Panel (B)). For each panel of LUAD and OV, PPI networks of specific immune component are shown from the left to right for T cells, monocytes, and neutrophils. The degree of each protein is shown by the node size with a bigger node representing a higher degree.

monocytes, and Bax-GATA3 for neutrophils. Interestingly, we discovered that the sign of Akt-PTEN is positively correlated to the T cell abundance. Specifically, when T-cell abundance is higher, Akt-PTEN is positive; vice-versa, Akt-PTEN is negative when T cell is scarce. It is well-established that PTEN suppresses Akt signaling and the loss of PTEN results in the hyper-activation of Akt in cancer cells and the low T cell abundance in lung cancer (Conciatori et al., 2020). In addition, we find Bid-PCNA edge is positively correlated with monocytes abundance. It has been shown that PCNA promotes Bid through caspase proteins and is crucial to immune evasion in cancers (Wang et al., 2021). Finally, we discover that Bax-GATA3 edge is positively correlated with neutrophil abundance. Recently, GATA3 has been found to down-regulate BCL-2 (Cohen et al., 2014), which inhibits the Bax protein (Antonsson et al., 1997), and neutrophils promotes the Bax to induce the apoptosis (Li et al., 2020). These findings highlight specific PPIs that are influenced by the abundance of immune components and suggest potential targets for further investigation of immunotherapy for lung cancer.

4.7 Discussion

In this Chapter, we develop a flexible Bayesian framework called robust Bayesian graphical regression (rBGR) to construct heterogeneous networks that accounts for covariate-specific information for non-normally distributed data. By accommodating the non-normal marginal tail behaviours through random scales, we construct covariate-specific graph through graphical regression-based approaches and characterize the edge dependencies through conditional sign independence (CSIx). For a specific covariate, the CSIx of two variables ensures the distribution of the sign of one variable is not affected by the information from the other variable given the remaining variables. Specifically, given random scales and covariates, we build a underlying multivariate Gaussian distribution and model the covariate-specific graph via

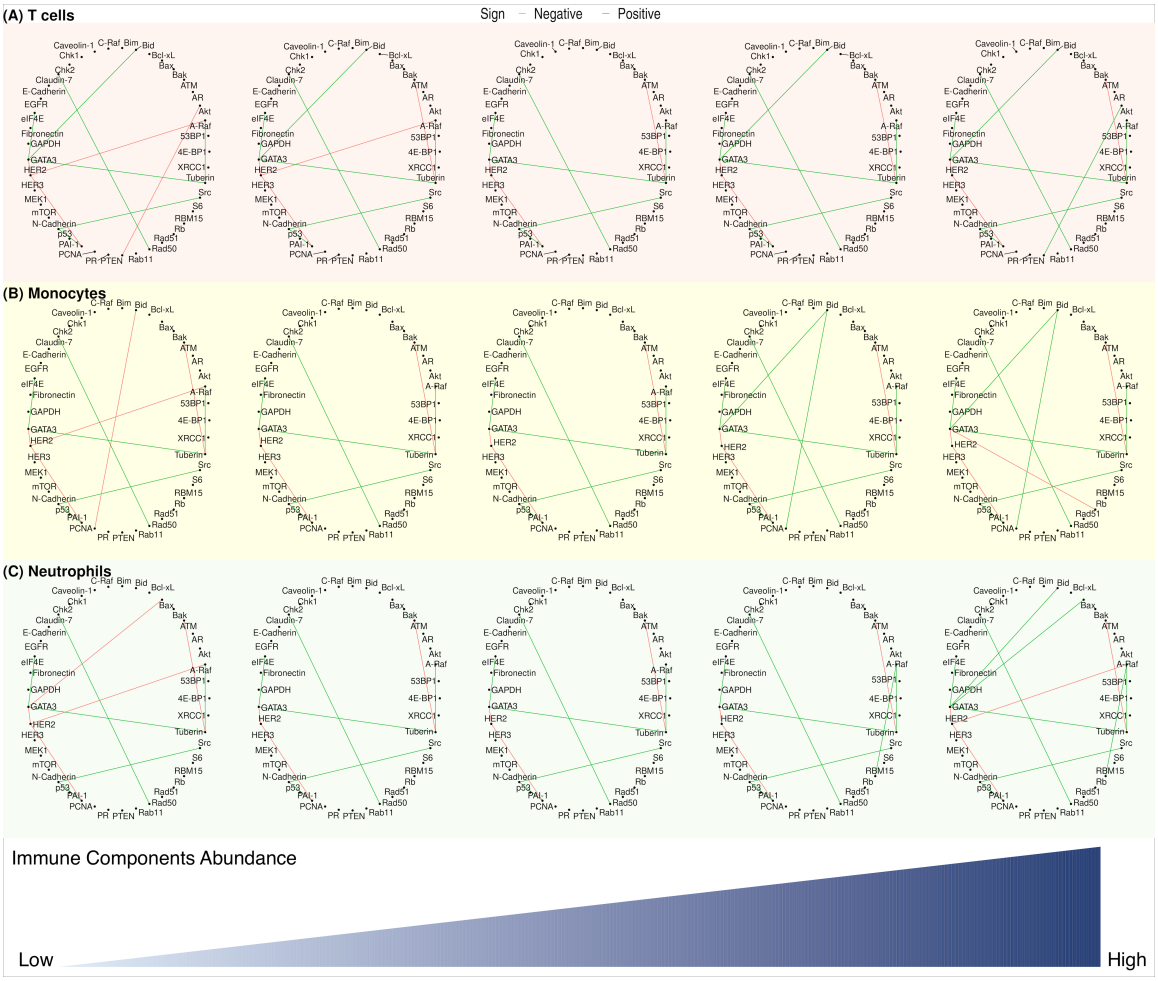


Figure IV.5: Networks of LUAD under five different percentiles immune component of (A) T cells, (B) monocytes and (C) neutrophils with the rest two components fixed at mean zero. The estimated network for varying immune components are shown from the left to right for 5, 25, 50 ,75, and 95-th percentiles. Edges are identified with signs (green: positive and red: negative) when the ePPs are bigger than $c_1 = 0.5$.

a functional precision matrix in (4.2). We estimate the functional precision matrix through regressions and relate the CSIx with the functional coefficients (CSIF) that vary with covariates. From Proposition 4.3.1, zero CSIF implies missing edges and we build covariate-specific graphs based on zero CSIF. We also propose an efficient Gibbs sampler for posterior inference. Empirically, rBRG outperforms other existing methods that construct the covariate-specific graphs under varying non-normality levels.

We employ rBGR on proteogenomic datasets in two cancers to build patient-specific PPI network and identify PPIs that are impacted by tumor heterogeneity. Specifically, we quantify the immune cell abundance to discover immunogeneic heterogeneity on aberrant PPI for lung and ovarian cancers that are triggered by different levels of immune responses. Our analyses align with existing biology along three major axes: (i) immune responses, (ii) hub proteins, and (iii) PPIs. For example, higher connections in LUAD are consistent with existing biology since LUAD belongs to the class of the immunlogically "hot" tumors. We identify a hub protein of HER2, which is associated with a poor survival in LUAD. Another example is a PPI of Akt-PTEN and PTEN down-regulates Akt. Our study further suggests PPIs that vary with specific immune component. For example, we discover PPIs of Akt-PTEN, Bid-PCNA, and Bax-GATA3 that varies positively on T cells, monocytes and neutrophils, respectively. These findings suggest potential future targets for immunotherapy in lung cancer.

In the current implementation of rBGR, we construct the random scales with two main underlying assumptions: (i) the independence among different random scales, and (ii) a mixture of parametric distributions that matches the tail behavior for each marginal distribution. Both assumptions can be generalized to include a broader class of non-normal distributions. For example, we can jointly consider a multivariate random scales without assuming independence. For the parametric assumption of

matching tail behavior, one can consider a nonparametric transformation such as the nonparanormal transformation with basis expansion (e.g. Mulgrave and Ghosal, 2022) or the copula model with the empirical cumulative density function (e.g. Dobra and Lenkoski, 2011). However, all these generalizations above impose difficulties in both interpretations and computations, especially when incorporating the subject-specific covariates. For the model aspect, we currently consider only linear effect of covariates to reduce the inferential and computation burden. It is possible to include the non-linear functionals through basis expansion techniques such as splines (Ni et al., 2019) – however this will increase the computational burden of fitting rBGR. Another possible extension is other types of graphs. For example, chain graph considers ordered multi-level structure via directed and undirected edges(e.g. Chakraborty et al., 2021). By introducing the random scales and generalizing the regression coefficients as functional coefficients, the model can include the covariates in the precision matrix to build the subject-specific chain graphs. Another direction could be include discrete nodes and the concept of CSIx can be extended for discrete data (Bhadra et al., 2018). All these directions are left for future investigations.

Code and Data Availability. We also provide a general purpose code in `R` that accompanies this Chapter along with all the necessary documentation and datasets required to replicate our results (see `https://github.com/bayesrx/rBGR`).

CHAPTER V

Summary and Future Directions

In this dissertation, I construct a family of Bayesian models tailored for structured covariances. More precisely, these models and the associated structured covariances are mainly devised to address two different biological dependencies (tree- and graph-based dependencies) that originated from the underlying scientific contexts in cancer research. In Chapter II, I formulate the dependency among different treatment mechanisms as a tree-structure covariance and measure the mechanism similarities based on the tree structure to infer the treatment effectiveness. Chapter III extends the exploration of ultrametric matrices and proposes a consistent Markovian prior for ultrametric matrices along with an efficient algorithm, providing uncertainty quantification alongside point estimates. In Chapter IV, I focus on graph-based dependency structures and construct the covariate-specific proteomic networks to address the immunogeneic heterogeneity using the non-normally distributed protein expression data. In summary, this dissertation spans two distinct biological dependencies and encompasses various covariance structures that capture the underlying scientific hypotheses to provide coherent estimation, inference and interpretations.

Several assumptions used in this dissertation can be generalized for both structured covariances in this dissertation, including the underlying theoretical considerations and computations. We describe the potential future directions for each of the

structured-covariances.

Tree-based covariance. One evident assumption is the normality assumption used in the tree structures of Chapter II and III. Since the ultrametric inequality remains valid under monotone transformations (McCullagh, 2006), utilizing random scale becomes a potential strategy to accommodate non-normality and explore larger classes of resulting distributions. Furthermore, these assume a common tree structure covariance is assigned to all subjects. This assumption can also be further generalized to allow the tree structures to vary based on different subject-specific information (e.g. covariates). Following the development of this dissertation, it is plausible to represent a covariate-specific tree structure as a functional ultrametric matrix with each element as a function that varies on different covariates. However, ensuring that the ultrametric inequalities are satisfied for every potential covariate realization presents challenges both for computation and interpretation.

Graph-based covariance. Regarding the graph-based covariances of Chapter IV, we estimate the graphs through the graphical regression model with the linear effect of the covariates. One potential generalization is to incorporate the non-linear functional of the covariates by the basis expansion techniques such as splines (e.g. Ni et al., 2019) or the kernel-based techniques (e.g. Liu et al., 2010). However, both methods increase the computation burden due to a larger number of possible parameters (depending on the parameterization). In the current implementation, we accommodate the non-normality by random scales with two main assumptions: (i) the independence among different random scales, and (ii) a mixture of parametric distributions that matches the tail behavior for each marginal distribution. It is possible to include a broader class of non-normal distributions by generalizing two assumptions above. For example, one can consider a joint random scale without assuming the independence among different random scales. For the parametric assumption of

matching tail behavior, we can construct a nonparametric transformation such as the nonparanormal transformation with basis expansion (e.g. Mulgrave and Ghosal, 2022) or the copula model with the empirical cumulative density function (e.g. Dobra and Lenkoski, 2011). However, all these generalizations would require non-trivial generalizations of our existing methodology and additional care needs to be taking with respect to the model interpretations and computations, especially when incorporating the subject-specific covariates.

All these directions generalize the current models used in this dissertation and remain open for future investigations.

APPENDICES

APPENDIX A

Appendix of Chapter II

A.1 Proof of Proposition 1

We provide a proof for a tree with four leaves (see Figure A.1) and extension to trees with a larger number of leaves follows by induction. The main idea is to merge subtrees backward and integrate out responses of internal nodes when merging subtrees.

Proof. Consider a subtree $\mathcal{T}'$ rooted at $(t_1, \boldsymbol{X}_1')$ with two leaves $(1, \boldsymbol{X}_1)$ and $(1, \boldsymbol{X}_2)$, and one internal node $(t_2, \boldsymbol{X}_2')$ (see Panel (A) of Figure A.1). Assume that the root $(t_1, \boldsymbol{X}_1')$ of the subtree is fixed, and responses $\boldsymbol{X}_i, \boldsymbol{X}_i' \in \mathbb{R}^J, J \geq 1, i = 1, 2$. With $\boldsymbol{t} = (t_1, t_2, t_3)^\mathsf{T}$, the conditional distribution for leaf responses would be $\boldsymbol{X}_i | \boldsymbol{X}_2', \mathcal{T}, \boldsymbol{t} \sim N_J(\boldsymbol{X}_2', (1 - t_2)\sigma^2 \boldsymbol{I}), i = 1, 2$. Since $\boldsymbol{X}_2' | \boldsymbol{X}_1', \mathcal{T}, \boldsymbol{t} \sim N_J(\boldsymbol{X}_1', (t_2 - t_1)\sigma^2 \boldsymbol{I})$, based on the conjugacy of the normal distribution, the marginal distribution is also normal. Conditional on $\boldsymbol{t}$ and $\mathcal{T}$, mean and covariance of $\boldsymbol{X}_i, i = 1, 2$ can be derived by the law of iterated expectations and results in the distribution of the subtree $\mathcal{T}'$ with two leaves:

$$
\begin{aligned}
E[\boldsymbol{X}_i] &= E[E[\boldsymbol{X}_i|\boldsymbol{X}_2']] = E[\boldsymbol{X}_2'] = \boldsymbol{X}_1', \quad i = 1, 2; \\
Var[\boldsymbol{X}_i] &= Var[E[\boldsymbol{X}_i|\boldsymbol{X}_2']] + E[Var[\boldsymbol{X}_i|\boldsymbol{X}_2']] \\
&= Var[\boldsymbol{X}_2'] + E[(1-t_2)\sigma^2\boldsymbol{I}] = (1-t_1)\sigma^2\boldsymbol{I}_J; \\
Cov[\boldsymbol{X}_1, \boldsymbol{X}_2] &= Cov[E[\boldsymbol{X}_1|\boldsymbol{X}_2'], E[\boldsymbol{X}_2|\boldsymbol{X}_2']] + E[Cov[\boldsymbol{X}_1, \boldsymbol{X}_2|\boldsymbol{X}_2']] \\
&= Var[\boldsymbol{X}_2'] + E[0] = (t_2 - t_1)\sigma^2\boldsymbol{I}_J;
\end{aligned}
$$

The marginal distribution for the subtree $\mathcal{T}'$ with two leaves is

$$
\begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 \end{bmatrix} \sim \mathsf{MN}_{J\times 2}\left(\begin{bmatrix} \boldsymbol{X}_1' & \boldsymbol{X}_1' \end{bmatrix}, \boldsymbol{I}_J, \sigma^2\boldsymbol{\Sigma}^{\mathcal{T}'} \right), \quad \boldsymbol{\Sigma}^{\mathcal{T}'} = \begin{bmatrix} 1-t_1 & t_2-t_1 \\ t_2-t_1 & 1-t_1 \end{bmatrix}.
$$

Therefore, we can merge two leaves responses $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. Similarly, we can also merge the other subtree $\mathcal{T}''$ to obtain.

$$
\begin{bmatrix} \boldsymbol{X}_3 & \boldsymbol{X}_4 \end{bmatrix} \sim \mathsf{MN}_{J\times 2}\left(\begin{bmatrix} \boldsymbol{X}_1' & \boldsymbol{X}_1' \end{bmatrix}, \boldsymbol{I}_J, \sigma^2\boldsymbol{\Sigma}^{\mathcal{T}''} \right), \quad \boldsymbol{\Sigma}^{\mathcal{T}''} = \begin{bmatrix} 1-t_1 & t_3-t_1 \\ t_3-t_1 & 1-t_1 \end{bmatrix}.
$$

Eventually, we can merge two subtrees (see Panel (B) of Figure A.1), $\mathcal{T}'$ and $\mathcal{T}''$. From conjugacy of the normal distribution, the resulting joint marginal distribution of $\boldsymbol{X}_i, i = 1, 2, 3, 4$ is normal. The mean and the variance can be derived along identical lines as above. The only term left is the covariance, and we need to (re-)compute

them for locations within and between the combined subtrees. Explicitly,

$$
\begin{aligned}
Cov[\boldsymbol{X}_1, \boldsymbol{X}_2] &= Cov[E[\boldsymbol{X}_1|\boldsymbol{X}_1'], E[\boldsymbol{X}_2|\boldsymbol{X}_1']] + E[Cov[\boldsymbol{X}_1, \boldsymbol{X}_2|\boldsymbol{X}_1']] \\
&= Var[\boldsymbol{X}_1'] + E[(t_2 - t_1)\sigma^2 \boldsymbol{I}_J] = t_2\sigma^2 \boldsymbol{I}_J \\
Cov[\boldsymbol{X}_1, \boldsymbol{X}_3] &= Cov[E[\boldsymbol{X}_1|\boldsymbol{X}_1'], E[\boldsymbol{X}_3|\boldsymbol{X}_1']] + E[Cov[\boldsymbol{X}_1, \boldsymbol{X}_3|\boldsymbol{X}_1']] \\
&= Var[\boldsymbol{X}_1'] + E[0] = t_1\sigma^2 \boldsymbol{I}_J \,.
\end{aligned}
$$

This ensures that

$$
\mathbf{X}^\mathsf{T} = \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 & \boldsymbol{X}_3 & \boldsymbol{X}_4 \end{bmatrix} \sim \mathsf{MN}_{J\times 4}\left(\begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}, \boldsymbol{I}_J, \sigma^2 \boldsymbol{\Sigma}^{\mathcal{T}} \right)
$$

$$
\boldsymbol{\Sigma}^{\mathcal{T}} = \begin{bmatrix} 1 & t_2 & t_1 & t_1 \\ t_2 & 1 & t_1 & t_1 \\ t_1 & t_1 & 1 & t_3 \\ t_1 & t_1 & t_3 & 1 \end{bmatrix},
$$

as required. Moreover, denote $t_{i,i'}$ as the most recent divergence time of leaves i and i'. We observe that $t_1 = t_{1,3} = t_{1,4} = t_{2,3} = t_{2,4}, t_2 = t_{1,2}$, and $t_3 = t_{3,4}$ and complete the Proposition 1. □

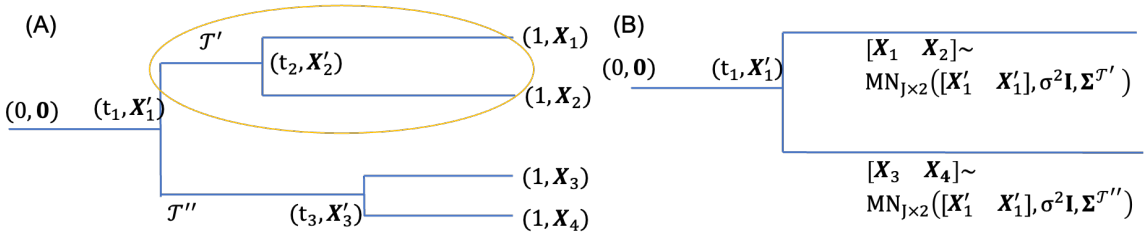


Figure A.1: Merging subtrees for the integration process. (A) First step of merging upper subtree, and (B) Final step of merging all subtrees.

A.2 Efficient Two-Stage Hybrid ABC-MH Algorithm

Here we offer details of two-stage algorithm with pseudo code. In the Section A.2.1, we describe the full algorithm of the ABC with the following posterior summary of Euclidean parameters (c, σ^2). The Section A.2.2 includes the implementation of the proposal function and the acceptance probability of MH stage. Pseudo code for the full two-stage algorithm is presented below in Algorithm 2

A.2.1 ABC Stage and the Posterior Summary of c and σ^2

The Section 3 of the Main Paper states the main idea of ABC and we offer the full algorithm of ABC including (i) the synthetic data generation process, (ii) the regression adjustment (Blum, 2010) of ABC, and (iii) posterior summary of the Euclidean parameters.

Data generation in ABC. Following Section 2 in the Main Paper, a synthetic data is generated from DDT as follows: (i) given $c_l \sim \mathsf{Gamma}(a_c, b_c)$, generate a tree $\mathcal{T}_l$ through the divergence function $a(t) = c_l(1-t)^{-1}$, and (ii) given $\mathcal{T}_l$ and $1/\sigma_l^2 \sim \mathsf{Gamma}(a_{\sigma^2}, b_{\sigma^2})$, generate triples $(t_j, \boldsymbol{X}_i', \boldsymbol{X}_i), i' = 1 \dots I-1, i = 1 \dots I$ by a scaled Brownian motion upon $\mathcal{T}_l$. After discarding $(\mathcal{T}_l, t_i, \boldsymbol{X}_i')$, the leaf locations $\boldsymbol{X}_i$ form an I by J observed data matrix $\mathbf{X}_l$. In Algorithm 2, ABC repeats the procedure above to generate N^{syn} synthetic data (see Figure A.2).

Regression adjustment in ABC. Originally proposed in Beaumont et al. (2002) and later generalized by Blum (2010), regression adjustment for ABC is performed in Step 8 of Algorithm 2. The motivation is to use smoothing technique to weaken the effect of the discrepancy between the summary statistic calculated from synthetic data and that from the observed data. We briefly describe the the procedure of c. Additional details can be found in Beaumont et al. (2002) and Blum (2010). Suppose we are given the observed summary statistics $\boldsymbol{S}_{\text{obs}}^{(c)}$ and unadjusted samples

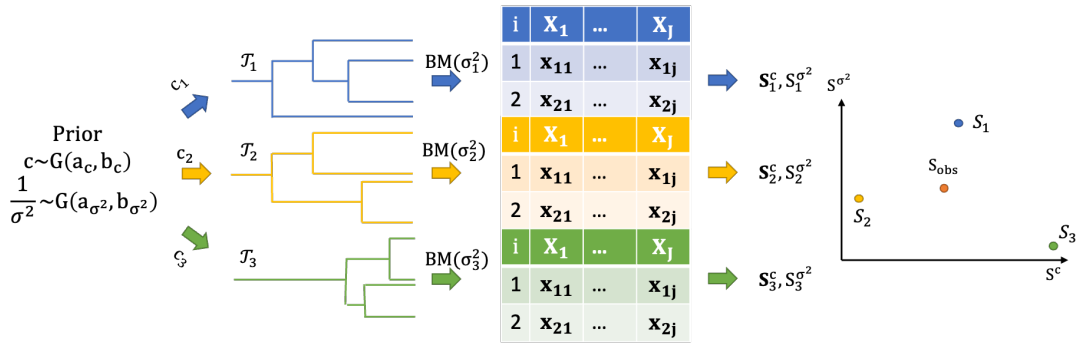


Figure A.2: Schematic diagram of synthetic data generation and the calculation of summary statistics (first stage of Algorithm 2). S_{obs} is calculated based on the actual observed data.

$(c_l^{\text{unadj}}, \boldsymbol{S}_l^{(c)}), l = 1, \ldots, k$, we can calculate the weight for each sample by

$$w_l^{(c)} = K_h(\|\boldsymbol{S}_l^{(c)} - \boldsymbol{S}_{\text{obs}}^{(c)}\|) \tag{A.1}$$

, where the bandwidth h is set at the largest value, such that $K_h(\max_{l=1\ldots k} \|\boldsymbol{S}_l^{(c)} - \boldsymbol{S}_{\text{obs}}^{(c)}\|) = 0$ to ensure non-zero importance weight for k samples (Sisson et al., 2019) and mean integrated square error consistency (Biau et al., 2015). Regression adjustment seeks to produce adjusted samples c_l but maintain the sample weights and thus assumes the following model for the unadjusted samples c^{unadj} with mean-zero i.i.d errors ϵ_l where $E(\epsilon_l^2) < \infty$ for $l = 1 \ldots, k$:

$$c_l^{\text{unadj}} = m(\boldsymbol{S}_l^{(c)}) + \epsilon_l \,. \tag{A.2}$$

The estimated regression function $\widehat{m}$ is then a kernel-based local-linear polynomial obtained as a solution of $\text{argmin}_{\alpha,\beta} \sum_{l=1}^{k} [c_l^{\text{unadj}} - (\alpha + \beta(\boldsymbol{S}_l^{(c)} - \boldsymbol{S}_{\text{obs}}^{(c)}))]^2 w_l^{(c)}$. Using the empirical residuals $\widehat{\epsilon}_l = c_l^{\text{unadj}} - \widehat{m}(\boldsymbol{S}_l^{(c)})$, we then construct the adjusted values $c_l = \widehat{m}(\boldsymbol{S}_{\text{obs}}^{(c)}) + \widehat{\epsilon}_l$.

Posterior summary of Euclidean parameters (c, σ^2)**.** The first stage of our ABC-MH algorithm produces weighted samples $\{c_\ell, w_l^{(c)}\}$, $\{\sigma_\ell^2, w_l^{(\sigma^2)}\}$, $l = 1, \ldots, k$,

and we summarize the weighted samples as follows. We illustrate the calculations with c, and the calculations for σ^2 follow similarly. We calculate the posterior median and 95% credible interval by finding the 50, 2.5 and 97.5% quantiles, and use the posterior median for the second stage of the proposed ABC-MH algorithm when sampling the tree. In general, for calculating the $q \times 100\%$ quantile, we fit an intercept-only quantile regression of c_ℓ with weights $w_l^{(c)}$; this is implemented by `rq` wrapped in the summary function `summary.abc` in the R package `abc`.

A.2.2 MH Algorithm for Updating the Tree in the DDT Model.

In the second stage of Algorithm 2, we have used existing MH tree updates (Knowles and Ghahramani, 2015). We briefly describe the proposal for generating a candidate tree $\mathcal{T}'$ from the current tree $\mathcal{T}$ and the acceptance probability. Given the current tree, a candidate tree is proposed in two steps: (i) detaching a subtree from the original tree, and (ii) reattaching the subtree back to the remaining tree (see Figure A.3). In Step i, let $(\mathcal{S}, \mathcal{R})$ be the output of the random detach function that divides the original tree $\mathcal{T}$ into two parts at the detaching point u, where $\mathcal{S}$ is the detached subtree and $\mathcal{R}$ is the remaining tree. In this paper, we generate the detaching point u by uniformly selecting a node and taking the parent of the node as the detaching point. In Step ii, for the re-attaching point v, we follow the divergence and branching behaviors of the generative DDT model by treating subtree $\mathcal{S}$ as a single datum and adding a new datum $\mathcal{S}$ to $\mathcal{R}$. Given the point v, a candidate tree $\mathcal{T}'$ results by re-attaching $\mathcal{S}$ back to $\mathcal{R}$ at point v. The time of re-attaching point t_v is then earlier than the time of the root of $\mathcal{S}$ to avoid distortion of $\mathcal{S}$: $t_v < t(\text{root}(\mathcal{S}))$. By choosing u and v as above, we have described the proposal distribution from $\mathcal{T}$ to $\mathcal{T}'$, $q(v, \mathcal{R})$, which is essentially the probability of diverging at v on the subtree $\mathcal{R}$.

The acceptance probability is then

$$\min\left\{1, \frac{f(\mathcal{T}', \mathbf{X})q(u, \mathcal{R})}{f(\mathcal{T}, \mathbf{X})q(v, \mathcal{R})}\right\} \tag{A.3}$$

, where $f(\mathcal{T}, \mathbf{X}) = f(\mathcal{T}, \mathbf{X}|c_0, \sigma_0^2) = P(\mathbf{X}|\mathcal{T}, \sigma_0^2)P(\mathcal{T}|c_0)$, $P(\mathbf{X}|\mathcal{T}, \sigma_0^2)$ is the likelihood of the tree structure (Proposition 1), $P(\mathcal{T}|c_0)$ is the prior for the tree (the first two terms in Equation (4)), and c_0 and σ_0^2 are representative value chosen from the posterior sample of c and σ^2, respectively.

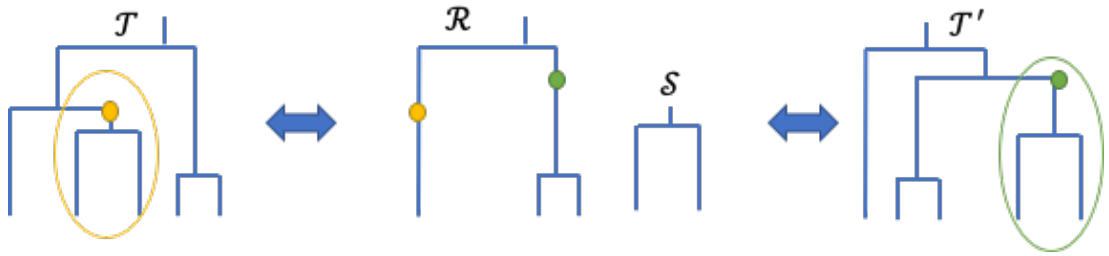


Figure A.3: Schematic diagram of proposing a candidate tree in MH. (Left) Current tree $\mathcal{T}$ with detach point u (yellow); (Middle) Intermediate subtrees with remaining tree $\mathcal{R}$ and the detached subtree $\mathcal{S}$; (Right) The proposed tree $\mathcal{T}'$ with reattached point v (green).

A.3 Tree Projection of Pairwise iPCP Matrix

In the Main Paper Section 3.2, we mentioned that a pairwise iPCP matrix $\mathbf{\Sigma}$ with entries $\text{iPCP}_{i,i'}, i, i' = 1, \ldots, I$ need not to be a tree-structured matrix and we address the projection of Σ on to the space of tree-structured matrices here. Given $L > 1$ posterior trees with I leaves and the corresponding pairwise iPCP matrix $\mathbf{\Sigma} = \left(\text{iPCP}_{i,i'}\right)$, each entry of iPCP matrix can be express as $\text{iPCP}_{i,i'} = \frac{\sum_{l=1}^{L} t_{i,i'}^{(l)}}{L}$, where $t_{i,i'}^{(l)}$ is the divergence time of leaves i and i' in the l-th posterior tree. Obviously, every entry of the iPCP matrix takes the element-wise Monte Carlo average over L tree-structured matrix and breaks the inequalities (2) and (3) in the Main Paper. Following the work of Bravo et al. (2009), by representing a tree as a tree-structured matrix, we can project Σ on to the closest tree-structured matrix in terms of Frobenius

Algorithm 2 Two-stage hybrid ABC-MH algorithm

Input:

(a) Observed data: $\mathbf{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_I]^\mathsf{T}$ consisting of I points in $\mathbb{R}^J$;

(b) Summary statistics $\boldsymbol{S}^{(c)}, S^{(\sigma^2)}$ defined in the Main Paper Section 3.1.1;

(c) Synthetic data of size N^{syn} and threshold $d \in (0, 1)$ with $k = \lceil N^{\text{syn}} d \rceil$, the number of nearest synthetic data sets to retain;

(d) Prior for model parameters: $c \sim \mathsf{Gamma}(a_c, b_c)$, $\frac{1}{\sigma^2} \sim \mathsf{Gamma}(a_{\sigma^2}, b_{\sigma^2})$;

(e) Univariate Kernel $K_h(\cdot)$ with bandwidth $h > 0$ and compact support.

Output:

(a) Posterior samples of c and σ^2 of size $k = N^{\text{syn}} d$;

(b) posterior samples of $(\mathcal{T}, \boldsymbol{t})$.

procedure EUCLIDEAN PARAMETERS(c, σ^2)
  for $l = 1 ... N^{\text{syn}}$ **do**
    Sample Euclidean parameters from prior $c_l \sim \mathsf{Gamma}(a_c, b_c), \sigma_l^2 \sim \mathsf{Gamma}(a_{\sigma^2}, b_{\sigma^2})$;
    Simulate data $\mathbf{X}_l$ from DDT using (c_l, σ_l^2);
    Compute: $\boldsymbol{S}_l^{(c)}$ and $S_l^{(\sigma^2)}$ along with $\|\boldsymbol{S}_l^{(c)} - \boldsymbol{S}_{\text{obs}}^{(c)}\|$ and $\|S_l^{(\sigma^2)} - S_{\text{obs}}^{(\sigma^2)}\|$.
  Choose $\{(c_{l_s}, \sigma_{l_s}^2), s = 1, \ldots, k\}$ corresponding to k smallest $\|\boldsymbol{S}_l^{(c)} - \boldsymbol{S}_{\text{obs}}^{(c)}\|$ and $\|S_l^{(\sigma^2)} - S_{\text{obs}}^{(\sigma^2)}\|$
  Calculate the sample weights $w_{l_s}^{(c)} = K_h(\|\boldsymbol{S}_{l_s}^{(c)} - \boldsymbol{S}_{\text{obs}}^{(c)}\|)$ and $w_{l_s}^{(\sigma^2)} = K_{h'}(\|S_{l_s}^{(\sigma^2)} - S_{\text{obs}}^{(\sigma^2)}\|)$ based on Equation (A.1);
  Compute regression adjusted samples c_{l_s} and $\sigma_{l_s}^2$ with weights $w_{l_s}^{(c)}$ and $w_{l_s}^{(\sigma^2)}$ with the model (A.2) and calculate posterior summary c_0 and σ_0^2 plugging the adjusted c_{l_s} and $\sigma_{l_s}^2$.
procedure TREE PARAMETERS$((\mathcal{T}, \boldsymbol{t}))$
  Follow the MH algorithm in Section A.2.2 with fixed c_0 and σ_0^2 at the posterior median values and compute acceptance probabilities with Equation A.3.

norm. The projection can be formulated as a constrained mixed-integer programming (MIP) problem:

$$\underset{\boldsymbol{\Sigma}^{\mathcal{T}}}{\text{argmin}} \quad ||\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^{\mathcal{T}}||_F$$

$$\text{s.t.} \quad \Sigma^{\mathcal{T}}_{i,i'} \geq 0;\ \Sigma^{\mathcal{T}}_{i,i} \geq \Sigma^{\mathcal{T}}_{i,i'};\ \Sigma^{\mathcal{T}}_{i,i'} \geq \min(\Sigma^{\mathcal{T}}_{i,i''}, \Sigma^{\mathcal{T}}_{i',i''}),\ \text{for all } i \neq i' \neq i''.$$

We applied the projection on the pairwise iPCP matrix from the breast cancer (panel (A)), colorectal cancer (panel (B)) and melanoma (panel (C)) data of NIBR-PDXE and show the result in the Figure A.4. In Figure A.4, the MAP tree, the tree representation of projected iPCP matrix (MIP tree), the original iPCP matrix and the projected iPCP matrix are shown in from the left to the right columns, respectively. From the left two columns of the tree structures, we found that trees from the MAP and MIP show similar pattern and the MIP tree allows a non-binary tree structure. For example, three combination therapies and two PI3K inhibitors (CLR457 and BKM120) framed by a box form a tight subtree in both MAP and MIP tree, but the subtree in the MIP is non-binary. For the iPCP matrix, high element-wise correlation $\text{Cor}(\Sigma^{\mathcal{T}}_{i,i'}, \Sigma_{i,i'})$ between the original iPCP Σ and the projected iPCP $\Sigma^{\mathcal{T}}$ are presented (BRCA: 0.9987; CRC: 0.9962; CM: 0.9918).

A.4 Simulation Studies of Euclidean Parameters

In this section, we empirically compare the Euclidean parameters of c and σ^2 from ABC of the proposed two-stage algorithm and single-stage MCMC. We organize this section as follows. We first compare other candidate summary statistics of c and σ^2 for ABC in Section A.4.1. In Section A.4.2, we illustrate the superior inference performance of Euclidean parameters from ABC than single-stage MCMC through a series of simulations. Section A.4.3 offers the diagnostic statistics and the sensitivity analysis for ABC stage of the proposed two-stage algorithm and checks the convergence of c and σ^2 for the single-stage MCMC.

Simulation setup. For illustrative purposes, we fixed the observed PDX data matrix with 50 treatments ($I = 50$) and 10 PDX mice ($J = 10$) in all simulation scenarios. In addition, we let c and σ^2 take values from $\{0.3, 0.5, 0.7, 1\}$ and $\{0.5, 1\}$ respectively to mimic the PDX data with tight and well-separated clusters. For each pair of (c, σ^2),

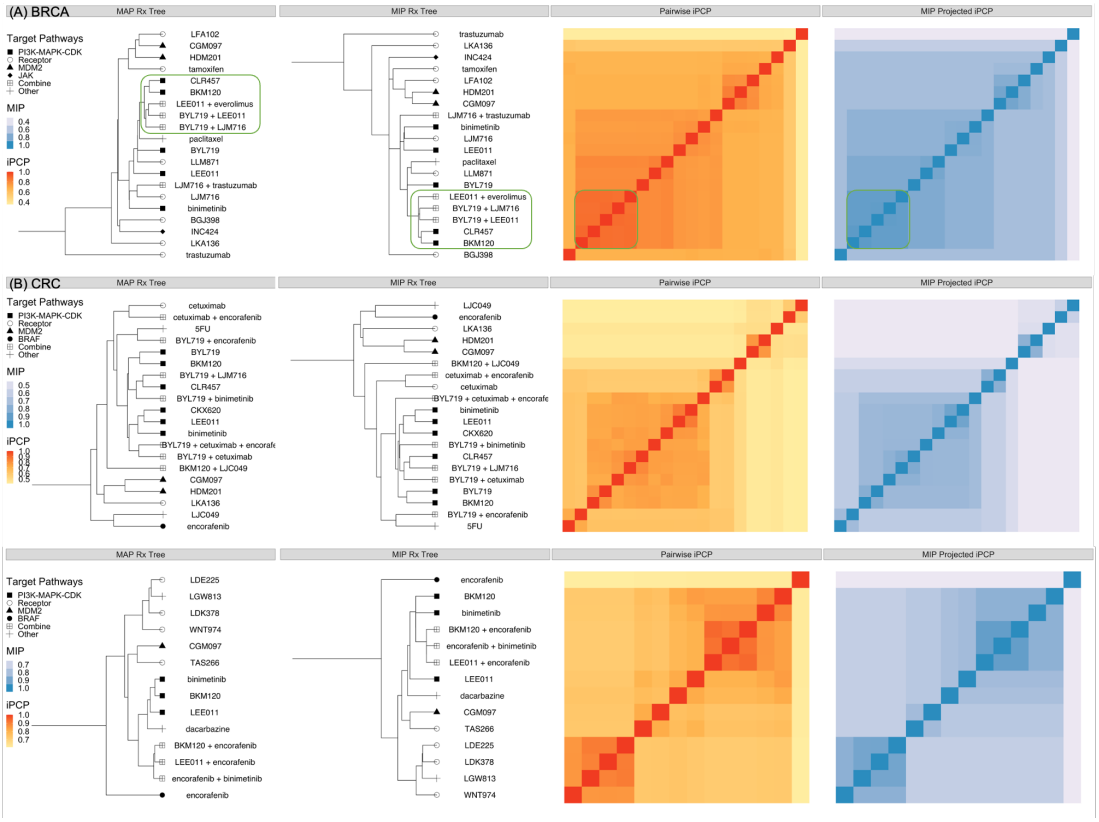


Figure A.4: Comparison between (Left two columns) the tree structure from the MAP and the projected iPCP matrix (MIP tree) and (Right two columns) the matrix from the original iPCP matrix and the projected iPCP matrix for (A) breast cancer, (B) colorectal cancer and (C) melanoma. The matrix from the original iPCP and the MIP projected iPCP matrix are aligned by the MIP tree.

200 replicated experiments with different tree and observed PDX data matrices were independently drawn according to the DDT generating model. We specify a prior distribution for $c \sim \mathsf{Gamma}(2, 2)$ with shape and rate parameterization. For diffusion variance σ^2, let $1/\sigma^2 \sim \mathsf{Gamma}(1, 1)$. We compare ABC-MH of the proposed two-stage algorithm against two alternatives based on single-stage MH algorithms (Neal, 2003) (see details in Section A.2.2). The first one initializes at the true parameter values and the true tree, referred to as $\mathrm{MH}_{\mathsf{true}}$. The idealistic initialization at the truth is a best case scenario in applying existing MH algorithm to inferring DDT models. The second alternative, referred to as $\mathrm{MH}_{\mathsf{default}}$, initializes (c, σ^2) by a random draw from the prior; the unknown tree is initialized by agglomerative hierarchical clustering

with Euclidean distance and squared Ward's linkage (Murtagh and Legendre, 2014) – thus providing a fair apples-to-apples comparison. For the ABC, we generated N^{syn} synthetic data of c and σ^2 and kept $k = \lceil N^{\text{syn}} d \rceil$ nearest samples in terms of the $\|\boldsymbol{S}_l^{(c)} - \boldsymbol{S}_{\text{obs}}^{(c)}\|$ and $\|S_l^{(\sigma^2)} - S_{\text{obs}}^{(\sigma^2)}\|$. We varied the number of synthetic data N^{syn} and the threshold parameter $d \in (0, 1)$ under different settings and we specified N^{syn} and d in each of the following sections. We ran two MH algorithms with 10,000 iterations and discarded the first 7,000 iterations.

Performance metrics for Euclidean parameters. We used two algorithm performance metrics to compare our algorithm to the classical single-stage MCMC algorithms. First we computed the effective sample sizes for each Euclidean parameter c and σ^2 (ESS_c and ESS_{σ^2}) given a nominal sample size (NSS) kept for posterior inference. ESS for each parameter represents the number of independent draws equivalent to NSS posterior draws of correlated ($\text{MH}_{\mathsf{true}}$ and $\text{MH}_{\mathsf{default}}$) or independent and unequally weighted samples (ABC stage of the proposed algorithm). We let NSS for MH algorithms be the number of consecutive posterior samples in a single chain after a burn-in period; let NSS for ABC be k as in Step 6, Algorithm 2. For c and σ^2, the ESS of MH (Gelman et al., 2013) is estimated by $\text{NSS}/(1+\sum_{t=1}^{\infty} \widehat{\rho}_t)$ where $\widehat{\rho}_t$ is the estimated autocorrelation function with lag t (Geyer, 2011). The ESS for ABC (Sisson et al., 2019) is the reciprocal of the sum of squared normalized weights, $1/\sum_{l=1}^{k} \widetilde{W}_l^2$, where $\widetilde{W}_l = w_l / \sum_{l'=1}^{k} w_{l'}$ (see weights, w_l, in Equation (A.1)). Second, we evaluated how well did the posterior distributions recover the true (c, σ^2). We computed the mean absolute percent bias for c and σ^2: $|\mathbb{E}\{c \mid \mathbf{X}\} - c|/c$ and $|\mathbb{E}\{\sigma^2 \mid \mathbf{X}\} - \sigma^2|/\sigma^2$, respectively. We also computed the empirical coverage rates of the nominal 95% credible intervals (CrI) for c and σ^2.

A.4.1 Other Choices of Summary Statistics

Proposition 1 points towards other potential summary statistics for the first stage of Algorithm 2 that uses ABC to produce weighted samples to approximate the posterior distributions for c and σ^2. Here we consider a few such alternatives with $N^{\text{syn}} = 600,000$ and $d = 0.5\%$ and empirically compare their performances to the summary statistics used in the Main Paper ($\boldsymbol{S}^{(c)}$ and $S^{(\sigma^2)}$) in terms of the mean absolute percent bias in recovering the true parameter values of c and σ^2.

Summary statistic for c. Unlike building $\boldsymbol{S}^{(c)}$ based on the inter-point distance, the off-diagonal terms of $\boldsymbol{T} = \sum_j \boldsymbol{X}_{\cdot,j}\boldsymbol{X}_{\cdot,j}^{\mathsf{T}}$ (see the definition of $\boldsymbol{T}$ in Lemma 1 in Main Paper) is another potential summary statistic for c. Since the divergence parameter c affects the marginal likelihood implicitly through the divergence time $\boldsymbol{t}$, the summary statistics for $\boldsymbol{t}$ is informative for c. From Proposition 1, $\boldsymbol{T}$ is sufficient for $\sigma^2\Sigma_{\mathcal{T}}$, where the off-diagonal terms of $\sigma^2\Sigma_{\mathcal{T}}$ taking the form $\sigma^2 t_d, d = 1 \dots n-1$ and containing unrelated information from σ^2. Let $\boldsymbol{Q_T}$ be a vector of the 10th, 25th, 50th, 75th and 90th percentiles of the off-diagonal terms of $\boldsymbol{T}$. Because $\boldsymbol{T}$ is sufficient for $\sigma^2\Sigma_{\mathcal{T}}$ and involves extra Gaussian diffusion variance parameter, we can design alternative summary statistics based on $\boldsymbol{Q_T}$ through (i) augmentation, $(\boldsymbol{Q_T}, S^{(\sigma^2)})$ or (ii) scaling, $\boldsymbol{Q_T}/S^{(\sigma^2)}$. From Figure A.5, $\boldsymbol{S}^{(c)}$ proposed in the Main Paper outperformed the summary statistics from $\boldsymbol{Q_T}$ by producing less biased posterior mean estimates.

Summary statistic for σ^2. Following Proposition 1, several matrix functionals on the data $\mathbf{X}$ or statistics $\boldsymbol{T}$ can be considered as alternatives to $S^{(\sigma^2)}$. We compare performance of three candidates: (i) average L_1 norm (AvgL1) of columns: $\frac{1}{J}\sum_{j=1}^{J}|\boldsymbol{X}_{\cdot,j}|_1$; (ii) Frobenius norm of $\mathbf{X}$; and, (iii) vector containing 10th, 25th, 50th, 75th and 90th percentiles of first principal component (PC1) of $\mathbf{X}$. From Figure A.6, the first three methods are comparable while ABC based on principal components shows larger bias due to the information loss.

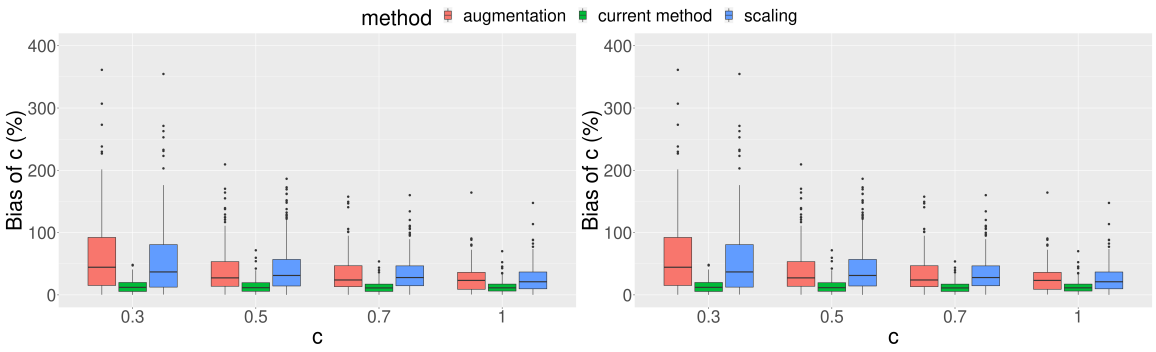


Figure A.5: Comparison among different summary statistics for c (red: $(\boldsymbol{Q_T}, S^{(\sigma^2)})$; green: $\boldsymbol{S}^{(c)}$; blue: $\boldsymbol{Q_T}/S^{(\sigma^2)}$) under different values of σ^2 in terms of the mean absolute percent bias. (Left) $\sigma^2 = 0.5$; (Right) $\sigma^2 = 1$.

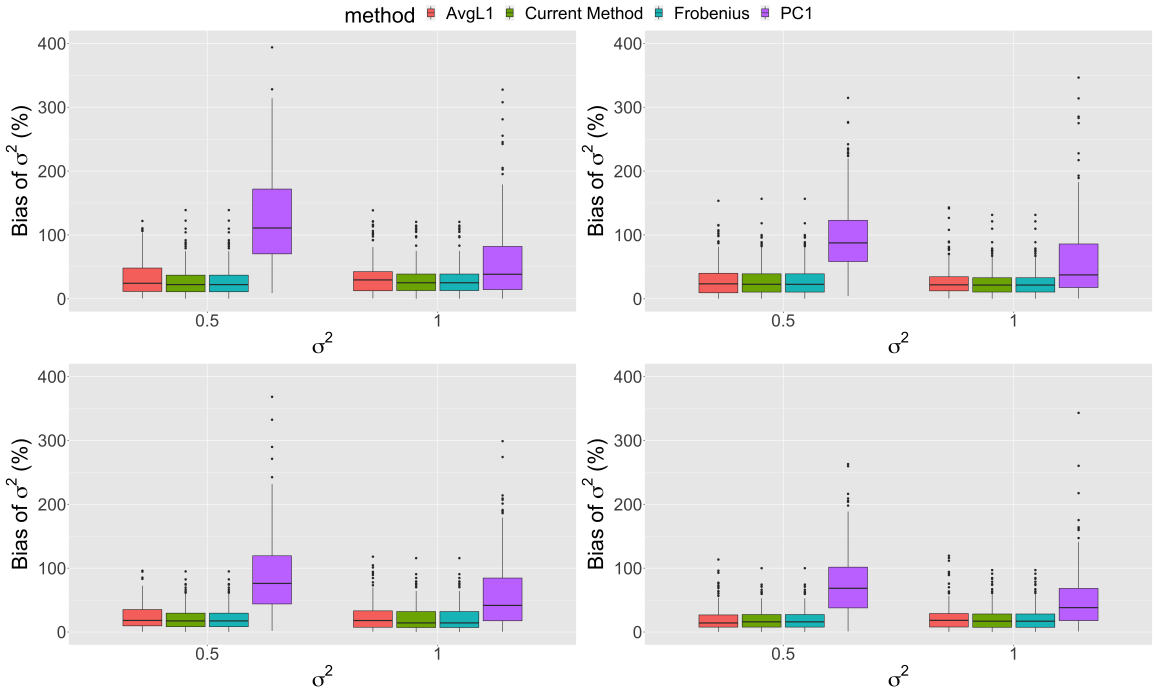


Figure A.6: Comparison among different summary statistics for σ^2 under different values of c in terms of the mean absolute percent bias. (Upper Left) $c = 0.3$; (Upper Right) $c = 0.5$; (Lower Left) $c = 0.7$; (Upper Right) $c = 1.0$.

A.4.2 Posterior Inference of Euclidean Parameters

In this section, we show that two-stage algorithm (ABC-MH) outperforms the single-stage MCMC (MH) for real parameters in terms of (i) stable effective sample size (ESS) for (c, σ^2); (ii) similar or better inference on (c, σ^2), as ascertained using mean absolute percent bias and nominal 95% credible intervals.

A.4.2.1 Stable Effective Sample Sizes of ABC-MH

We calculated ESS-to-NSS ratios at varying truths of c and σ^2. To illustrate, we matched the NSS budget of ABC with that of MH (NSS $= 3,000$) by keeping $d = 0.5\%$ of $N^{\text{syn}} = 600,000$ synthetic data sets that are closest to the observed data in terms of the summary statistic for each parameter (Step 6 of Algorithm 2). Table A.1 shows that the ESS_c/NSS and ESS_{σ^2}/NSS ratio from ABC is stable between 0.64 to 0.68 and around 0.83 across different c and σ^2 values, respectively. In contrast, the ESS_c/NSS ratio for MH quickly deteriorates ($\text{MH}_{\mathsf{true}}$: 0.97 to 0.41; $\text{MH}_{\mathsf{default}}$: 0.73 to 0.35) as c increases from 0.3 to 1 and ESS_{σ^2}/NSS for MH are extremely poor (< 0.06) across different values of c and σ^2. MH produced very good ESS_c under small value $c = 0.3$ but poor ESS_c under $c = 1$. As a result, under larger values of c, MH algorithms must run longer to reach a target ESS_c. Although ESS_c for ABC is not as high as $\text{MH}_{\mathsf{true}}$ or $\text{MH}_{\mathsf{default}}$ at $c = 0.3$, the stability of ESS_c of ABC means that a predictably constant NSS is needed for conducting posterior inference across different values of c. Finally, the ESS_{σ^2} for the diffusion variance parameter from MH algorithms are strikingly smaller than ABC, indicating ABC should be preferred.

A.4.2.2 Superior Quality Posterior Inference of ABC-MH

Does ABC give better posterior inference with a fixed computational budget? To make fair comparisons, we fixed a total CPU time and used the same computing processor to run the ABC (1st stage of Algorithm 1) and MH algorithms. Let t_{MH} and t_{ABC} be the estimated CPU time for generating one iteration in MH and one synthetic data in ABC on the same processor. Note, t_{MH} includes the additional time for proposing a valid tree. By varying the number of synthetic samples, we can match the total CPU time used by ABC with that of MH algorithms which were run for $10,000$ iterations. We generated $10,000 t_{\mathsf{MH}}/t_{\mathsf{ABC}} = 17,345$ synthetic data sets and took $d = 5\%$ with summary statistics $\boldsymbol{S}^{(c)}$ and $S^{(\sigma^2)}$ (see different values of d in Section

Table A.1: ESS-to-NSS ratios between ABC-MH ($d = 0.5\%$), $\text{MH}_{\textsf{true}}$, and $\text{MH}_{\textsf{default}}$. All values here are obtained from 200 independent replications. For each random replication at (c, σ^2). All methods were controlled to produce identical NSS with size $3,000$.

		ESS/NSS(sd) for c		ESS/NSS(sd) for σ^2	
c	method	$\sigma^2 = 0.5$	$\sigma^2 = 1$	$\sigma^2 = 0.5$	$\sigma^2 = 1$
0.3	ABC-MH	0.68(0.032)	0.67(0.027)	0.83(0.0048)	0.83(0.0042)
	$\text{MH}_{\textsf{true}}$	0.97(0.11)	0.96(0.13)	0.051(0.061)	0.056(0.072)
	$\text{MH}_{\textsf{default}}$	0.73(0.33)	0.67(0.34)	0.028(0.043)	0.038(0.08)
0.5	ABC-MH	0.66(0.02)	0.65(0.018)	0.83(0.0047)	0.83(0.0044)
	$\text{MH}_{\textsf{true}}$	0.85(0.23)	0.83(0.24)	0.034(0.042)	0.045(0.067)
	$\text{MH}_{\textsf{default}}$	0.66(0.35)	0.62(0.34)	0.033(0.051)	0.041(0.067)
0.7	ABC-MH	0.65(0.017)	0.64(0.017)	0.83(0.0047)	0.83(0.004)
	$\text{MH}_{\textsf{true}}$	0.63(0.31)	0.67(0.32)	0.024(0.027)	0.029(0.038)
	$\text{MH}_{\textsf{default}}$	0.53(0.33)	0.51(0.35)	0.028(0.039)	0.038(0.072)
1.0	ABC-MH	0.65(0.017)	0.64(0.017)	0.83(0.0044)	0.83(0.0041)
	$\text{MH}_{\textsf{true}}$	0.41(0.3)	0.44(0.32)	0.019(0.026)	0.019(0.023)
	$\text{MH}_{\textsf{default}}$	0.35(0.29)	0.35(0.29)	0.022(0.026)	0.022(0.027)

A.4.3.3) for ABC. Table A.2 shows that ABC produced posterior samples that confer comparable inferences about c in terms of the bias and coverage of nominal 95% CrIs. The posterior mean of c from ABC is comparable to that from $\text{MH}_{\textsf{true}}$ and less biased than $\text{MH}_{\textsf{default}}$ for all settings. The coverage rates of the nominal 95% CrIs from ABC are comparable to $\text{MH}_{\textsf{true}}$ but higher than $\text{MH}_{\textsf{default}}$. $\text{MH}_{\textsf{true}}$, however, is initialized at true values and is unrealistic in practice. We observed $\text{MH}_{\textsf{true}}$ sometimes failed to converge (Table A.3), stuck around the initial true values and resulted in deceptively low biases and good coverage rates. Turning to the inference of σ^2, ABC offers a much better alternative to MH algorithms in terms of smaller bias in the posterior mean and better coverage of the 95% credible intervals (Table A.2). This is primarily caused by the difficulty of MH in exploring the posterior distribution of σ^2 resulting in chains with high auto-correlations. The squeezed boxplots in Figure A.7 indicate that the chains for σ^2 in $\text{MH}_{\textsf{true}}$ and $\text{MH}_{\textsf{default}}$ were almost always slowly mixing and stuck around the initial values. In addition, unlike the serial nature of MH, ABC can

be further parallelized to reduce the wall clock time to a fraction of what is required by MH using multicore processors. Although parallelizing MH with techniques such as consensus MCMC (e.g., Scott et al., 2016) is possible, the parallelized ABC does not require data splitting and will not trade the quality of posterior inference for computational speed.

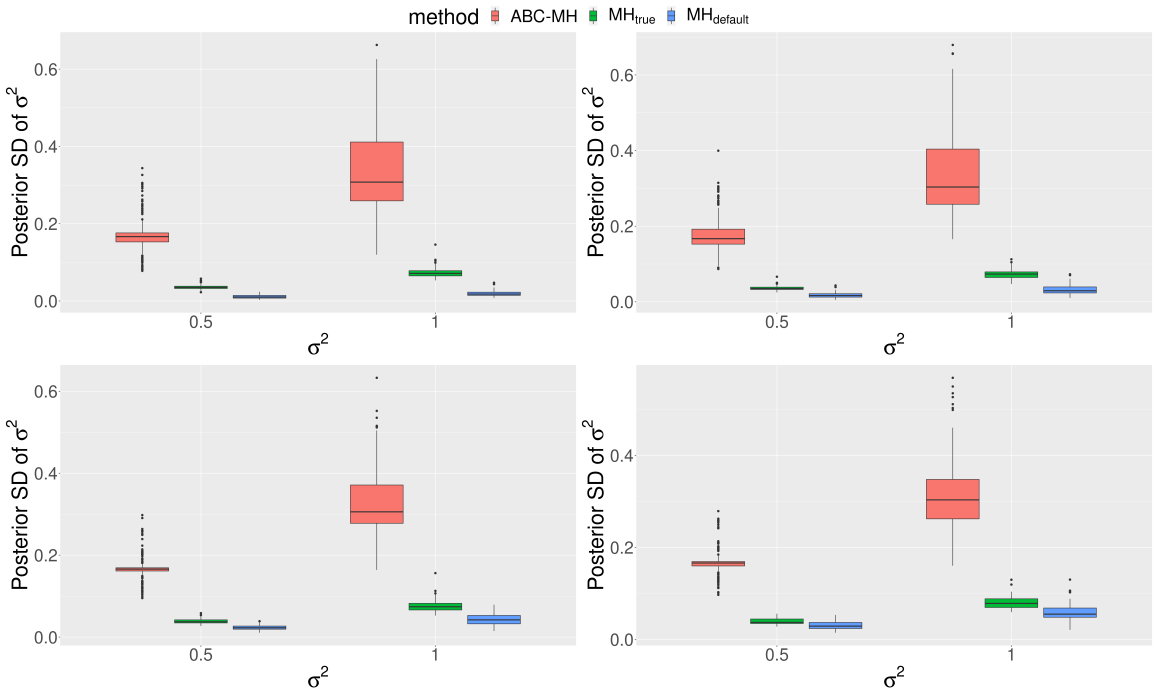


Figure A.7: (Upper left) $c = 0.3$; (Upper right) $c = 0.5$; (Lower left) $c = 0.7$; (Lower right) $c = 1.0$. The posterior standard deviation of σ^2 from MH (green and blue) are close to zero across different true c showing MH is stuck. Results are based on 200 replications.

A.4.3 Algorithm Diagnostics

Here we examine the convergence of MH through the Geweke statistics (Geweke, 1992) and the goodness of fit for ABC. Specifically, two important hyper-parameters are involved in ABC: (i) the kernel bandwidth h for samples weights in Equation A.1 and (ii) the threshold d for $k = \lceil N^{\text{syn}} d \rceil$ nearest samples in the Step 6 of Algorithm 2. We follow the test from Prangle et al. (2014) to justify the kernel bandwidth h and conduct the sensitivity analysis for threshold d to understand how threshold d

Table A.2: Comparison of inferential performance for c and σ^2 between ABC-MH ($d = 5\%$), MH_{true}, and $\text{MH}_{\text{default}}$. All values here are obtained from 200 independent replications. For each random replication at (c, σ^2), all methods were run for identical total CPU time and only converged chains from MH algorithms were included.

		Percent Bias(sd) for c		Coverage(sd) for c		Percent Bias(sd) for σ^2		Coverage(sd) for σ^2	
c	method	$\sigma^2 = 0.5$	$\sigma^2 = 1$	$\sigma^2 = 0.5$	$\sigma^2 = 1$	$\sigma^2 = 0.5$	$\sigma^2 = 1$	$\sigma^2 = 0.5$	$\sigma^2 = 1$
	ABC-MH	12(9.4)	13(9.9)	98(0.99)	99(0.71)	28(25)	31(25)	90(2.1)	88(2.3)
0.3	MH_{true}	13(9.8)	12(9.5)	94(2)	95(1.9)	9.4(7.1)	9(6.6)	80(3.4)	82(3.3)
	$\text{MH}_{\text{default}}$	45(20)	46(20)	33(5.5)	30(6.1)	71(12)	72(11)	0(0)	0(0)
	ABC-MH	15(11)	15(11)	92(1.9)	93(1.8)	28(26)	27(22)	90(2.2)	94(1.7)
0.5	MH_{true}	11(9)	11(8.6)	97(1.7)	97(1.6)	8.6(6.7)	9.9(7)	80(4)	78(4.1)
	$\text{MH}_{\text{default}}$	33(18)	31(19)	60(5.5)	57(5.7)	54(17)	57(16)	1.2(1.2)	1.3(1.3)
	ABC-MH	13(10)	14(11)	96(1.5)	93(1.8)	21(18)	22(20)	97(1.2)	94(1.6)
0.7	MH_{true}	12(9.1)	12(9.1)	95(2.6)	96(2.1)	11(8.3)	11(8.3)	70(5.3)	68(4.9)
	$\text{MH}_{\text{default}}$	25(15)	27(16)	73(5.5)	69(5.9)	38(17)	41(19)	12(4)	8.1(3.5)
	ABC-MH	14(11)	14(13)	95(1.5)	94(1.6)	19(18)	21(18)	98(1.1)	96(1.5)
1.0	MH_{true}	11(7.6)	13(11)	97(2)	92(3.5)	13(9.1)	10(7.7)	64(5.8)	85(4.6)
	$\text{MH}_{\text{default}}$	14(11)	16(14)	93(3.5)	89(3.8)	24(15)	27(16)	35(6.5)	21(5.1)

affects the result in terms of the inferential performance.

A.4.3.1 Convergence of MH Chains in Simulations

In all of our simulations, we ran MH for $10,000$ iterations. Table A.3 shows that the percentages of the converged MH chains for 200 replications are between 12.5 and 68.5% within a total $10,000$ iterations (based on Geweke statistic). Running the chains longer will increase these percentages. In contrast, with appropriate choice of bandwidth and the fraction of synthetic samples to keep, ABC does not involve convergence issues and according to Section A.4.2 achieves better ESS for a fixed NSS and similar or better quality posterior inference for fixed CPU time.

Table A.3: Percentage of converged chains for (i) MH initialized at true (c, σ^2) ($\text{MH}_{\mathsf{true}}$), and (ii) MH initialized randomly from prior ($\text{MH}_{\mathsf{default}}$). All values here are obtained from 200 independent replications.

		Convergence % for c		Convergence % for σ^2	
c	method	$\sigma^2 = 0.5$	$\sigma^2 = 1$	$\sigma^2 = 0.5$	$\sigma^2 = 1$
0.3	$\text{MH}_{\mathsf{true}}$	68.0	68.5	16.5	22.5
	$\text{MH}_{\mathsf{default}}$	36.5	28.5	12.5	16.0
0.5	$\text{MH}_{\mathsf{true}}$	50.0	52.0	23.0	29.5
	$\text{MH}_{\mathsf{default}}$	40.5	37.5	18.5	23.5
0.7	$\text{MH}_{\mathsf{true}}$	38.0	46.0	26.5	27.0
	$\text{MH}_{\mathsf{default}}$	33.5	31.0	14.5	25.0
1.0	$\text{MH}_{\mathsf{true}}$	35.0	30.5	20.5	30.5
	$\text{MH}_{\mathsf{default}}$	27.5	33.0	14.0	25.0

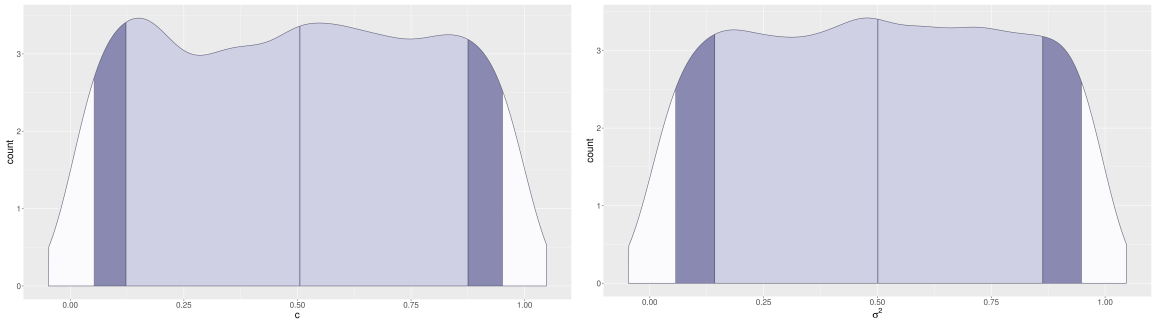
A.4.3.2 Diagnostics for ABC

We empirically justify the choice of the kernel bandwidth h and the goodness of approximation in ABC algorithm by the calibration method from Prangle et al. (2014) based on the coverage property of the credible interval. Suppose we generated pseudo-observed data $\mathbf{X}_e$ in the eth replication from the DDT model with parameter

(c_e, σ_e^2), where c_e and σ_e^2 are random draws from the prior ($c_e \sim \mathsf{Gamma}(a_c, b_c), 1/\sigma_e^2 \sim \mathsf{Gamma}(a_{\sigma^2}, b_{\sigma^2})$) and $e = 1 \dots E$. Once the tuning parameters (N^{syn}, d, h) are decided, Algorithm 2 will output regression adjusted sample (c_ℓ, σ_ℓ^2) with size $\ell = 1, \dots, k; k = \lceil N^{\text{syn}} d \rceil$ based on the input data D. We describe diagnostics for c, and note that an identical description applies to σ^2 as well. According to Cook et al. (2006), the ABC procedure produces reliable approximations of the posterior if the random variables $q_e^{(c)} := \frac{1}{k} \sum_{l=1}^{k} \mathbb{I}_{\{c_\ell > c_e\}}$ follow a uniform distribution over the interval $(0, 1)$. Accordingly, Prangle et al. (2014) suggest a goodness-of-fit test $H_0 : q_e^{(c)} \sim \mathsf{Unif}(0, 1)$ as a diagnostic in order to calibrate ABC. If the test fails to reject the null hypothesis, the empirical quantiles can be viewed as being indistinguishable from the uniform distribution, and the credible interval from the posterior samples would show the asserted coverage. We use the Kolmogorov–Smirnov statistic to carry out the test, follow the simulation setting with $I = 50$ and $J = 10$, and reuse $600,000$ synthetic data sets. The synthetic data is randomly split into two non-overlapping subsets: training data with size $597,000$ and pseudo-observed data with size $E = 3,000$. Again, we run the ABC part of Algorithm 2 by treating each of the pseudo-observed data sets as the actually observed data with $N^{\text{syn}} = 597,000$ and $d = 0.5\%$. We obtained statistically non-significant KS statistics for c and σ^2 (p-values: 0.61 for c, 0.71 for σ^2). The 95% credible intervals from ABC showed 94.9% and 95.93% empirical coverage rates which are close to the nominal level.

A.4.3.3 Sensitivity Analysis of k Nearest Samples

In the previous section, we have used a simple diagnostic procedure to show the choice of bandwidth parameter h is reasonable. Here we focus on conducting additional simulations to investigate how does varying values of d in the Step 6 of Algorithm 2 impact the inferential performance of ABC. We focus on c to illustrate the main points. Similar to Table A.2 in the Section A.4.2 where $d = 5\%$, in the



(a) Empirical quantiles at true c **(b)** Empirical quantiles at true σ^2

Figure A.8: The empirical quantiles at the true value follow the standard uniform distribution indicating calibrated ABC. Results are based on 3,000 independent draws from the prior.

following we show the results for $d = 0.5\%$ and $d = 1\%$ in Table A.4. First, for ABC itself, the bias in the posterior mean is similar, e.g. the mean bias is 14% for all three different d when $c = 1.0$ and $\sigma^2 = 0.5$. For each pair of (c, σ^2), the empirical coverage rate of the 95% credible interval decreases when d increases from 0.5% to 5%. Specifically, the empirical coverage range from 92% to 99% for $d = 5\%$, 88% to 97% for $d = 1\%$ and 84% to 94% for $d = 0.5\%$. This is likely caused by a smaller sample size $k = \lceil N^{\text{syn}} d \rceil$ and a higher posterior variance under a similar level of bias.

A.4.4 Sensitivity Analysis of the Number of Synthetic Data in ABC

To our knowledge, only two packages available from: (i) Neal (2003) on the website `https://www.cs.toronto.edu/~radford/dft.software.html` and (ii) Knowles and Ghahramani (2015) on the Github `https://github.com/davidaknowles/pydt`. Neal's code is implemented on `R`, and does not implement the inference algorithm, while Knowles programmed the C++ code from scratch including the library for the tree structure. However, the C++ libraries from Knowles and Ghahramani (2015) are deprecated and require additional updates for the version updates of the C++ compiler. Without additional documentation, the C++ code is hard to adapt in our context. Thus, we implemented our algorithm in `R` based on the existing libraries for

Table A.4: Sensitivity analysis of d for ABC-MH. We compare the inferential performance for c among ABC-MH with $d = 5\%$, ABC-MH with $d = 1\%$, ABC-MH with $d = 0.5\%$, $\text{MH}_{\mathsf{true}}$, and $\text{MH}_{\mathsf{default}}$. All values here are obtained from 200 independent replications. For each random replication at (c, σ^2), all methods were run for identical total CPU time and only converged chains from MH algorithms were included.

		Percent Bias(sd)		Coverage(sd)	
c	method	$\sigma^2 = 0.5$	$\sigma^2 = 1$	$\sigma^2 = 0.5$	$\sigma^2 = 1$
0.3	ABC-MH with $d = 5\%$	12(9.4)	13(9.9)	98(0.99)	99(0.71)
	ABC-MH with $d = 1\%$	13(9.8)	14(10)	97(1.2)	96(1.5)
	ABC-MH with $d = 0.5\%$	14(10)	15(11)	94(1.6)	92(1.9)
	$\text{MH}_{\mathsf{true}}$	13(9.8)	12(9.5)	94(2)	95(1.9)
	$\text{MH}_{\mathsf{default}}$	45(20)	46(20)	33(5.5)	30(6.1)
0.5	ABC-MH with $d = 5\%$	15(11)	15(11)	92(1.9)	93(1.8)
	ABC-MH with $d = 1\%$	15(12)	16(12)	88(2.3)	90(2.1)
	ABC-MH with $d = 0.5\%$	16(12)	16(12)	84(2.6)	86(2.4)
	$\text{MH}_{\mathsf{true}}$	11(9)	11(8.6)	97(1.7)	97(1.6)
	$\text{MH}_{\mathsf{default}}$	33(18)	31(19)	60(5.5)	57(5.7)
0.7	ABC-MH with $d = 5\%$	13(10)	14(11)	96(1.5)	93(1.8)
	ABC-MH with $d = 1\%$	13(10)	13(11)	94(1.7)	90(2.1)
	ABC-MH with $d = 0.5\%$	13(11)	14(11)	90(2.1)	89(2.2)
	$\text{MH}_{\mathsf{true}}$	12(9.1)	12(9.1)	95(2.6)	96(2.1)
	$\text{MH}_{\mathsf{default}}$	25(15)	27(16)	73(5.5)	69(5.9)
1.0	ABC-MH with $d = 5\%$	14(11)	14(13)	95(1.5)	94(1.6)
	ABC-MH with $d = 1\%$	14(10)	14(13)	88(2.3)	92(1.9)
	ABC-MH with $d = 0.5\%$	14(11)	15(13)	86(2.4)	86(2.4)
	$\text{MH}_{\mathsf{true}}$	11(7.6)	13(11)	97(2)	92(3.5)
	$\text{MH}_{\mathsf{default}}$	14(11)	16(14)	93(3.5)	89(3.8)

tree structure (e.g. `ape` and `phylobase`) and the ABC algorithm (e.g. `ABC`).

The main computation bottleneck for our algorithm on `R` is the ABC stage (141 hours for 600,000 synthetic data), which is much slower than the MH stage (1.7 hours for 10,000 iterations) and the single stage MCMC (2.5 hours for 10,000 iterations). However, the ABC can be easily parallelized to reduce the wall-clock time given a sufficient number of CPU cores. In addition, we may reduce the number of synthetic data (N^{Syn}) in ABC to further improve speed. We have now conducted a simulation study to empirically demonstrate the acceleration of the ABC through the reduction

of N^{Syn}. Specifically, we ran the ABC and measured the posterior median under a lower N^{Syn}.

We show the simulation results in Table A.5. From Table A.5, the $\widehat{\sigma}^2$ are relatively stable in terms of the mean and standard deviation under a lower N^{Syn}. On the other hand, the standard deviation of $\widehat{c}$ grows rapidly ($sd : 0.0280$ for $N^{\text{Syn}} = 600,000$ and $sd : 0.260$ for $N^{\text{Syn}} = 5,000$) when the N^{Syn} decreases. For our main analyses, we went with the conservative choice of $N^{\text{Syn}} = 600,000$ for the confirmatory results.

N^{Syn}	Total CPU Hour	$\widehat{c}$ (sd)	$\widehat{\sigma}^2$ (sd)
600,000	141	1.18 (0.0280)	1.87 (0.245)
300,000	70.5	1.18 (0.0278)	1.87 (0.246)
100,000	23.5	1.16 (0.0429)	1.87 (0.246)
50,000	11.8	1.18 (0.0707)	1.86 (0.235)
10,000	2.35	1.17 (0.159)	1.88 (0.240)
5,000	1.18	1.25 (0.260)	1.84 (0.249)

Table A.5: The total CPU time and the median of the real parameters (mean and the standard deviation in the bracket) under different numbers of synthetic data (N^{Syn}) for the ABC stage. All values are obtained from 30 independent replicates from the correct specified data generating mechanism. The underlying true $c = 1.220$ and $\sigma^2 = 1.755$.

A.5 Additional Simulation Results of $\mathbf{R_x}$-Trees

In this Section, we provide more simulation results for the Section 4.2 in the Main Paper. We empirically compare the the proposed two-stage ABC-MH with the single-stage MCMC in terms of the MAP tree estimation (Section A.5.1) and recovery of pairwise treatment similarities (Section A.5.2).

Simulation setup. For the following simulations, we followed the same setup as in Section A.4 with $I = 50$ and $J = 10$, and let c and σ^2 take values from $\{0.3, 0.5, 0.7, 1.0\}$ and $\{0.5, 1.0\}$, respectively. For each pair of (c, σ^2), 50 pairs of tree and data on the leaves were independently drawn based on the DDT model. For ABC, we generated $N^{\text{syn}} = 600,000$ synthetic data sets from the DDT model with threshold

parameter $d = 0.5\%$. We assigned priors on $c \sim \mathsf{Gamma}(2, 2)$ and $1/\sigma^2 \sim \mathsf{Gamma}(1, 1)$ with shape and rate parameterization. We compare the proposed algorithm against two alternatives based on MH algorithms ($\text{MH}_{\mathsf{true}}$ and $\text{MH}_{\mathsf{default}}$. We ran MH algorithms (the 2nd stage of the proposed algorithm, $\text{MH}_{\mathsf{true}}$ and $\text{MH}_{\mathsf{default}}$) with 10,000 iterations and discarded the first 7,000 iterations.

Performance metrics. We assess the accuracy of tree estimation using Billera – Holmes– Vogtmann (BHV) distance (Billera et al., 2001) between the true tree and the *maximum a posteriori* (MAP) tree obtained from ABC-MH, $\text{MH}_{\mathsf{true}}$ and $\text{MH}_{\mathsf{default}}$, or between the true tree and the dendrogram obtained from hierarchical clustering, respectively. For the pairwise similarities, we follow the Section 4.1 and calculate iPCPs for all pairs of treatments and evaluate the iPCPs by correlation of correlation for estimated similarities and true branching time and the Frobenius norm for the overall matrix.

A.5.1 Recovery of the True Tree

The proposed two-stage algorithm decoupled the real and tree parameters, produced better inference for Euclidean parameters (See Section A.4.2), resulting in better inference for the unknown treatment tree. In particular, Figure A.9 shows that, in terms of the BHV distance, the MAP tree estimates from ABC-MH better recovers the trees than $\text{MH}_{\mathsf{default}}$ and hierarchical clustering with Euclidean distance and squared Ward linkage (Hclust). On average, MAP from $\text{MH}_{\mathsf{true}}$ is the closest to the true underlying tree. However, $\text{MH}_{\mathsf{true}}$ requires knowledge about the truth and is unrealistic in practice. In addition, we observed that the chains from $\text{MH}_{\mathsf{true}}$ in fact did not mix well and were stuck at the initial values hence falsely appearing accurate. The second stage MH for sampling the tree built on the high-quality posterior samples of c and σ^2 obtained from the 1st stage ABC and produced better MAP tree estimates that are on average closer to the simulation truths than $\text{MH}_{\mathsf{default}}$ and

Hclust.

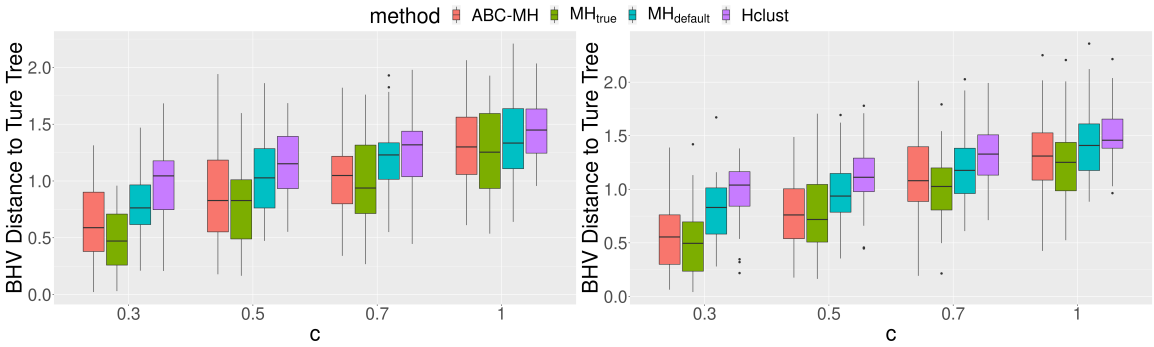


Figure A.9: (Left) $\sigma^2 = 0.5$; (Right) $\sigma^2 = 1$. The BHV distance between the MAP estimate and the underlying tree for each algorithm. Results are based on 50 replications.

A.5.2 Estimation of Treatment Similarities

The two-stage algorithm also produces better iPCPs due to decoupling strategy and superior inference for Euclidean parameters in the first stage. Similar to the results for MAP, pairwise iPCPs from ABC-MH better recover the true branching time than $\text{MH}_{\text{default}}$, Hclust and Pearson correlation and reach similar quality to the iPCPs from MH_{true} (See Figure A.10). Since MH_{true} requires unrealistic true parameters, MH_{true} is not attainable. From the simulations above, MAP and iPCPs from ABC-MH outperform $\text{MH}_{\text{default}}$ and take care of overall and local tree details, respectively. We apply the ABC-MH to obtain posterior DDT samples for the real data analysis section.

A.5.3 Computation Time of the Gaussian Likelihood Evaluation

Computationally, the complexity for the belief propagation is faster in theory, but the computation time also replies on the implementation. We empirically compare the running time of the evaluation of Gaussian likelihood on `R` for (i) the naive method of the Cholesky decomposition and (ii) the belief propagation algorithm. Specifically, we ran the `dmvnorm` function for naive method from the package `mvtnotm` and the

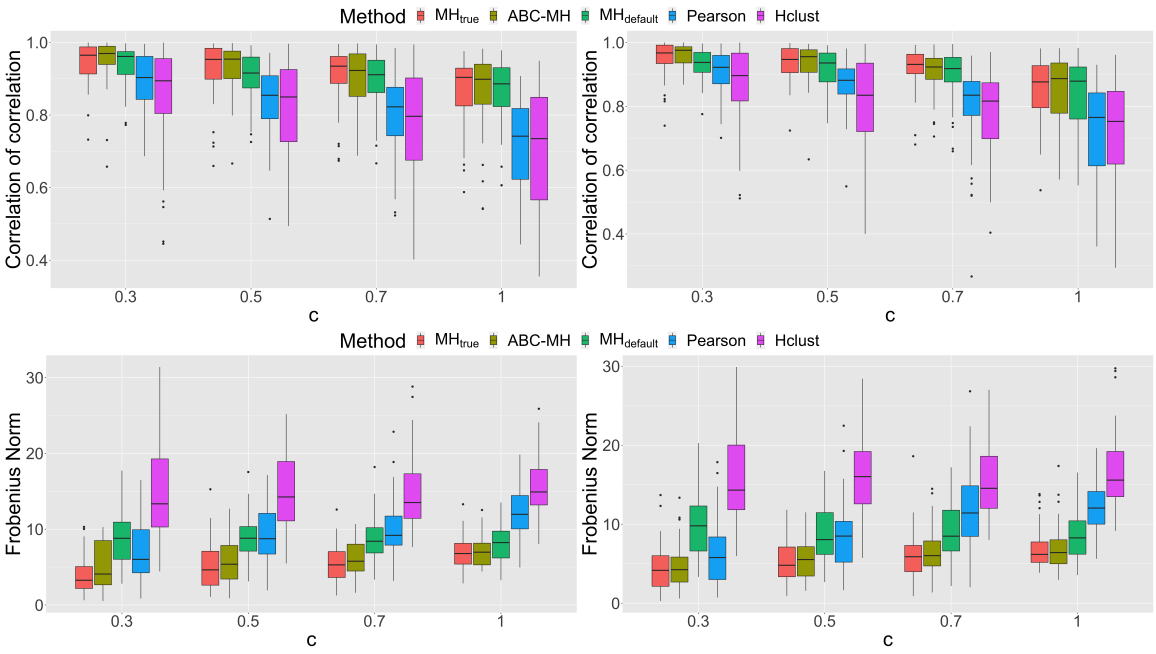


Figure A.10: Under different c and σ^2, two-stage algorithm better estimates the pairwise similarities than classical single-stage MCMC in terms of correlation of correlation (upper panels) and Frobenius norm (lower panels). (Left) $\sigma^2 = 0.5$; (Right) $\sigma^2 = 1$. Results are based on 50 replications.

`Marginals` function for belief propagation from the package `BayesNetBP` (Yu et al., 2020). To our knowledge, the package `BayesNetBP` is the only `R` package implements exact belief propagation for the Gaussian data without commercial dependencies (Yu et al., 2020). We ran each function 500 times on the Breast cancer data with the dimension of 20×38 given the same tree structure. All computation are executed on the same local computer of the Mac mini with M1 CPU and 8Gb memory. On `R`, the belief propagation (0.0566 second) is slower than the naive likelihood calculation (0.000148 second). The hindered belief propagation might be the result of the for-loop, which is slow in `R` (Burns, 2011).

A.5.4 Inference using the Whole Posterior Samples of c and σ^2

Our algorithm runs the approximate Bayesian computation (ABC) rejection algorithm (Sisson et al., 2019) to obtain the posterior samples of c and σ^2 and uses

the posterior median of c and σ^2 as the common and fixed input for different chains of the MH algorithm. The ABC merges all synthetic data into a larger dataset and re-use the same synthetic data for different chains of the MH, which is advocated by Bertorelle et al. (2010) and Blum et al. (2013). Under the ABC framework, the same synthetic data results in the identical posterior samples of c and σ^2 as the common input for different chains of the MH.

Once MH algorithm receives the posterior samples, another viable option is to use the whole posterior sample instead of using the fixed representative statistics only. We provide a set of simulations to empirically compare two algorithms using: (i) fixed posterior median only and (ii) the whole posterior samples. The algorithm (i) plugins the fixed posterior medians of c and σ^2, while the algorithm (ii) randomly picks one posterior sample at each iteration in MH. Specifically, given L weighted posterior samples of c_l and σ_l^2, with the weights w_l^c and $w_l^\sigma, l = 1 \ldots, L$, algorithm (ii) draws a posterior sample of c_l and σ_l^2 with corresponding weights at each iteration. Eventually, we measure the results through the pairwise similarity with the correlation of correlation and the Frobenius norm.

We show our simulation results in Figure A.11 using pairwise similarity. In Figure A.11, the algorithm (i) (DDT) and (ii) (DDT.all) perform similarly in terms of the correlation of correlation (mean for DDT: $(0.944, 0.971, 0.981, 0.882)$ and DDT.all: $(0.945, 0.966, 0.979, 0.877)$) and the matrix norm (mean for DDT: $(1.154, 1.415, 1.558, 1.811)$ and DDT.all: $(1.156, 1.503, 1.516, 1.814)$) under four different data generating scenarios.

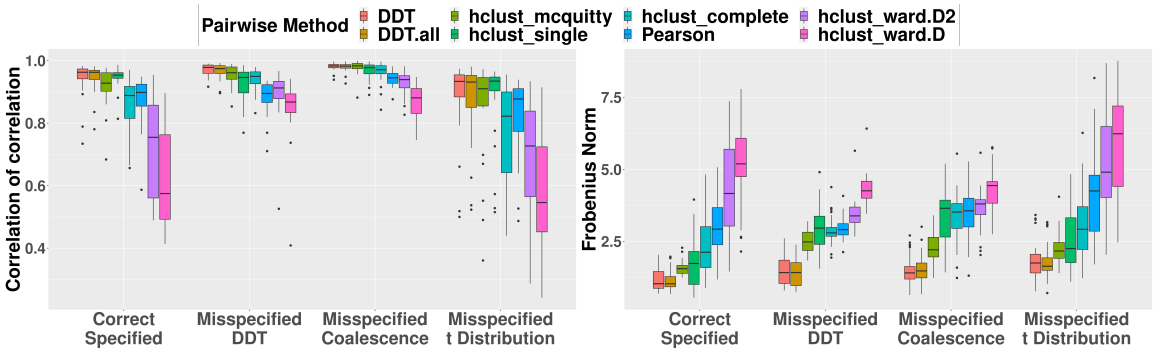


Figure A.11: Simulation studies for comparing the quality of estimated treatment similarities based on DDT (DDT: median of (c, σ^2) and DDT.all: re-sample from the whole posterior samples of (c, σ^2)), hierarchical clustering, and empirical Pearson correlation. Two performance metrics are used: (Left) Correlation of correlation (higher values are better); (Right) Matrix distances with Frobenius norm for pairwise similarity and max norm for three-way similarity (lower values are better). DDT captures true similarity best under four levels of misspecification scenarios.

A.5.5 PDX Experiment with a Smaller Dimension

We investigated the performance of proposed method on smaller scale simulated datasets. Specifically, we applied our algorithm to two datasets with smaller dimensions (treatments, patients): 5×5 and 10×15. We show the simulation results in Figure A.12 through the pairwise similarity (the correlation of correlation and the Frobenius norm). Overall, our algorithm outperforms the distance based hierarchical clustering (hclust) and the pairwise Pearson correlation in terms of the pairwise similarity except for two cases. Specifically, our algorithm is the best or the second best except for two cases: (i) the correlation of correlation under the scenario of the misspecified t-distribution with the dimension of 5×5 and (ii) the Frobenius norm under the scenario of the misspecified DDT with the dimension of 10×15. However, even under these two cases, our algorithm still have a highest lower bound in case (i) and a lowest upper bound of the Frobenius norm in case (ii), which indicates the advantage of avoiding the worst case for our algorithm. In summary, under the $1 \times 1 \times 1$ experimental design, we recommend our algorithm even under an extremely small dataset such as the dimension of 5 by 5, given enough computation resources.

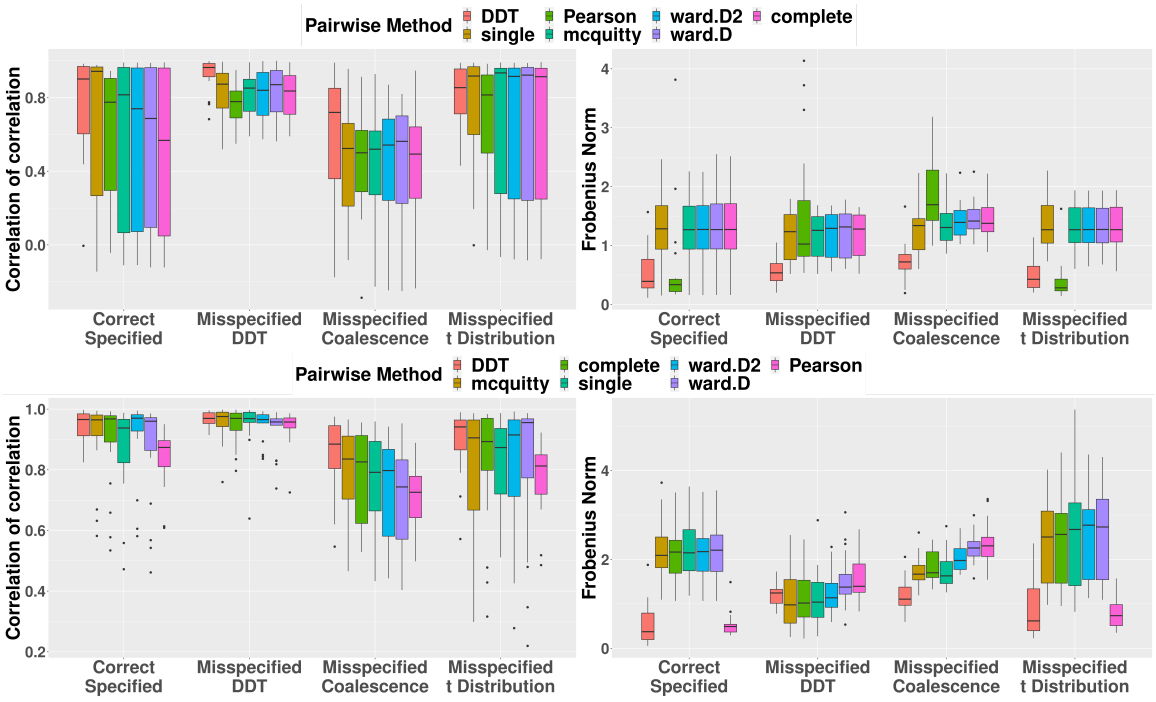


Figure A.12: The pairwise similarity for the PDX experiment with a small number of dimensions. (top): 5 treatments and 5 patients; (bottom): 10 treatments and 15 patients. The results are obtained through 30 replicates.

A.6 Additional Results for PDX Analysis

In this section, we provide the pre-processing procedures of NIBR-PDXE and present the results for non-small lung cancer (NSCLC) and pancreatic ductal adenocarcinoma (PDAC) with tables including treatment and pathway information.

A.6.1 PDX Data Pre-Processing

We followed pre-processing procedure in Rashid et al. (2020) and imputed the missing data by k-nearest neighbor method. We take the best average response (BAR) as the response and scale the BAR by the standard deviation over all patients, treatments and across five cancers. Since the scaled BAR contains missing values, we impute the missing data by the k-nearest neighbor with $k = 10$ and compare all treatments to the untreated group. Specifically, we take $x_{ij} = \text{BAR}_{ij} - \text{BAR}_{0j}, i =$

$1 \dots I, j = 1 \dots J$ as the observed data, where $\text{BAR}_{0,j}$ is the untreated BAR for patient j.

A.6.2 Test for Distributional Assumption

Our main interest of the paper is the tree-structured covariance that models the treatment similarity. The relevant class of distributions for modeling thus consists of those whose properties are fully described through a tree-structured covariance matrix (with mean equal to zero). A natural candidate is the parameterized family of mean-zero symmetric elliptical distributions indexed by tree-structured covariance matrices, which includes the Gaussian as a special case.

From a methodological perspective, restriction in the paper to the Gaussian setting is to be viewed as a first step towards modelling using the more general elliptical family, mainly driven by computational considerations and interpretability within the context of the scientific application. Notwithstanding this, the Gaussian setup, which facilitates scalable and explicit computations, does not appear unreasonable: multivariate normality tests with the multivariate qq-plot (Figure A.13) demonstrate that BRCA (panel (A)) and CM (panel (B)) roughly fall on the 45-degree line, but CRC (panel (C)) slightly deviates from the the 45-degree lines indicating some departure from normality; this is further corroborated with the Doornik-Hansen (Doornik and Hansen, 2008) multivariate normal test, which resulted in p-values 0.0969 (BRCA), 0.0833 (CM) and <0.001 (CRC) for testing the null hypothesis that the responses were Gaussian.

With an eye towards future extensions to the elliptical family, we carried out hypothesis tests to assess the multivariate elliptical symmetry assumption; using the test proposed by Babic et al. (2021a) available in the `R` package `ellipticalsymmetry` (Babic et al., 2021b), we fail to reject the null hypothesis of elliptical symmetry with the p-values of 0.6805 for BRCA, 0.8679 for CRC, and 0.4385 for CM.

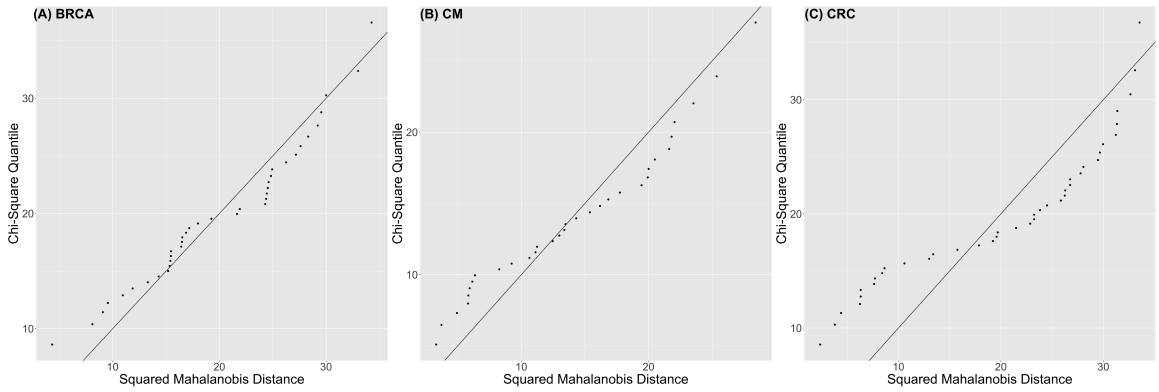


Figure A.13: The multivariate normality QQ-plot for (A) breast cancer, (B) melanoma, and (C) colorectal cancer

A.6.3 Threshold of the Co-Clustering

Generally, it is hard to recommend a universal threshold for co-clustering without considering unique patterns in each dataset. For example, different cancers may respond differently to treatments, resulting in varying degrees of tumor size shrinkage. This is reflected by the varying distributions for all the pairwise iPCPs obtained from datasets for three cancers (BRCA, CRC and CM); See the three sets of different empirical quantiles in Table A.6. Recognizing the practical utility of iPCP cutoffs, in the following, we use pairwise iPCPs to illustrate a practical strategy for determining such cut-offs; similarly for multi-way iPCPs.

First, for a "fully-exploratory" analysis, where one does not assume any prior knowledge about multiple monotherapies that share the same mechanism, we recommend ranking all the pairwise iPCPs as in Table A.6 and setting the cut-off at the 75-th percentile.

Second, for a "partially-exploratory" analysis, where one incorporates prior knowledge by assuming the PDX dataset contains two or more specific monotherapies with known and the same mechanism, we recommend using a cut-off determined by their corresponding iPCP. For example, two treatments (BKM120 and BYL719) are both PI3K inhibitors and were tested in the BRCA data with a pairwise iPCP of 0.8002,

which we recommend as a practical cut-off. If multiple such iPCPs are available for other pairs of treatments with a common mechanism, we recommend the lowest iPCP as the cut-off. In this scenario, a question may be raised regarding whether the biologically-motivated cut-off is similar to the cut-off determined by the empirical 75 percentile and which one to use. In fact, we observed that two cut-offs were practically similar. For example, the 75-th percentile of all pairwise iPCPs for BRCA is 0.753 and two treatments (binimetinib and BKM120) targeting the same pathway PI3K-MAPK-CDK have a pairwise iPCP of 0.7427. As another example, in the CM data set, the 75-th percentile of pairwise iPCPs is 0.801; the two treatments (LEE011, binimetinib) targeting the same pathway PI3K-MAPK-CDK have a iPCP of 0.8210. In practice, when both are available, we recommend using the 75 percentile cut-off for fully-exploratory analyses and using the biologically-motivated cut-off for partially-exploratory analyses.

Cancer	Min	25-th	Median	75-th	Max
BRCA	0.357	0.664	0.680	0.753	0.899
CRC	0.420	0.441	0.515	0.687	0.862
CM	0.610	0.723	0.742	0.801	0.939

Table A.6: The descriptive statistics for all possible pairs of pairwise iPCP for the breast cancer (top), colorectal cancer (middle) and the melanoma (bottom).

A.6.4 Additional Results for Monotherapy

In Main Paper, we listed the results for monotherapies targeting the cell regulated pathways. We offer more monotherapies targeting the rest two categories of the pathways.

ERBB3 and tubulin inhibitors. Our model also found high iPCP values among tubulin, ERBB3, and PI3K-MAPK-CDK inhibitors in BRCA. ERRB3 inhibitor, LJM716, exhibits high pairwise iPCP values with PI3K (BKM120: 0.7501, BYL719: 0.7513, CLR457: 0.7500), MAPK (binimetinib: 0.7811), CDK (LEE011: 0.7847) and tubulin

(paclitaxel: 0.7505) inhibitors. Since PI3K and MAPK are downstream pathways of ERBB3 (Balko et al., 2012) and CDK works closely with PI3K and MAPK (Kurtzeborn et al., 2019; Repetto et al., 2018), high iPCPs between ERBB3 inhibitor and PI3K-MAPK-CDK inhibitors are not surprising. For ERBB3 and tubulin, ERBB3 is a critical regulator of microtubule assembly (Wu et al., 2021) and tubulin plays an important role in building microtubules. Since microtubules form the skeletons of cells and are essential for cell division (Gunning et al., 2015; Haider et al., 2019), tubulin inhibitor, paclitaxel, kills cancer cell by interfering cell division and is an FDA-approved treatment. In congruence with the above results, tubulin inhibitor paclitaxel also shares high iPCPs with PI3K (BKM120: 0.8076, BYL719: 0.8063, CLR457: 0.8076), MAPK (binimetinib: 0.7433), CDK (LEE011: 0.7587) and ERBB3 (LJM716: 0.7505). In addition, another CDK4 inhibitor BPT also inhibits tubulin (Mahale et al., 2015) and PI3K inhibitor BKM120 inhibits the formation of microtubule (Bohnacker et al., 2017). Both offer additional reasons for high iPCP between tubulin and PI3K-MAPK-CDK inhibitors.

MDM2 inhibitors. We found two drugs: CGM097 and HDM201 share high iPCP values in BRCA (0.8365) and CRC (0.7860). Since CGM097 and HDM201 target the same pathway, MDM2, high iPCPs suggest a high similarity between CGM097 and HDM201 and show consistent results between our model and underlying biological mechanism. MDM2 negatively regulates the tumor suppressor, p53 (Zhao et al., 2014) and if MDM2 is suppressed by inhibitors, p53 is able to prevent tumor formation. Both CGM097 and HDM201 entered phase I clinical trial (Konopleva et al., 2020) for wild-type p53 solid tumors and leukemia, respectively.

A.6.5 R_x-Tree for Non-Small Lung Cancer (NSCLC) and Pancreatic Ductal Adenocarcinoma (PDAC)

We applied the R_x-tree on the rest two cancers in the data: non-small lung cancer (NSCLC) and pancreatic ductal adenocarcinoma (PDAC). Similar to the Figure 5 in the Main Paper, R_x-tree, pairwise iPCP and (scaled) Pearson correlation are shown in the left, middle and right panels in Figure A.14, respectively. Again, we observe that the R_x-tree and the pairwise iPCP matrix show the similar clustering patterns. For example, three PI3K inhibitors (BKM120, BYL719 and CLR457) and a combination therapy (BKM120 + binimetibin) in NSCLC form a tight subtree and are labeled by a box in the R_x-tree of Figure A.14 and a block with higher values of iPCP among therapies above also shows up in the corresponding iPCP matrix. The R_x-tree roughly clusters monotherapies targeting oncogenic process (PI3K-MAPK-CDK, MDM2 and JAK) and agrees with the biology mechanism. For example, three PI3K inhibitors (BKM120, BYL719 and CLR457) belong to a tighter subtree in both cancers. Following the same idea as the Main Paper, we further quantify the treatment similarity through iPCP. However, compared to three cancers (BRCA, CRC and CM) in the Main Paper, different problems of model fitting or interpretation lie in NSCLC and PDAC: NSCLC deviates from the normal assumption of Equation (4) (Figure A.13) and PDAC shows lower iPCP (average iPCP of PDAC: 0.4119 < BRCA: 0.6734, CRC: 0.5653, CM: 0.7535, NSCLC: 0.5817). For concerns raised above, we only verify the model through the monotherapies with known biology for each cancer.

Non-small cell lung cancer. Our model suggests high iPCP values for treatments share the same target. For example, our model shows a high iPCP among three PI3K (BKM120, BYL719 and CLR457) inhibitors: (BKM120, BYL719): 0.8402, (BKM120, CLR457): 0.8321, (BYL719, CLR457): 0.8710. For treatments with different targets, our model also exhibits a high iPCP values. For example, the monotherapy HSP990 that inhibits the heat shock protein 90 (HSP90) shows a high iPCP

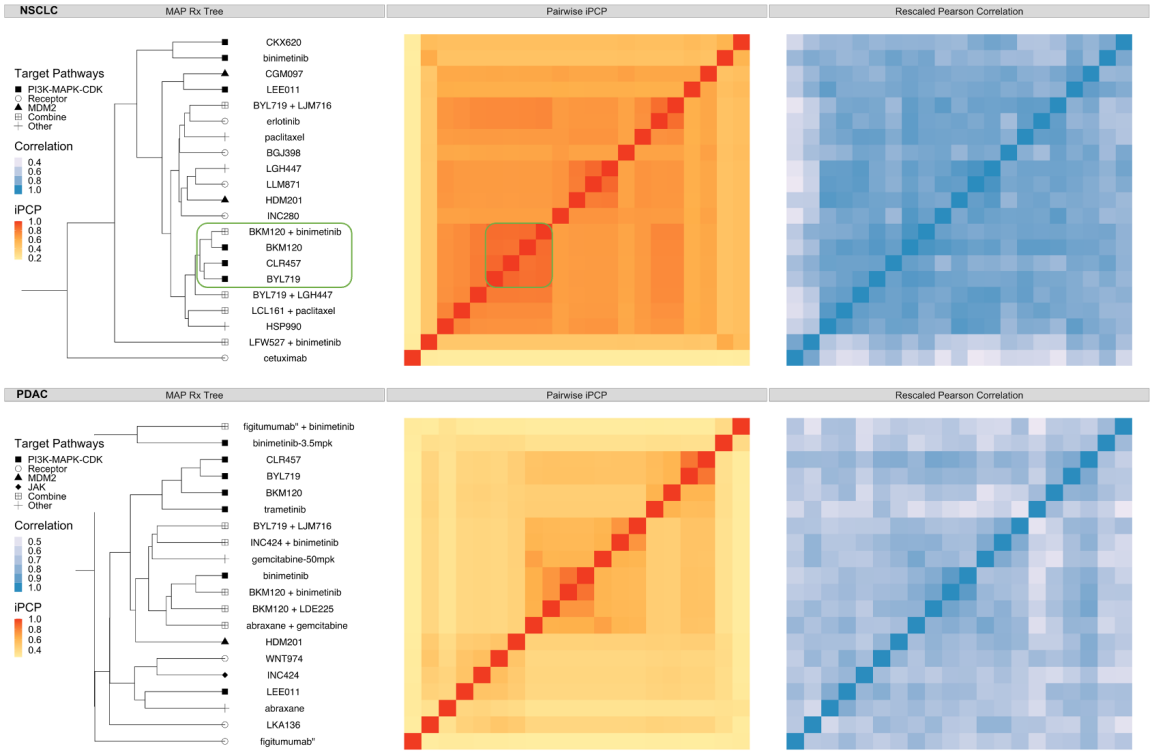


Figure A.14: The R_x-tree and iPCP for non-small cell lung cancer (NSCLC, top row) and pancreatic ductal adenocarcinoma (PDAC, lower row). Three panels in each row represent: (left) estimated R_x-tree (MAP); distinct external target pathway information is shown in distinct shapes for groups of treatments on the leaves; (middle) Estimated pairwise iPCP, i.e., the posterior mean divergence time for pairs of entities on the leaves (see the result paragraph for definition for any subset of entities); (right) Scaled Pearson correlation for each pair of treatments. The Pearson correlation $\rho \in [-1, 1]$ was scaled by $\frac{\rho+1}{2}$ to fall into $[0, 1]$. Note that the MAP visualizes the hierarchy amongst treatments; the iPCP is not calculated based on the MAP, but based on posterior tree samples (see definition in Main Paper Section 3.2)

with PI3K inhibitors ((BKM120, HSP990): 0.7108, (BYL719, HSP990): 0.7114, (CLR457,HSP990): 0.7109). Since the inhibiting of HSP90 also suppresses PI3K (Giulino-Roth et al., 2017), it is not surprising to ses a high iPCP between PI3K and HSP90 inhibitors.

Pancreatic ductal adenocarcinoma. For PDAC, our model overall suggests a lower iPCP (average iPCP of PDAC: 0.4119). Out of 91 pairs of monotherapies, only BYL719 and CLR457 share a higher iPCP (0.8415). The higher iPCP can be explained by the common target PI3K of BYL719 and CLR457.

A.6.6 R Shiny Application

We illustrate the input and outputs of the proposed method via a R Shiny application hosted on the web (Figure A.15). The visualizations are based on completed posterior computations for illustrative purposes. A user needs to specify the following inputs:

(A) Cancer type to choose the subset of data for analysis

(B) Number of treatments of interest in the subset $\mathcal{A}$ to evaluate synergy via iPCP

(C) Names of the treatments in the subset $\mathcal{A}$

Given the inputs above, the Shiny app visualizes the outputs:

(D) *maximum a posteriori* treatment tree for all the available treatments

(E) $\text{PCP}_{\mathcal{A}}(t)$ curve for the subset of treatments, $\mathcal{A}$

(F) $\text{iPCP}_{\mathcal{A}}$ value calculated from the corresponding $\text{PCP}_{\mathcal{A}}(t)$

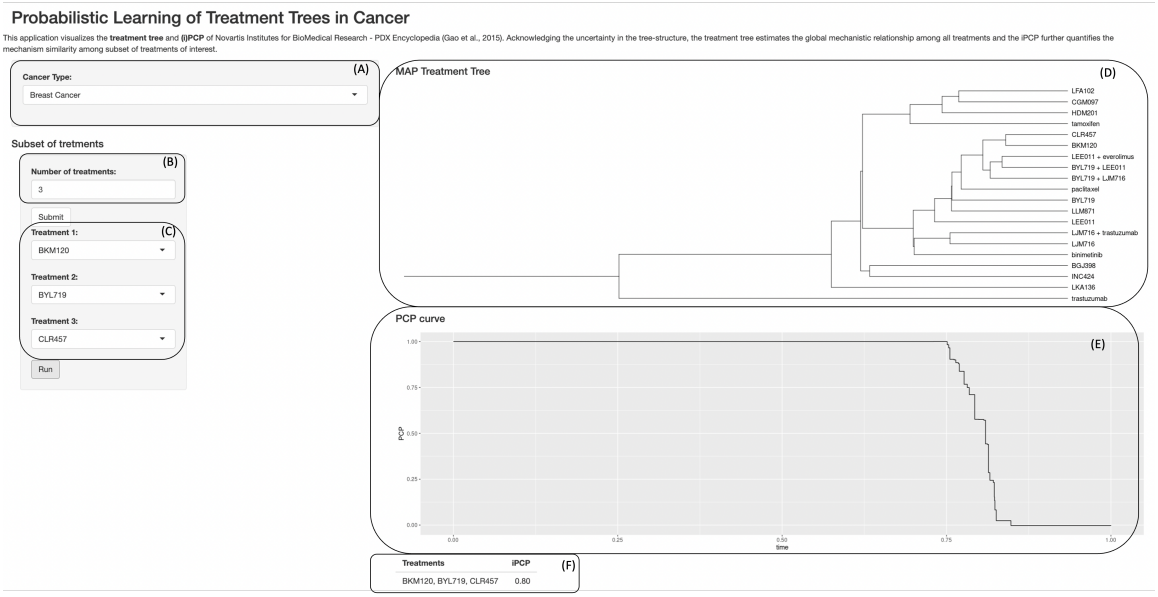


Figure A.15: R Shiny app screenshot for illustrating model inputs and outputs for analyzing PDX data (20 treatments for breast cancer); the PCP curve and iPCP value are computed for a subset of three selected treatments.

A.7 Random Effects Model for Multiple Animals Design

The current work is built upon the $1 \times 1 \times 1$ design, but multi-replicate experiment set-up is extremely relevant in practice, and is an interesting direction for future work. Several possible modeling options can extend our work to adapt to the multi-replicate experimental design. Following the comment, we consider two different scenarios for the response: (i) homogeneous and (ii) heterogeneous responses. First, recent literature (Evrard et al., 2020) suggests robustness for PDX studies (including BAR and other tumor volume measurements) under different protocol and mice replicates and implies the homogeneous responses. When the responses are homogeneous, we can simply average the outcomes over the replicates, which makes out method directly applicable. Alternatively, when the responses are heterogeneous, we can use random effects for multiple replicates nested within each patient. We can incorporate the random effects either in the mean structure or in the variance structure. Specifically, given a PDX experiment with I treatments and J patients, for each treatment, we consider K_j independent mice replicates for the j-th patient, $j = 1, \ldots, J$. Let $\boldsymbol{X}_{\cdot jk} = [X_{1jk}, \ldots, X_{Ijk}] \in \mathbb{R}^I$ be a vector of BAR response across I treatments from the k-th replicate of patient j. Following Proposition 1, we may consider adding random effects in the mean structure:

$$\boldsymbol{X}_{\cdot jk} \overset{iid}{\sim} \boldsymbol{N}_I(\boldsymbol{\mu}_{jk}, \sigma^2 \boldsymbol{\Sigma}^{\mathcal{T}});\ \boldsymbol{\mu}_{jk} \sim \boldsymbol{N}_I(\boldsymbol{0}, \boldsymbol{\Omega}),\ j = 1, \ldots, J; k = 1, \ldots, K_j,$$

where the $\boldsymbol{\mu}_{jk} = [\mu_{1jk}, \ldots, \mu_{Ijk}]$ is the normal random effect with mean zero and a variance $\boldsymbol{\Omega}$. We assume $\boldsymbol{\Omega}$ to be diagonal to maintain the ultrametric property for the marginal variance of $\mathrm{Var}(\boldsymbol{X}_{\cdot jk}) = \sigma^2 \boldsymbol{\Sigma}^{\mathcal{T}} + \boldsymbol{\Omega}$.

One may instead include random effects in the variance and the corresponding tree-structured matrix. Following the same notation, we can formulate the distribu-

tion as

$$\boldsymbol{X}_{.jk} \overset{iid}{\sim} \boldsymbol{N}_I(\mathbf{0}, \sigma^2 \boldsymbol{\Sigma}_k^{\mathcal{T}}),\ k = 1, \ldots, K_j,$$

where $\boldsymbol{\Sigma}_k^{\mathcal{T}}$ is the tree-structured matrix for each replicate. We can further consider two cases with (i) pooling all tree-structured matrix of $\boldsymbol{\Sigma}_k^{\mathcal{T}} = \boldsymbol{\Sigma}^{\mathcal{T}}$ for all $k = 1, \ldots, K_j$ and (ii) assigning different $\boldsymbol{\Sigma}_k^{\mathcal{T}}$ for each k. The case (i) of pooling all tree-structured matrix is the same as the original Proposition 1 and ignores the heterogeneity of the responses. For the case (ii), we can further assign a prior distribution on each tree-structured matrix and include the external covariate information (e.g. heterogeneity of the response) in the prior distribution.

Table A.7: Full CPUs series used for computations.

Intel Xeon X series	X5660@2.80GHz X5680@3.33GHz
Intel Xeon E series	E5-24400@2.40GHz E5-24700@2.30GHz E5-24500@2.10GHz E5-2650v3@2.30GHz E5-2650v4@2.20GHz E5-2690v4@2.60GHz E5-2690v4@2.60GHz

Table A.8: Pathways full names and the corresponding abbreviations.

Abbreviation	Target Name
PI3K	Phosphoinositide 3-kinases
CDK	Cyclin-dependent kinases
MAPK	Mitogen-activated protein kinases
JAK	Janus kinase
MDM2	Murine double minute 2
BRAF	Serine/threonine-protein kinase B-Raf
MTOR	Mechanistic target of rapamycin
EGFR/ERBB	Epidermal growth factor receptor
SMO	Smoothened
TNKS	Tankyrase
PIM	Proto-oncogene serine/threonine-protein kinase Pim-1
BIRC2	Baculoviral IAP repeat-containing protein 2
IGF1R	Insulin-like growth factor 1 receptor

Table A.9: Monotherapy names with targets. Different target groups are labeled differently in the Figure 5 and Figure A.14.

Treatment name	Other names	Trade name	Target	Target Group
5FU	Fluorouracil	Adrucil	chemotherapy	Other
abraxane	nab-paclitaxel	abraxane	Tubulin	Other
BGJ398	Infigratinib		FGFR	Receptor
binimetinib	MEK162	Mektovi	MAPK	PI3K-MAPK-CDK
BKM120	Buparlisib		PI3K	PI3K-MAPK-CDK
BYL719	Alpelisib	Piqray	PI3K	PI3K-MAPK-CDK
cetuximab		Erbitux	EGFR	Receptor
CGM097			MDM2	MDM2
CKX620			MAPK	PI3K-MAPK-CDK
CLR457			PI3K	PI3K-MAPK-CDK
dacarbazine		DTIC-Dome	chemotherapy	Other
encorafenib	LGX818	Braftovi	BRAF	BRAF
erlotinib	Erlotinib hydrochloride	Tarceva	EGFR	Receptor
figitumumab	CP-751871		IGF1R	Receptor
gemcitabine		Gemzar	chemotherapy	Other
HDM201	Siremadlin		MDM2	MDM2
HSP990			HSP90	Other
INC280	Capmatinib	Tabrecta	MET	Receptor
INC424	Ruxolitinib	Jakafi and Jakavi	JAK	JAK
LDE225	Sonidegib	Odomzo	SMO	Receptor
LDK378	Ceritinib	Zykadia	ALK	Receptor
LEE011	Ribociclib	Kisqal	CDK	PI3K-MAPK-CDK
LFA102			PRLR	Receptor
LGH447			PIM	Other
LGW813			IAP	Other
LJC049			TNKS	Other
LJM716	Elgemtumab		ERBB3	Receptor
LKA136			NTRK	Receptor
LLM871			FGFR2/4	Receptor
paclitaxel		Taxol	Tubulin	Other
tamoxifen		Nolvadex	ESR1	Receptor
TAS266			DR5	Receptor
trametinib	GSK1120212	Mekinist	MAPK	PI3K-MAPK-CDK
trastuzumab		Herceptin	ERBB2	Receptor
WNT974			PORCN	Receptor

Table A.10: Combination therapy full names with known targets.

Combination Therapies	Known Target Pathways	Cancer
abraxane+gemcitabine	Tubulin+chemotherapy	PDAC
BKM120+binimetinib	PI3K+MAPK	NSCLC,PDAC
BKM120+encorafenib	PI3K+BRAF	CM
BKM120+LDE225	PI3K+SMO	PDAC
BKM120+LJC049	PI3K+TNKS	CRC
BYL719+binimetinib	PI3K+MAPK	CRC
BYL719+cetuximab	PI3K+EGFR	CRC
BYL719+cetuximab+encorafenib	PI3K+EGFR+BRAF	CRC
BYL719+encorafenib	PI3K+BRAF	CRC
BYL719+LEE011	PI3K+CDK	BRCA
BYL719+LGH447	PI3K+PIM	NSCLC
BYL719+LJM716	PI3K+ERBB3	BRCA,CRC,NSCLC,PDAC
cetuximab+encorafenib	EGFR+BRAF	CRC
encorafenib+binimetinib	BRAF+MAPK	CM
figitumumab+binimetinib	IGF1R+MAPK	PDAC
INC424+binimetinib	JAK+MAPK	PDAC
LCL161+paclitaxel	BIRC2+Tubulin	NSCLC
LEE011+encorafenib	CDK+BRAF	CM
LEE011+everolimus	CDK+MTOR	BRCA
LFW527+binimetinib	IGF1R+MAPK	NSCLC
LJM716+trastuzumab	ERBB3+ERBB2	BRCA

APPENDIX B

Appendix of Chapter III

B.1 Details of BHV Space as a CAT(0) Space

A CAT(0) space entails the nonpositive curvature space for $\mathcal{U}_p$. Curvature in this context is determined by comparing the triangles within the space to those in Euclidean space (Bridson and Haefilger, 1999). Specifically, a triangle in a CAT(0) space is not "thicker" than the corresponding Euclidean triangle with the same side lengths. Visually speaking, the flat edges of Euclidean triangles represent zero curvature for Euclidean space, while edges in the CAT(0) triangle are curved and drawn inside the triangle, resulting in the CAT(0) triangle being "thinner" than the Euclidean triangle. For the BHV space, the common boundary of different orthants are glued isometrically by considering an equivalent relation that locates trees on the boundary of different orthants at the same point in the BHV space. By doing so, the manifestation of nonpositive curvature is observed in the boundary of different orthants. Consider three vertices lying in three nearby orthants that share a common boundary (Panel (F) in Figure III.2). These three vertices form a triangle with all edges contained within the space, resulting in a triangle thinner than the Euclidean

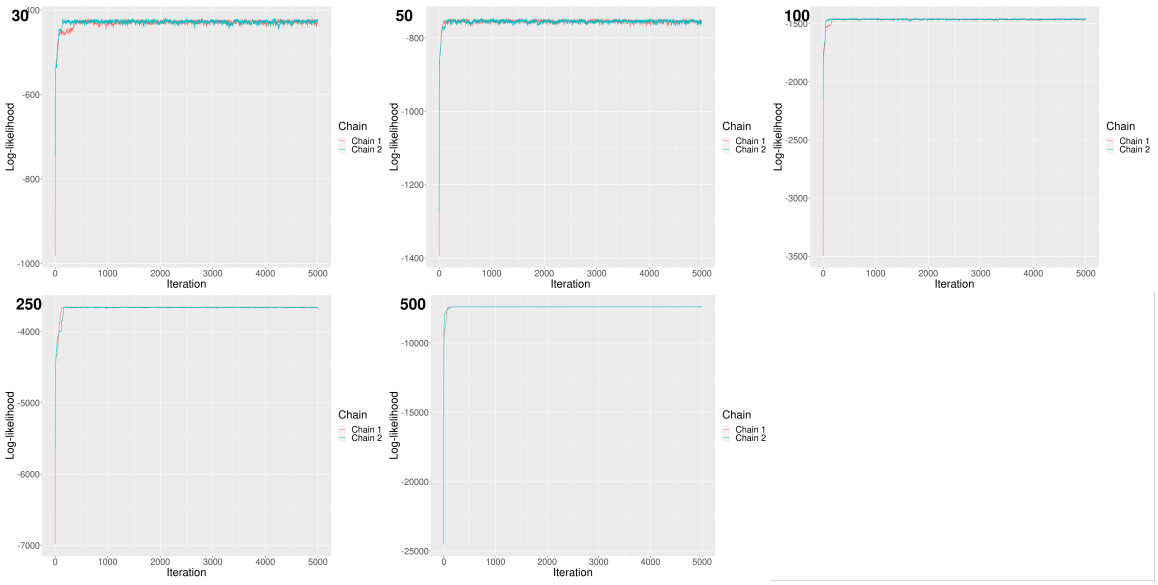


Figure B.1**:** Convergence diagnostics for the algorithm using the likelihood trace plot. Two chains of the same algorithm are initiated by different trees, shown by two colors, across five different sample sizes of $n \in \{30, 50, 100, 250, 500\}$.

triangle with the nonpositive curvature of the BHV space and allowing the algorithm to find the compatible splits easily.

B.2 Additional Simulation Results of Ultrametric Matrices

B.2.1 Convergence Diagnostics

We examine the convergence of our algorithm through the likelihood trace plot and ensure convergence likelihood from different initializations. Specifically, given the same dataset, we run the same algorithm with two chains initiated by two different trees and plot the likelihood over iterations. We randomly chose five likelihood trace plots, each from different sample sizes of $n \in \{30, 50, 100, 250, 500\}$, as shown in Figure B.1. Obviously, the likelihood increases rapidly and remains at a relatively high plateau. More importantly, both chains converge to a similar level of likelihood, indicating the convergence of the algorithm.

B.2.2 Comparing Different Algorithms

We compare our algorithm to the algorithm from Nye (2020) and empirically investigate the rate of convergence through the likelihood trace plot shown in Figure B.2. The algorithm from Nye (2020) enables the algorithm to propose a candidate edge set that is the same as the edge set from the previous iteration, resulting in slower convergence. We briefly describe Nye's algorithm and refer the reader to the original paper for more details. Essentially, Nye's algorithm is still an MH update and changes the proposal function illustrated in Step 3 to 6 in Algorithm 1. Instead of directly removing the internal split in Step 3, Nye's algorithm generates a branch length difference from a normal distribution of $\delta \sim N(0, \sigma_{\mathcal{E}})$ and lets $c = |e_A| + \delta$. When $c > 0$, the algorithm will stay in the same edge set with $e_B = e_A$ and $|e_B| = c$. Otherwise, when $c \leq 0$, the algorithm will propose a candidate split e_B from nearby orthants and assign the branch length of $|e_B| = -c$. However, Nye's algorithm does not exclude the original split from the candidates, resulting in a positive probability of staying in the same edge set and, therefore, slower convergence. We implement Nye's algorithm and compare it to our algorithm. Figure B.2 empirically compares the rate of convergence. Obviously, our algorithm converges faster than Nye's algorithm under five different sample sizes.

B.2.3 Element-wise Coverage for t-dsitribution

We show the nominal coverage for element-wise 95% credible intervals under the mis-specified t-distribution. Specifically, the results of nominal coverage for t_4 and t_3 for five different sizes are shown in Figure B.3 and B.4, respectively. We observe a similar pattern for the nominal coverages from t_4 and t_3 distributions. Elements in the last row and column correspond to zero elements in the true covariance and result in an estimated coverage close to one. For non-zero elements in the true covariance, we observe that the nominal coverages for t-distribution are moderate when the sample

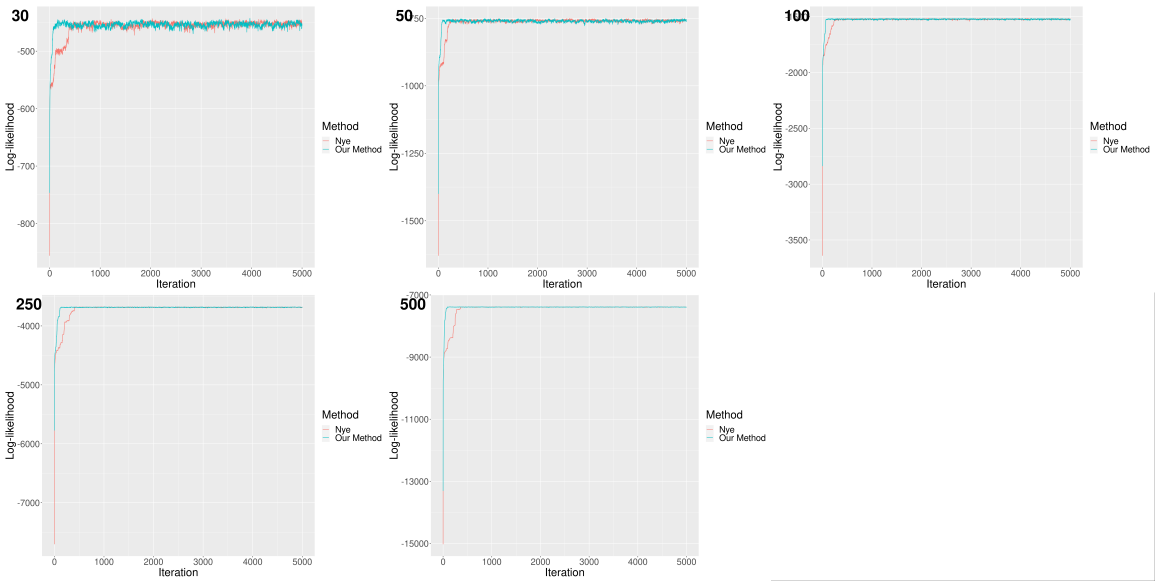


Figure B.**2:** Empirical comparison of our proposed algorithm (blue) and the algorithm from Nye (2020) (red) in terms of the rate of convergence under five different sample sizes of $n \in \{30, 50, 100, 250, 500\}$.

size is small ($n = 30$). However, when the sample size increases, the nominal coverage corresponds to non-zero elements in the true covariance worsens. Unfortunately, when our algorithm is mis-specified, it does not generate posterior matrices that converge to the true matrix. We conjecture that our algorithm converges to incorrect branch lengths when the model is mis-specified. This conjecture is supported by Table 3.1, which indicates that our algorithm still generates posterior samples with the correct topology for t-distribution.

B.2.4 Topology Trajectory for the Proposed Method

In this section, we examine the trajectory of our proposed algorithm in BHV space. Specifically, we initiate our algorithm with trees that are far away (in terms of BHV distance) from the true tree and track the topologies generated by our algorithm. Figure B.5 presents the trajectory of our algorithm with 15 different initial trees in terms of the BHV distance between the estimated tree and the true tree. For each posterior tree, we color the tree based on the corresponding topology. Obviously,

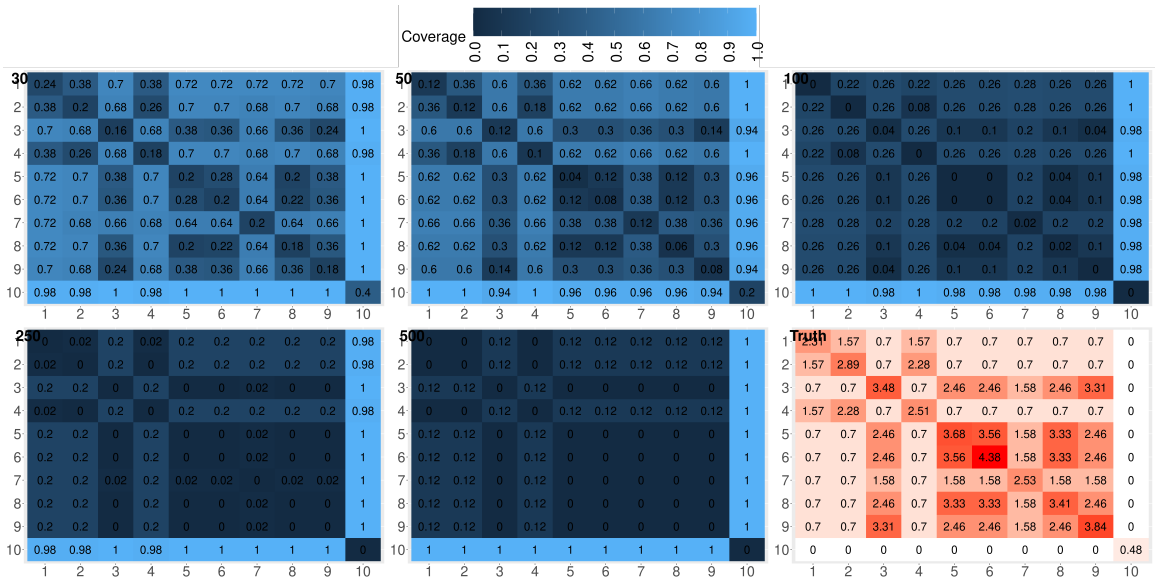


Figure B.**3:** Element-wise coverage from the 95% credit interval for the mis-specified t-distribution of degree of freedom four under five different sample sizes. The true underlying covariance is shown in the lower right panel.

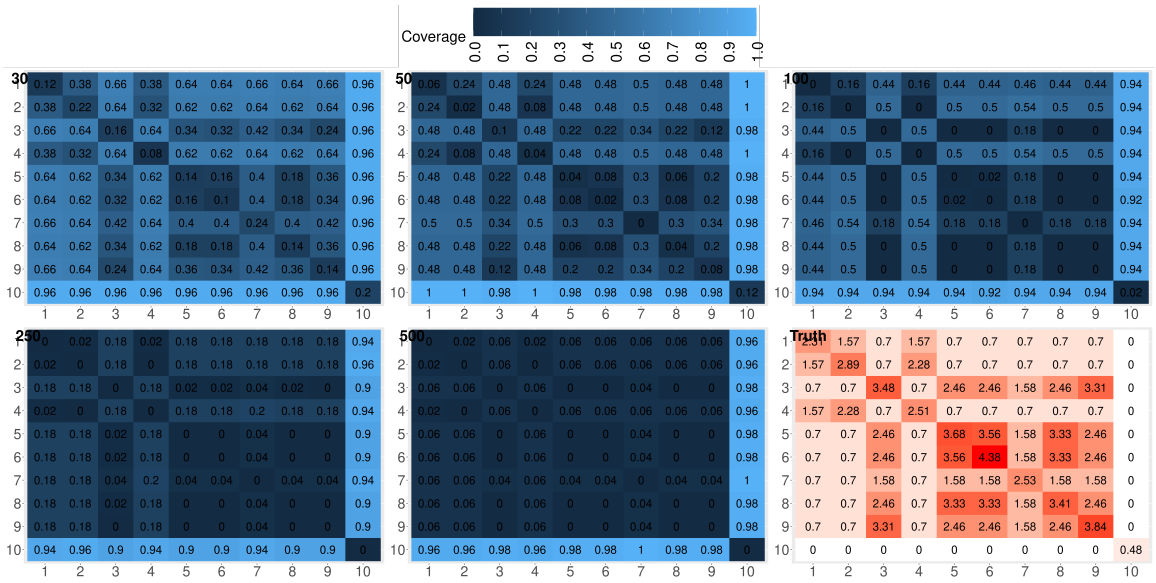


Figure B.**4:** Element-wise coverage from the 95% credit interval for the mis-specified t-distribution of degree of freedom three under five different sample sizes. The true underlying covariance is shown in the lower right panel.

all initial trees are distant from the true tree, with higher BHV distances. Over iterations, our algorithm traverses different nearby orthants and quickly moves to the correct topology (topology 1) with a smaller BHV distance to the true tree.

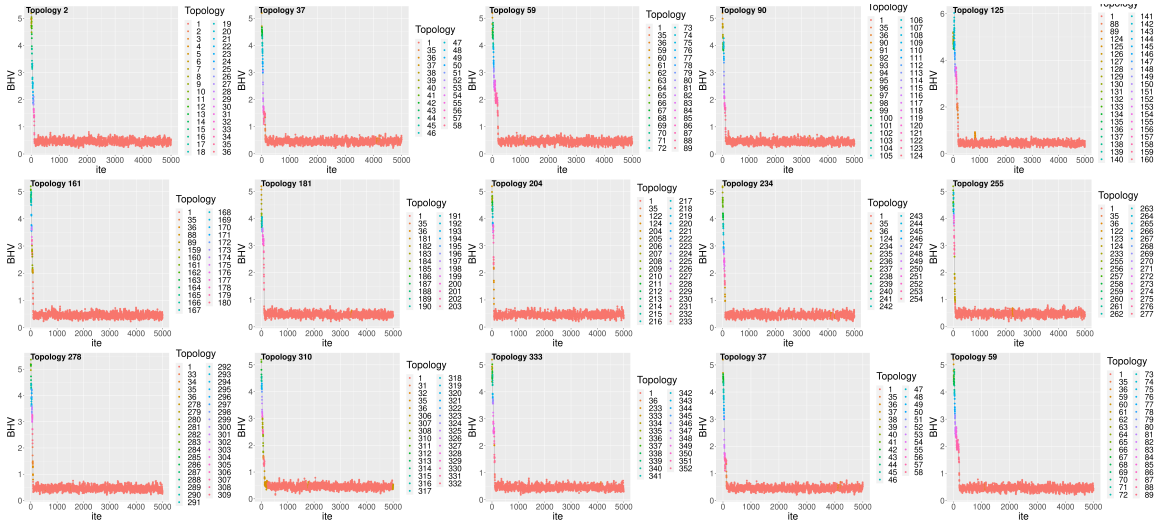


Figure B.5: Trajectory of our algorithm in terms of BHV distances. Over iterations, BHV distances between each posterior tree and the true tree are measured. Each posterior sample is colored according to the corresponding topology. The same algorithm is initiated by 15 different trees that are far away (in terms of the BHV distance) from the true tree. Our algorithm traverses different orthants and arrives at the true topology (topology 1) quickly after a few iterations.

B.2.5 Simulation Results for Underlying Trees from the Ultrametric Tree

In this Section, we demonstrate additional simulation results when all leaves in the true underlying tree are equidistant to the root. Specifically, we obtain a tree from the coalescence model implemented by the function `rcoal` in the `R` package `ape` and generate the data from the normal and t-distribution described in Main Paper Section 3.5. We run Algorithm 1 without restricting the prior on the branch lengths for $5,000$ iterations and discard the first $4,000$ iterations. We summarize the posterior samples by the point estimation and the quantify the uncertainty through the element-wise 95% credible interval. The performance of the point estimator is compared to the projection-based method of Bravo et al. (2009) and the sample covariance under the measurement of the BHV distance (Owen and Provan, 2011) and the matrix norm. For the uncertainty quantification, we calculate the nominal coverage of the 95% credible interval. All results were obtained from 50 independent replicates.

The results of the point estimators are shown in Figure B.6. For the point esti-

mator, the mean and MAP trees from our method are comparable to the estimated matrix from Bravo et al. (2009) and sample covariance in terms of BHV distance and matrix norm across different data generating mechanisms and sample sizes. When the model is correctly specified, all methods benefit from the increase in the sample size, resulting in a smaller distance to the true tree. For the mis-specified scenario, the advantage from the larger sample size is moderate. Essentially, when all leaves in the true underlying tree are equidistant to the root, our algorithm still generates posterior samples that are comparable to existing methods with a similar level of performance in terms of the BHV distance and matrix norm.

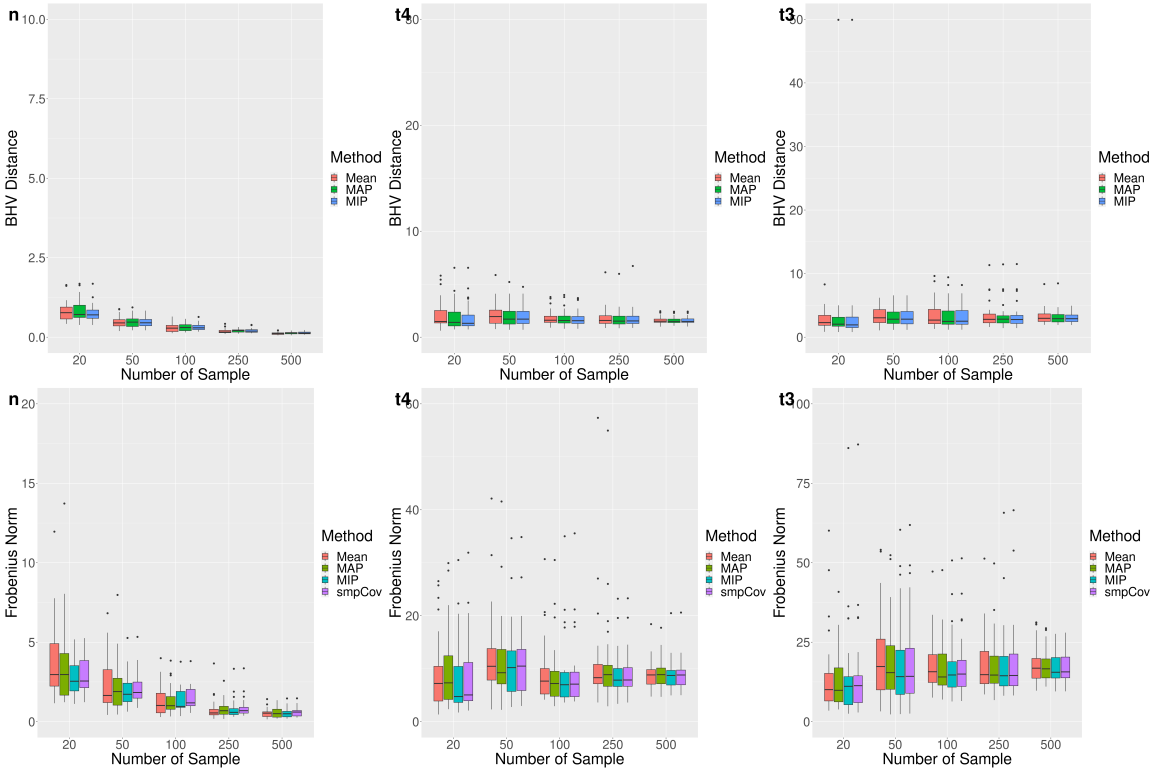


Figure B.6: Distances between the estimated matrix and the true matrix under different data generating mechanism and sample sizes. The mean (red) and MAP tree (green) from our method is comparable to competing methods (blue for MIP and purple for sample covariance) in terms of the BHV distance (top row) and matrix norm (bottom row).

Figure B.7 demonstrates the nominal coverage of the element-wise 95% credible interval with the true generating covariance in the lower right panel. As we expected, the diagonal elements in the true underlying covariance are equivalent, implying that

all leaves in the true underlying tree are equidistant to the root. Similar to the results shown in Main Paper Section 3.5, the 95% credible interval gives a high nominal coverage (around 0.73 to 1), and the estimated coverage is higher when the sample size increases. In summary, our algorithm can efficiently draw posterior samples of matrices under different conditions imposed on the branch lengths of the true underlying tree.

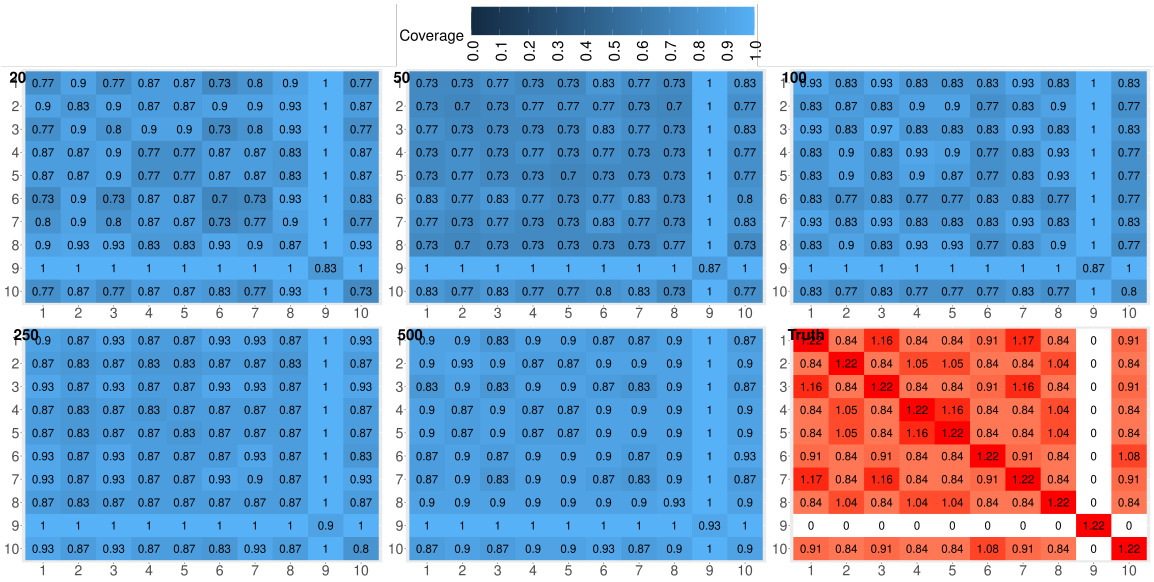


Figure B.7: Element-wise coverage from the 95% credit interval for the correct specified normal distribution with five different sample sizes with the true underlying covariance in the lower right panel. Equal diagonal elements in the true underlying covariance indicate that all leaves in the true underlying tree are equidistant to the root.

APPENDIX C

Appendix of Chapter IV

C.1 Proof for Proposition 4.3.1

We provide a detailed proof for Proposition 4.3.1 in the Main Paper. We proceed this proof through two steps. First, we show the conditional sign independence for the undirected graph and the equivalent graphical regression model without covariates. We then can extend to result to the regression model with the covariates.

Proof. We first show the undirected case with scalar coefficients β. Denote the $\boldsymbol{D} = \text{diag}(\frac{1}{d_1}, \ldots, \frac{1}{d_p})$ a diagonal matrix of dimension p by p. Following the assumption of normal conditional distribution of (4.3), the joint distribution of $\boldsymbol{Y}\boldsymbol{D} = [\frac{Y_1}{d_1}, \ldots, \frac{Y_p}{d_p}]$ is a multivariate normal distribution $\boldsymbol{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. From the Proposition (C.5) of Lauritzen (1996), we can first partition the joint distribution with

$$\boldsymbol{Y}\boldsymbol{D} \mid \boldsymbol{D} = \begin{bmatrix} Y_j/d_j \\ Y_{j'}/d_{j'} \\ \boldsymbol{Y}_{V\setminus\{j,j'\}}\boldsymbol{D}_{V\setminus\{j,j'\}} \end{bmatrix} \sim \boldsymbol{N}\left(\begin{bmatrix} \mu_j \\ \mu_{j'} \\ \boldsymbol{\mu}_{V\setminus\{j,j'\}} \end{bmatrix}, \begin{bmatrix} \kappa_{jj} & \kappa_{jj'} & \boldsymbol{\kappa}_{j.}^{\mathsf{T}} \\ \kappa_{jj'} & \kappa_{jj} & \boldsymbol{\kappa}_{j'.}^{\mathsf{T}} \\ \boldsymbol{\kappa}_{j.} & \boldsymbol{\kappa}_{j'.} & \mathcal{K}_{V\setminus\{j',j\}} \end{bmatrix}^{-1} \right)$$

where $\boldsymbol{\Sigma} = \mathcal{K}^{-1}$, $\boldsymbol{\kappa}_{j\cdot} = [\kappa_{jv}]$ and $\boldsymbol{\kappa}_{j'\cdot} = [\kappa_{j'v}], v \in V \setminus \{j, j'\}$. Thus, the conditional distribution of $\frac{Y_j}{d_j}$ and $\frac{Y_{j'}}{d_{j'}}$ is a bivariate normal distribution:

$$\begin{bmatrix} Y_j/d_j \\ Y_{j'}/d_{j'} \end{bmatrix} \Bigg| \boldsymbol{Y}_{V\setminus\{j,j'\}}, \boldsymbol{D} \sim \boldsymbol{N}_2 \left(\begin{bmatrix} \mu_{jD} \\ \mu_{j'D} \end{bmatrix}, \mathcal{K}_{jj'}^{-1} \right), \tag{C.1}$$

where $\mathcal{K}_{jj'} = \begin{bmatrix} \kappa_{jj} & \kappa_{jj'} \\ \kappa_{jj'} & \kappa_{j'j'} \end{bmatrix}$ and $\begin{bmatrix} \mu_{jD} \\ \mu_{j'D} \end{bmatrix} = \begin{bmatrix} \mu_j \\ \mu_{j'} \end{bmatrix} - \mathcal{K}_{jj'}^{-1} \begin{bmatrix} \boldsymbol{\kappa}_{j\cdot}^{\mathsf{T}} \\ \boldsymbol{\kappa}_{j'\cdot}^{\mathsf{T}} \end{bmatrix} (\boldsymbol{Y}_{V\setminus\{j,j'\}} \boldsymbol{D}_{V\setminus\{j,j'\}} - \boldsymbol{\mu}_{V\setminus\{j,j'\}}) = \begin{bmatrix} \mu_j \\ \mu_{j'} \end{bmatrix} - \mathcal{K}_{jj'}^{-1} \begin{bmatrix} \sum_{v\in V\setminus\{j,j'\}} \kappa_{jv}(Y_v/d_v - \mu_v) \\ \sum_{v\in V\setminus\{j,j'\}} \kappa_{j'v}(Y_v/d_v - \mu_v) \end{bmatrix}$.

Now, we can show the univariate distribution of $\frac{Y_j}{d_j}$:

$$\frac{Y_j}{d_j} \Bigg| \boldsymbol{Y}_{V\setminus\{j\}}, \boldsymbol{D} \sim N\big(\widetilde{\mu}_{jD}, \kappa_{jj}^{-1}\big),$$

where $\widetilde{\mu}_{jD} = \mu_j - \frac{1}{\kappa_{jj}} \sum_{v\in V\setminus\{j\}} \kappa_{jv}(Y_v/d_v - \mu_v)$. When Y_j/d_j and $Y_{j'}/d_{j'}$ are independent, $\mu_{jD} = \mu_j - \kappa_{jj}^{-1} \sum_{v\in V\setminus\{j,j'\}} \kappa_{jv}(Y_v/d_v - \mu_v) = \widetilde{\mu}_{jD}$ and $\kappa_{jj'} = 0$. Thus,

$$p(Y_j/d_j | \boldsymbol{Y}_{V\setminus\{j,j'\}}, \boldsymbol{D}) = p(Y_j/d_j | \boldsymbol{Y}_{V\setminus\{j\}}, \boldsymbol{D}). \tag{C.2}$$

However, the conditional independence does not hold after integrating out the random scaling $\boldsymbol{D}$. Specifically, the integration of $\boldsymbol{D}$ conditioning on $\boldsymbol{Y}_{V\setminus\{j,j'\}}$ and $\boldsymbol{Y}_{V\setminus\{j\}}$ are different. We can see that from the following expectation values.

$$\mathbb{E}[Y_j | \boldsymbol{Y}_{V\setminus\{j,j'\}}] = \mathbb{E}_{\boldsymbol{D}|\boldsymbol{Y}_{V\setminus\{j,j'\}}}[\mathbb{E}[Y_j | \boldsymbol{Y}_{V\setminus\{j,j'\}}, \boldsymbol{D}]] = \mathbb{E}_{\boldsymbol{D}|\boldsymbol{Y}_{V\setminus\{j,j'\}}}[d_j \mu_{jD}]$$

$$\mathbb{E}[Y_j | \boldsymbol{Y}_{V\setminus\{j\}}] = \mathbb{E}_{\boldsymbol{D}|\boldsymbol{Y}_{V\setminus\{j\}}}\mathbb{E}[[Y_j | \boldsymbol{Y}_{V\setminus\{j\}}, \boldsymbol{D}]] = \mathbb{E}_{\boldsymbol{D}|\boldsymbol{Y}_{V\setminus\{j\}}}[d_j \mu_{jD}]$$

Since the conditional distributions of $\boldsymbol{D}|\boldsymbol{Y}_{V\setminus\{j\}}$ and $\boldsymbol{D}|\boldsymbol{Y}_{V\setminus\{j,j'\}}$ are not equal, the expectation values are different. Of note, the conditional sign independence still hold

due to following equations:

$$\begin{aligned}
\mathbb{P}(Y_j < 0|\boldsymbol{Y}_{V\setminus\{j,j'\}}) &= \mathbb{E}_{\boldsymbol{D}|\boldsymbol{Y}_{V\setminus\{j,j'\}}}\left[\mathbb{P}(Y_j < 0|\boldsymbol{Y}_{V\setminus\{j,j'\}}, \boldsymbol{D})\right] \\
&= \mathbb{E}_{\boldsymbol{D}|\boldsymbol{Y}_{V\setminus\{j,j'\}}}\left[\mathbb{P}(Y_j/d_j < 0|\boldsymbol{Y}_{V\setminus\{j,j'\}}, \boldsymbol{D})\right] \\
&= \mathbb{E}_{\boldsymbol{D}|\boldsymbol{Y}_{V\setminus\{j,j'\}}}\left[\mathbb{P}(\kappa_{jj}^{1/2}(Y_j/d_j - \mu_{jD}) < -\kappa_{jj}^{1/2}\mu_{jD}|\boldsymbol{Y}_{V\setminus\{j,j'\}}, \boldsymbol{D})\right] \\
&= \mathbb{E}_{\boldsymbol{D}_{V\setminus\{j,j'\}}|\boldsymbol{Y}_{V\setminus\{j,j'\}}}\left[\Phi(-\kappa_{jj}^{1/2}\mu_{j'D})\right] \\
&= \mathbb{E}_{\boldsymbol{D}_{V\setminus\{j\}}|\boldsymbol{Y}_{V\setminus\{h\}}}\left[\Phi(-\kappa_{jj}^{1/2}\mu_{jD})\right] \\
&= \mathbb{P}(Y_j < 0|\boldsymbol{Y}_{V\setminus\{j\}}),
\end{aligned}$$

where $\Phi(.)$ is the cdf of standard univariate normal distribution. The fourth and the fifth equivalence hold since $\mu_{jD} = \mu_j - \kappa_{jj}^{-1}\sum_{v\in V\setminus\{j,j'\}}\kappa_{jv}(Y_v/d_v - \mu_v)$ does not depend on $(d_j, d_{j'}, Y_j, Y_{j'})$.

By comparing the conditional distribution of (C.2) with the Equation (4.3), we can view the conditional distribution of (C.2) as a regression model with dependent variable Y_j/d_j, the independent variable Y_v/d_v and $\beta_{jv} = -\kappa_{jv}/\kappa_{jj}$ for every $v \in V \setminus \{j\}$. Obviously, $\beta_{jv} = 0$ when $\kappa_{jv} = 0$ implying the following conditional independence:

$$p(Y_j/d_j|\boldsymbol{Y}_{V\setminus\{j\}}, Y_{j'}, \boldsymbol{D}) = p(Y_j/d_j|\boldsymbol{Y}_{V\setminus\{j\}}, \boldsymbol{D}).$$

and the conditional sign independence:

$$\begin{aligned}
\mathbb{P}(Y_j < 0|\boldsymbol{Y}_{V\setminus\{j\}}, Y_{j'}) &= \mathbb{E}_{D|\boldsymbol{Y}_{V\setminus\{j\}}, Y_{j'}}[\mathbb{P}(Y_j < 0|\boldsymbol{Y}_{V\setminus\{j\}}, Y_{j'}, \boldsymbol{D})] \\
&= \mathbb{E}_{D|\boldsymbol{Y}_{V\setminus\{j\}}, Y_{j'}}[\Phi(-\kappa_{jj}^{1/2}\mu_{jD})] \\
&= \mathbb{E}_{D|\boldsymbol{Y}_{V\setminus\{j\}}}[\Phi(-\kappa_{jj}^{1/2}\mu_{jD})] \\
&= \mathbb{P}(Y_j < 0|\boldsymbol{Y}_{V\setminus\{j\}})
\end{aligned}$$

Last, we include covariates in the model with functional coefficients $\beta(\boldsymbol{X})$. Assume

the joint distribution of $\boldsymbol{YD}$ follows a multivariate normal distribution with a mean zero and a functional precision depending on the covariates $\boldsymbol{X} = [X_1, \ldots, X_q]^\mathsf{T}$. Specifically, the joint distribution can be written as

$$\boldsymbol{YD} \mid \boldsymbol{D}, \boldsymbol{X} = \begin{bmatrix} Y_j/d_j \\ Y_{j'}/d_{j'} \\ \boldsymbol{Y}_{V\setminus\{j,j'\}}\boldsymbol{D}_{V\setminus\{j,j'\}} \end{bmatrix} \sim \boldsymbol{N}\left(\boldsymbol{0}, \boldsymbol{\Omega}(\boldsymbol{X})^{-1}\right),$$

where $\boldsymbol{\Omega}(\boldsymbol{X})$ is the functional precision matrix of

$$\boldsymbol{\Omega}(\boldsymbol{X}) = \begin{bmatrix} \kappa_{jj}(\boldsymbol{X}) & \kappa_{jj'}(\boldsymbol{X}) & \boldsymbol{\kappa}_{j\cdot}(\boldsymbol{X})^\mathsf{T} \\ \kappa_{jj'}(\boldsymbol{X}) & \kappa_{j'j'}(\boldsymbol{X}) & \boldsymbol{\kappa}_{j'\cdot}(\boldsymbol{X})^\mathsf{T} \\ \boldsymbol{\kappa}_{j\cdot}(\boldsymbol{X}) & \boldsymbol{\kappa}_{j'\cdot}(\boldsymbol{X}) & \mathcal{K}_{V\setminus\{j,j'\}}(\boldsymbol{X}) \end{bmatrix}.$$

We can therefore have conditional distribution of $\frac{Y_j}{d_j}$ as:

$$\left.\frac{Y_j}{d_j}\right| \boldsymbol{Y}_{V\setminus\{j\}}, \boldsymbol{D}, \boldsymbol{X} \sim N\left(\widetilde{\mu}_{jD}(\boldsymbol{X}), \kappa_{jj}^{-1}(\boldsymbol{X})\right), \tag{C.3}$$

where $\widetilde{\mu}_{jD}(\boldsymbol{X}) = -\frac{1}{\kappa_{jj}(\boldsymbol{X})}\sum_{v\in V\setminus\{j\}\cup\{j'\}}\kappa_{jv}(\boldsymbol{X})(Y_v/d_v)$. By comparing the conditional distribution of (C.3) and (4.3), we can define the functional coefficients $\beta_{jv}(\boldsymbol{X}) = -\kappa_{jv}(\boldsymbol{X})/\kappa_{jj}(\boldsymbol{X})$. Therefore, the covariance of the joint distribution becomes

$$\boldsymbol{\Sigma}(\boldsymbol{X}) = \begin{bmatrix} \kappa_{11}(\boldsymbol{X}) & \kappa_{12}(\boldsymbol{X}) & \ldots & \kappa_{1p}(\boldsymbol{X}) \\ \kappa_{12}(\boldsymbol{X}) & \kappa_{22}(\boldsymbol{X}) & \ldots & \kappa_{2p}(\boldsymbol{X}) \\ \kappa_{13}(\boldsymbol{X}) & \kappa_{32}(\boldsymbol{X}) & \ldots & \kappa_{3p}(\boldsymbol{X}) \\ \vdots & \vdots & \ddots & \vdots \\ \kappa_{1p}(\boldsymbol{X}) & \kappa_{2p}(\boldsymbol{X}) & \ldots & \kappa_{pp}(\boldsymbol{X}) \end{bmatrix}.$$

Following the derivation above, we replace the scalar β by the functional coefficients

$\beta(\boldsymbol{X})$ and have the bivariate normal as (C.1) with functional mean:

$$\begin{bmatrix} \mu_{jD}(\boldsymbol{X}) \\ \mu_{j'D}(\boldsymbol{X}) \end{bmatrix} = - \begin{bmatrix} \kappa_{jj}(\boldsymbol{X}) & \kappa_{jj'}(\boldsymbol{X}) \\ \kappa_{jj'}(\boldsymbol{X}) & \kappa_{j'j'}(\boldsymbol{X}) \end{bmatrix}^{-1} \begin{bmatrix} \sum_{v \in V \setminus \{j,j'\}} \kappa_{jv}(\boldsymbol{X}) Y_v / d_v \\ \sum_{v \in V \setminus \{j,j'\}} \kappa_{j'v}(\boldsymbol{X}) Y_v / d_v \end{bmatrix}.$$

When $\kappa_{jj'}(\boldsymbol{X}) = 0$, $\beta_{jv}(\boldsymbol{X}) = -\kappa_{jv}(\boldsymbol{X})/\kappa_{jj}(\boldsymbol{X}) = 0$ implying the conditional independence with

$$p(Y_j/d_j | \boldsymbol{Y}_{\mathrm{pa}(j|\boldsymbol{X})}, Y_{j'}, \boldsymbol{D}, \boldsymbol{X}) = p(Y_j/d_j | \boldsymbol{Y}_{V\setminus\{j\}}, \boldsymbol{D}, \boldsymbol{X}).$$

and the corresponding conditional sign independence

$$\begin{aligned} \mathbb{P}(Y_j < 0 | \boldsymbol{Y}_{V\setminus\{j\}}, Y_{j'}, \boldsymbol{X}) &= \mathbb{E}_{D|\boldsymbol{Y}_{V\setminus\{j\}}, Y_{j'}, \boldsymbol{X}}[\mathbb{P}(Y_j < 0 | \boldsymbol{Y}_{V\setminus\{j\}}, Y_{j'}, \boldsymbol{D}, \boldsymbol{X})] \\ &= \mathbb{E}_{D|\boldsymbol{Y}_{V\setminus\{j\}}, Y_{j'}, \boldsymbol{X}}[\Phi(-\kappa_{jj}^{1/2} \mu_{jD}(\boldsymbol{X}))] \\ &= \mathbb{E}_{D|\boldsymbol{Y}_{V\setminus\{j\}}, \boldsymbol{X}}[\Phi(-\kappa_{jj}^{1/2} \mu_{jD}(\boldsymbol{X}))] \\ &= \mathbb{P}(Y_j < 0 | \boldsymbol{Y}_{V\setminus\{j\}}, \boldsymbol{X}) \end{aligned}$$

□

C.2 Posterior Inference

In this Section, we present the posterior inference procedure for rBGR including the MCMC algorithm and the symmetrization. We first provide details of the parameter expansion for the covariate coefficients $\alpha_{j,k,h}$ in Section C.2.1. Section C.2.2 describes the MCMC algorithm including the derivation of Gibbs sampler for the thresholded parameters. In Section C.2.3, we offer the rules used for symmetrizing both the covariate coefficients and the edges.

C.2.1 Parameter Expansion

In rBGR, we assign a spike-and-slab for covariate coefficients $\alpha_{j,k,h}$ with the parameter expansion technique (Geyer, 2011) to improve the mixing of MCMC. Let $\alpha_{j,k,h} = \eta_{j,k,h}\xi_{j,k,h}$. We impose a spike-and-slab prior on $\eta_{j,k,h} \sim N(0, s_{j,k,h})$ with $s_{j,k,h} = \gamma_{j,k,h}\nu_{j,k,h}$, $\nu_{j,k,h} \sim InvGa(a_\nu, b_\nu)$, and $\gamma_{j,k,h} \sim \rho_j\delta_1(\gamma_{j,k,h}) + (1-\rho_j)\delta_{v_0}(\gamma_{j,k,h})$, where v_0 is a small pre-specified hyperparameter. Obviously, the prior results in a binary scenario in terms of $\gamma_{j,k,h}$. When $\gamma_{j,k,h} = v_0$ (spike), $s_{j,k,h}$ is close to zero and results in negligible $\eta_{j,k,h}$ and $\alpha_{j,k,h}$ implying no effect from covariate X_h on the edge between nodes j and k. When $\gamma_{j,k,h} = 1$ (slab), $\alpha_{j,k,h}$ is non-zero with a linear effect from X_h on the edge between nodes j and k. We then assign a beta distribution on $\rho_j \sim Beta(a_\rho, b_\rho)$. For $\xi_{j,k,h}$, we assign a mixture of two normal distribution, $\xi_{j,k,h} \sim N(m_{j,k,h}, 1)$ with $m_{j,k,h} \sim 0.5\delta_1(m_{j,k,h}) + 0.5\delta_{-1}(m_{j,k,h})$. The bimodal mixture distribution encourage $\alpha_{j,k,h}$ to be away from zero, which has been shown to improve selection (Scheipl et al., 2012).

C.2.2 MCMC Algorithm

At each iteration, the MCMC algorithm for rBGR updates parameters that consists of three parts: (i) thresholded parameters of $\alpha_{j,k,h}$ and t_j, (ii) random scales of d_{ij}, and (iii) hyperparameters. The closed-form of the full conditional distribution for thresholded parameters and hyper-parameters are available and enables the Gibbs sampler. On the other hand, we implement the Metropolis–Hastings algorithm for random scales. However, the derivation of the closed-form of the full conditional distribution for thresholded parameters is not straightforward. We briefly describe the general form of the thresholded parameters with Algorithm 3 and refer to Li et al. (2023+) for more details. We then apply Algorithm 3 to the thresholded parameters in rBGR. We summarize the whole MCMC algorithm in Algorithm 4.

General algorithm for the thresholded parameter Consider a random variable θ. Let $f_j(\theta) = a_{1j}\theta^2 + a_{2j}\theta + a_{3j}$ and $g_k(\theta) = b_{1k}\theta^2 + b_{2k}\theta + b_{3k}$. Consider the density of θ to be proportional to

$$\exp\left\{\sum_{j=1}^{J} f_j(\theta)\mathbb{I}(\theta > L_j) + \sum_{k=1}^{K} g_k(\theta)\mathbb{I}(\theta < U_k)\right\},$$

where $L_j, j = 1, \ldots, J$ are lower bounds for f_j and $U_k, k = 1, \ldots, K$ are upper bounds for g_k. We can classified θ into three different mixture distributions based on the values of coefficients in f_j and g_k:

1. If at least one of $\{a_{1j}, \ldots, a_{1J}, b_{1k}, \ldots, b_{1K}\}$ is non-zero, then θ follows a mixture of truncated normal distributions.
2. If $a_{1j} = b_{1k} = 0, \forall j, k$ and at least one of $\{a_{2j}, \ldots, a_{2J}, b_{2k}, \ldots, b_{2K}\}$ is non-zero, then θ follows a mixture of exponential distributions.
3. If $a_{1j} = b_{1k} = a_{2j} = b_{2k} = 0, \forall j, k$ and at least one of $\{a_{3j}, \ldots, a_{3J}, b_{3k}, \ldots, b_{3K}\}$ is non-zero, then θ follows a mixture of uniform distributions.

The key idea is to exhaust the real line into mutually exclusive intervals and update the random variable θ within each interval. We start by dissecting the real line into $J + K + 1$ intervals using the lower or upper bounds as endpoints. For each interval, the truncation mechanism for all functions of f_j and g_k is determined, and we only need to consider the coefficients from non-zero functions of f_j and g_k. With the given coefficients, we can easily derive the distribution within each interval. Finally, we collect distributions from all intervals and normalize the distribution. We implement this idea in Algorithm 3.

From Algorithm 3, it is obvious that the conjugacy of θ can be achieved by assigning priors for different values for $f_j(\theta)$ and $g_k(\theta)$. Specifically, when $a_{1j} = a_{2j} = b_{1k} = b_{2k} = 0$ for all $j = 1, \ldots, J$ and $k = 1, \ldots, K$, we can assign a uniform prior with $c_1 = c_2 = 0$ and $c_3 \neq 0$ resulting in a mixture of uniform distribution. Since the prior

Algorithm 3 Full Condition for θ

Input:

(a) $\{L_j\}_{j=1}^{J}, \{U_k\}_{k=1}^{K}, \{f_j(\theta) = a_{1j}\theta^2 + a_{2j}\theta + a_{3j}\}_{j=1}^{J}$ and $\{g_k(\theta) = g_k(\theta) = b_{1k}\theta^2 + b_{2k}\theta + b_{3k}\}_{k=1}^{K}$.

(b) The prior on θ with the kernel $\exp(c_1\theta^2 + c_2\theta + c_3)$.

Output: The full condition distribution of θ.

Sort the bounds of $\{L_1, \ldots, L_J, U_1, \ldots, U_K\}$ in ascending order with $J + K + 1$ intervals dissected from the real line $\mathbb{R} = \cup_{i=1}^{J+K+1}\mathcal{I}_i$.
for Each interval $\mathcal{I}_i, i = 1...J + K + 1$ **do**
    Initialize $D_i = c_1, E_i = c_2$ and $F_i = c_3$.
    for $j = 1...J, k = 1, \ldots, K$ **do**
        if $\mathcal{I} \subset [L_j, \infty)$ **then**
            Update $D_i = D_i + a_{1j}, E_i = E_i + a_{2j}$ and $F_i = F_i + a_{3j}$.
        if $\mathcal{I} \subset (-\infty, U_k]$ **then**
            Update $D_i = D_i + b_{1j}, E_i = E_i + b_{2j}$ and $F_i = F_i + b_{3j}$.
    if $D_i \neq 0$ **then**
        $\theta \sim N_{\mathcal{I}_i}(-\frac{E_i}{2D_i}, -\frac{1}{D_i})$ for $\theta \in \mathcal{I}_i$.
    if $D_i = 0$ and $E_i \neq 0$ **then**
        $\theta \sim Exp_{\mathcal{I}_i}(E_i)$ for $\theta \in \mathcal{I}_i$.
    if $D_i = E_i = 0$ and $F_i \neq 0$ **then**
        θ follows a uniform distribution on $\mathcal{I}_i$.
Normalize the whole distribution θ, which is proportional to $\sum_{i=1}^{J+K+1} M_i h_i(\theta)$ and M_i is the normalizing constant independent of θ for the distribution $h_i(\theta)$ on interval $\mathcal{I}_i$.

is a special case of the mixture of uniform distribution with only one component, the conjugacy is attainable. Meanwhile, if we assign a normal prior with $c_1 \neq 0$, we obtain a mixture of truncated normal, which grants the conjugacy for θ with normal prior. Given the Algorithm, we then derive the full condition distribution for thresholded parameters from rBGR with the Gibbs sampler.

Covariate coefficients. We first derive the full condition for $\eta_{j,k,h}$ and $\xi_{j,k,h}$. We only show the full condition for $\eta_{j,k,h}$ since both are normally distributed, and the distribution of $\xi_{j,k,h}$ can be analogously derived.

$$
\begin{aligned}
p(\eta_{j,k,h} \mid \boldsymbol{Y}, \boldsymbol{X}, \Theta_{-\eta_{j,k,h}}) \propto \exp \Bigg[& \sum_{i: X_{ih} \geq 0} g_i(\eta_{j,k,h}) \left\{ \mathbb{I}(\eta_{j,k,h} \geq T_{i1}) + \mathbb{I}(\eta_{j,k,h} < T_{i2}) \right\} \\
& + \sum_{i: X_{ih} < 0} g_i(\eta_{j,k,h}) \left\{ \mathbb{I}(\eta_{j,k,h} \geq T_{i2}) + \mathbb{I}(\eta_{j,k,h} < T_{i1}) \right\} \Bigg] g_i(\eta_{j,k,h}) \\
& = a_{1i} \eta_{j,k,h}^2 + a_{2i} \eta_{j,k,h} \qquad \text{(C.4)} \\
a_{1i} & = -\frac{X_{ih}^2 Y_{ik}^2 \xi_{j,k,h}^2}{2\sigma_j^2 d_{ik}^2} - \frac{1}{2 s_{j,k,h}} \\
a_{2i} & = -\frac{X_{ih} \xi_{j,k,h}}{\sigma_j^2} \left[\frac{Y_{ik}^2}{d_{ik}^2} \sum_{l \neq h}^{q} \alpha_{j,k,l} X_{il} + \frac{Y_{ik}}{d_{ik}} \left(-\frac{X_{ij}}{d_{ij}} + \sum_{m \neq k}^{p} \beta_m(\boldsymbol{X}_i) Y_{im} \right) \right] \\
T_{i1} & = \frac{t_j - \sum_{l \neq h}^{q} \alpha_{j,k,l} X_{il}}{\xi_{j,k,h} X_{ih}}; \; T_{i2} = \frac{-t_j - \sum_{l \neq h}^{q} \alpha_{jl} X_{il}}{\xi_{j,k,h} X_{ih}}
\end{aligned}
$$

where $g_i(\eta_{j,k,h}) = a_{1i}\eta_{j,k,h}^2 + a_{2i}\eta_{j,k,h}$ is a quadratic function of $\eta_{j,k,h}$ and T_{i1} and T_{i2} are independent of $\eta_{j,k,h}$. Therefore, the full condition distribution of $\eta_{j,k,h}$ belongs to the first category with a mixture of normal distribution. When we assign a normal prior on $\eta_{j,k,h}$ and $\xi_{j,k,h}$, we obtain the conjugacy with the Gibbs sampler shown in Algorithm 3.

Threshold parameter. The same idea can be used on the threshold parameter t_j. Specifically, the full condition of the threshold parameter is

$$
p(t_j \mid \boldsymbol{Y}, \boldsymbol{X}, \Theta_{-t_j}) \propto \exp \left\{ -\sum_{i=1}^{n} \sum_{k \neq j}^{p} \mathbb{I}\left(t_j < |\theta_{j,k}(\boldsymbol{X}_i)|\right) \frac{P_{ik} + Q_{ik}}{2\sigma_j^2} \right\} \frac{\mathbb{I}(0 \leq t_j \leq t_{\max})}{t_{\max}} \qquad \text{(C.5)}
$$

$$
Q_{ik} = 2\theta_{j,k}(\boldsymbol{X}_i) \frac{Y_{ik}}{d_{ik}} \sum_{k' \neq k} \theta_{j,k'}(\boldsymbol{X}_i) \frac{Y_{ik'}}{d_{ik'}} \mathbb{I}\left(t_j < |\theta_{j,k'}(\boldsymbol{X}_i)|\right) \qquad \text{(C.6)}
$$

$$
P_{ik} = \theta_{j,k}^2(\boldsymbol{X}_i) \frac{Y_{ik}^2}{d_{ik}^2} - 2\theta_{j,k}(\boldsymbol{X}_i) \frac{Y_{ik}}{d_{ik}} \frac{Y_{ij}}{d_{ij}}. \qquad \text{(C.7)}
$$

Given all $\theta_{j,k}(\boldsymbol{X}_i)$, we claim that both P_{ik} and Q_{ik} are constant with respect to t_j. Obviously, P_{ik} does not depend on t_j. For any given interval, we also find that Q_{ik} is independent of t_j. Therefore, Q_{ik} is also independent of t_j, and the full condition for t_j falls into the third category with the mixture of the uniform distribution.

Now, we can present the whole MCMC algorithm as follows:

Algorithm 4 MCMC algorithm for rBGR

(a) Update $\eta_{j,k,h}$ and $\xi_{j,k,h}$ by Algorithm 3 with (C.4);

- Rescale $\eta_{j,k,h}$ and $\xi_{j,k,h}$ with $\eta_{j,k,h} \to \eta_{j,k,h}|\xi_{j,k,h}|$ and $\xi_{j,k,h} \to \eta_{j,k,h}/|\xi_{j,k,h}|$.

(b) Update t_j by Algorithm 3 with (C.5);

(c) Update $m_{j,k,h}$ by Gibbs: $p(m_{j,k,h} = 1 \mid \xi_{j,k,h}) = \frac{1}{1+\exp(-2\xi_{j,k,h})}$;

(d) Update $\gamma_{j,k,h}$ by Gibbs: $\frac{p(\gamma_{j,k,h}=1|\eta_{j,k,h},\nu_{j,k,h},\rho_{j,k,h})}{p(\gamma_{j,k,h}=v_0|\eta_{j,k,h},\nu_{j,k,h},\rho_j)} = \frac{\sqrt{v_0}\rho_j}{1-\rho_j}\exp(\frac{-v_0\eta^2_{j,k,h}}{2v_0\nu_{j,k,h}})$;

(e) Update $\nu_{j,k,h}$ by Gibbs: $p(\nu_{j,k,h} \mid \eta_{j,k,h}, \gamma_{j,k,h})\text{InvGa}(a_\nu + 1/2, b_\nu + \frac{\eta^2_{j,k,h}}{2\gamma_{j,k,h}})$;

(f) Update ρ_j by Gibbs: $p(\rho_j \mid \gamma_{j,k,h}) = \text{Beta}(a_\rho + \sum_{k,h}\mathbb{I}(\gamma_{j,k,h} = 1), b_\rho + \sum_{k,h}\mathbb{I}(\gamma_{j,k,h} = v_0))$;

(g) Update d_{ij} by MH algorithm with a proposal as prior.

(h) Update π_j by Gibbs: $p(\pi_j \mid D_{ij}) = \text{Beta}(a_\pi + \sum_i \mathbb{I}(D_{ij=1}), b_\pi + \sum_i \mathbb{I}(D_{ij\neq 1}))$;

C.2.3 Details of Covariate and Edge Selection

The estimated coefficients from rBGR of (4.3) do not guarantee the symmetry required in undirected graph. Moreover, due to the introduction of random factors with the CSI characterization, we only focus on the sign of the edge. In this section, we describe algorithms to symmetrize the estimated covariate coefficients $\widehat{\alpha}_{j,k,h}$ and the sign of graph edges of $\widehat{\beta}_{j,k}(\boldsymbol{X}_i)$. Denote $\mathbb{P}^{\alpha}_{j,k,h} = \mathbb{P}(\widehat{\alpha}_{j,k,h} \neq 0)$ as the posterior inclusion probability (PIP) of $\widehat{\alpha}_{j,k,h}$ and let $\widetilde{\alpha}_{j,k,h}$ be the covariate coefficients for the undirected graph between node j and k for covariate h. We formulate the symmetrization rules via choosing the direction with a lower PIP:

$$\widetilde{\alpha}_{j,k,h} = \widehat{\alpha}_{j,k,h}\mathbb{I}(\mathbb{P}^{\alpha}_{k,j,h} > \mathbb{P}^{\alpha}_{j,k,h}) + \widehat{\alpha}_{k,j,h}\mathbb{I}(\mathbb{P}^{\alpha}_{j,k,h} \geq \mathbb{P}^{\alpha}_{k,j,h}). \tag{C.8}$$

Given a cutoff c_0, Equation C.8 requires both directions to have PIPs bigger than c_0 implying a network with less edges. Another possible symmetrization is

$$\widetilde{\alpha}_{j,k,h} = \widehat{\alpha}_{j,k,h}\mathbb{I}(\mathbb{P}^{\alpha}_{j,k,h} > \mathbb{P}^{\alpha}_{k,j,h}) + \widehat{\alpha}_{k,j,h}\mathbb{I}(\mathbb{P}^{\alpha}_{k,j,h} \geq \mathbb{P}^{\alpha}_{j,k,h}). \tag{C.9}$$

Obviously, Equation (C.9) is less conservative and requires at least one PIP bigger than c_0. Similar symmetrization rules can be seen in Zhang and Li (2022) if we replace the PIP with the absolute value of coefficients. For rules with absolute value of coefficients, both rules are asymptotically equivalent (Meinshausen and Bühlmann, 2006; Zhang and Li, 2022), but the rule of (C.8) performs better given finite samples (Meinshausen and Bühlmann, 2006). We use the rule (C.8) for the rest paper.

For the edge $\widehat{\beta}_{j,k}(\boldsymbol{X}_i)$, we first calculate the estimated linear function of $\widetilde{\theta}_{j,k}(\boldsymbol{X}_i)$ and symmetrize the edge posterior probability (ePP) of the sign of $\widehat{\beta}_{j,k}(\boldsymbol{X}_i)$. Specifically, $\widetilde{\theta}_{j,k}(\boldsymbol{X}_i) = \sum_{h=1}^{q} \widetilde{\alpha}_{j,k,h} X_{ih}$ and $\widetilde{\theta}_{j,k}(\boldsymbol{X}_i)$ is symmetric since $\widetilde{\alpha}_{j,k,h}$ from (C.8) is symmetric. Denote $\mathbb{P}^{\beta}_{j,k}(\boldsymbol{X}_i) = \mathbb{P}(\widehat{\beta}_{j,k}(\boldsymbol{X}_i) \neq 0)$ as the ePP of a directed edge from node k to j. In this paper, we symmetrize the sign of the edge by taking the maximum of the ePP from two directions through

$$\widetilde{\mathbb{P}}^{\beta}_{j,k}(\boldsymbol{X}_i) = \max(\mathbb{P}^{\beta}_{j,k}(\boldsymbol{X}_i), \mathbb{P}^{\beta}_{k,j}(\boldsymbol{X}_i)), \tag{C.10}$$

where $\widetilde{\mathbb{P}}^{\beta}_{j,k}(\boldsymbol{X}_i)$ is ePP of an undirected edge between node j and k. Given a threshold c_1, we then call an undirected edge if $\widetilde{\mathbb{P}}^{\beta}_{j,k}(\boldsymbol{X}_i) > c_1$. Alternatively, we can take the minimum as

$$\widetilde{\mathbb{P}}^{\beta}_{j,k}(\boldsymbol{X}_i) = \min(\mathbb{P}^{\beta}_{j,k}(\boldsymbol{X}_i), \mathbb{P}^{\beta}_{k,j}(\boldsymbol{X}_i)). \tag{C.11}$$

Clearly, (C.11) is more conservative and needs both $\mathbb{P}^{\beta}_{j,k}(\boldsymbol{X}_i)$ and $\mathbb{P}^{\beta}_{j,k}(\boldsymbol{X}_i)$ bigger than c_1 to call an edge, while (C.10) requires only one of the posterior probability bigger

than c_1.

Once we symmetrize the ePP, we can decide the sign for edges given that the ePP is bigger than the cutoff $\widetilde{\mathbb{P}}^{\beta}_{j,k}(\boldsymbol{X}_i) > c_1$. Without loss of generality, assume that we chose a specific direction as undirected edge with $\widetilde{\mathbb{P}}^{\beta}_{j,k}(\boldsymbol{X}_i) = \mathbb{P}^{\beta}_{j,k}(\boldsymbol{X}_i)$. We estimate the sign of the edge by comparing the posterior probability of positive and negative for the chosen direction. Specifically, given the direction of $\widetilde{\mathbb{P}}^{\beta}_{j,k}(\boldsymbol{X}_i) = \mathbb{P}^{\beta}_{j,k}(\boldsymbol{X}_i)$, we estimate the sign of the edge by the following rule:

$$\text{sign}(\beta_{j,k}(\boldsymbol{X}_i)) = \begin{cases} 1 & \text{if } \mathbb{P}(\widehat{\beta}_{j,k}(\boldsymbol{X}_i) > 0) > \mathbb{P}(\widehat{\beta}_{j,k}(\boldsymbol{X}_i) < 0) \\ -1 & \text{if } \mathbb{P}(\widehat{\beta}_{j,k}(\boldsymbol{X}_i) > 0) \leq \mathbb{P}(\widehat{\beta}_{j,k}(\boldsymbol{X}_i) < 0) \end{cases} \tag{C.12}$$

Remark 5. Both rules of (C.10) and (C.11) leave the value of $\widehat{\beta}_{j,k}(\boldsymbol{X}_i)$ to be asymmetric. One might symmetrize edges through symmetrizing both linear function and the threshold parameter. However, matching the threshold parameter results in a common $\widehat{t}_j = \widehat{t}$ for all $j = 1, \ldots, p$, which imposes strict constraints. For this paper, we do not require the value of $\widehat{\beta}_{j,k}(\boldsymbol{X}_i)$ from two directions equal and only need to ensure that the sign of edges from two directions agrees.

C.3 Additional Simulation Results of rBGR

C.3.1 Details of Data Generating Mechanism

We generate the data from an underlying multivariate normal distribution with precision matrix representing the undirected graph and transform the latent normal data with random scale to obtain the observed non-normal data. Specifically, we first generate the covariates $\boldsymbol{X}_i \overset{iid}{\sim} U(-1, 1)$ and obtain the latent data from a multivariate normal distribution. By multiplying the latent data by random scales, we acquire the observed non-normal data. We set the sample size and the dimension of $\boldsymbol{Y}_i$ and $\boldsymbol{X}_i$

as $(n, p, q) = (250, 50, 3)$, and generate the latent data from the following procedures:

$$\boldsymbol{Y}_i^* = \left[Y_{i1}^*, \ldots, Y_{ip}^*\right]^\mathsf{T} \overset{iid}{\sim} \boldsymbol{N}_p(\boldsymbol{0}, \boldsymbol{\Omega}^{-1}(\boldsymbol{X}_i)), i = 1, \ldots, n$$

where $\boldsymbol{\Omega}^{-1}(\boldsymbol{X}_i)$ is the true precision matrix. For true precision matrix, we assign unit diagonal elements and randomly pick 2% of the off-diagonal to be non-zero. Given a threshold parameter t^0, each non-zero precision depends linearly on the covariates and is truncated to zero if the absolute value is smaller than the threshold parameter t^0. Specifically, we set the non-zero precision as $\omega^{j,k}(\boldsymbol{X}_i) = r^{j,k}(\boldsymbol{X}_i)\mathbb{I}(|r^{j,k}(\boldsymbol{X}_i)| > t^0)$ and $r^{j,k}(\boldsymbol{X}_i) = \sum_{h=1}^{q} X_{ih}\nu_{j,k,h}$, where $\nu_{j,k,h} \sim U(-0.5, -0.35) \cup U(0.35, 0.5)$. We set $t^0 = 0.15$ to filter around half of the non-zero off-diagonal elements. The final precision matrix might not be positive semi-definite, and we repeat the whole process till the precision matrix is positive semi-definite. We obtain the random scales from a mixture distribution of the point mass at one and a inverse gamma distribution with shape and scale parameters $d_{ij}^2 \overset{iid}{\sim} (1-\pi)\delta_1 + \pi InvGa(a_{d_j}, b_{d_j})$. We assign three different levels of π representing three different levels of non-normal contamination: $\pi \in \{0, 0.5, 0.8\}$. Given the latent data from the multivariate normal distribution, we multiply the random scales, d_{ij}, to generate the observed data of $[Y_{i1}, \ldots, Y_{ip}] = [Y_{i1}^* d_{i1}, \ldots, Y_{ip}^* d_{ip}]$.

C.3.2 Convergence Diagnostics of MCMC

One important issue for the Bayesian method is to ensure that the MCMC converged to draw the samples from the target posterior distribution. We investigate the convergence of the MCMC through the Geweke statistics (Geweke, 1992). Specifically, we check the Geweke statistics of the covariate coefficients $\alpha_{j,k,h}$. After the burn-in period, we take the first and the last 20% of the posterior samples and calculate the Geweke statistics. We require p-values for all $\alpha_{j,k,h}$ to be insignificant after the Bonferroni correction (Armstrong, 2014) to ensure the convergence of the

algorithm.

C.3.3 Simulation Results of Different cut-off of c_0 and c_1 Controlling for False Discovery Rates

Another possible way to decide the cut-off of c_0 and c_1 is by controlling the false discovery rate (FDR) (Storey and Tibshirani, 2003) α. Consider a sorted vector Q of dimension N in decreasing order with each element as a probability. Denote $Q_{(k)}$ as the k-th largest element in Q. We first calculate $\xi = \max\{K : K^{-1}\sum_{k=1}^{K}(1 - Q_{(k)}) < \alpha\}$ and set the cut-off as $c^{\alpha} = Q_{(\xi)}$. In this Section, we fixed the FDR at $\alpha = 0.1$ and obtain the cut-off for the PIP from $\alpha_{j,k,h}$ and the ePP from $\beta_{jk}(\boldsymbol{X}_i)$.

Panel (A) of Figure C.1 show the results for covariate selection when we use the cut-off controlling for the false discovery rate. Comparing to the cut-off at $c_0 = 0.5$ used in Main Paper of Chapter IV, we observe that rBRG and BGR generate a higher TPR and TNR but a lower MCC for covariate selection. Specifically, rBGR outperforms both BGR and RegGMM in TPR across different non-normality levels. For TNR, rBGR performs slightly worse than BGR and RegGMM for across all non-normality levels, but the disadvantage of rBGR decreases when the non-normality level increases. Moreover, all three methods select correct covariates and edges ($>$ 90%) with small difference ($<$ 10%) in terms of TNR. We observe that rBGR achieves a lower MCC comparing to BGR and RegGMM when the data is normally distributed. However, rBGR surpasses BGR and RegGMM in terms of MCC when the level of non-normality increases. Similar to Main Paper of Chapter IV, modeling the non-normality from random scales in rBGR is favored compared to models without random scales in terms of covariate selection.

We show the graph recovery for the edge selection using the cut-off controlling for the false discovery rate in Panel (B) of Figure C.1. We observe that using the cut-off controlling for the false discovery rate results in a higher TPR, but lower

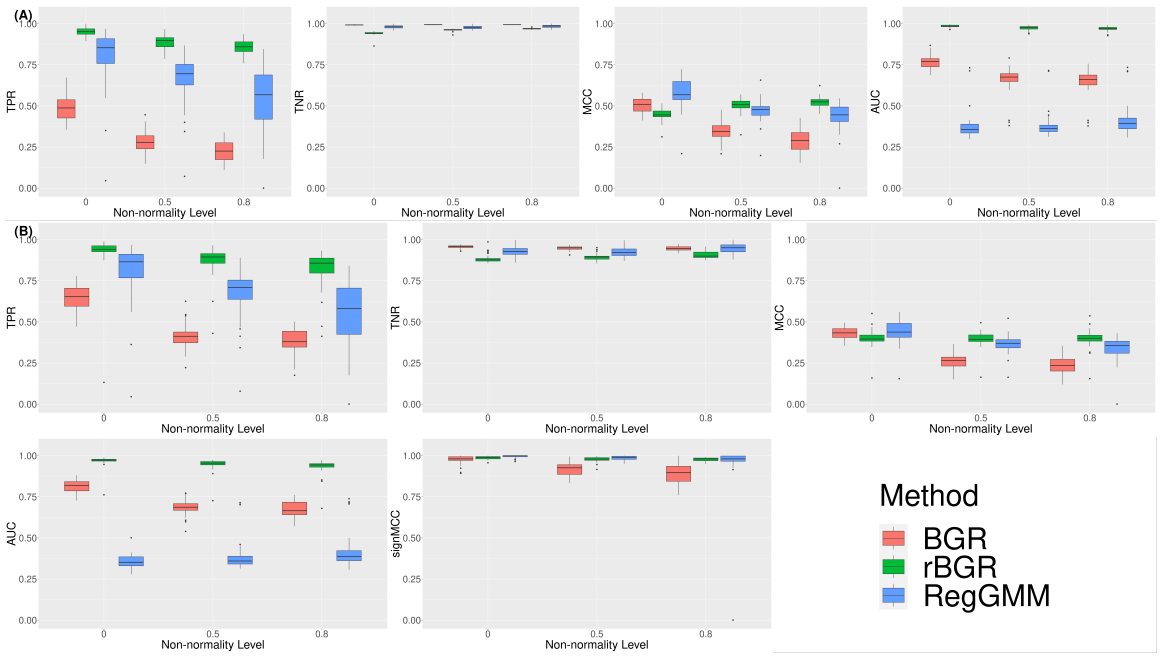


Figure C.1: Graph recovery for BGR (red), rBGR (green) and RegGMM (blue) under different levels of non-normality in terms of (A) covariates selection (top row) and (B) edge selection (bottom two rows). Panel (A) measures the covariate selection through four metrics (from left to right: TPR, TNR, MCC and AUC) are measured under three different levels of non-normality. Panel (B) demonstrates the edge selection by four criteria (from upper left to lower right: TPR, TNR, MCC, AUC) and the sign consistency by sign-MCC (lower left) for non-zero edges. All values for TPR, TNR and MCC are measured at a cut-off controlling for false discovery rate.

TNR and MCC for rBGR. Specifically, rBGR has the best performance in terms of TPR comparing to other benchmarks of BGR and RegGMM under all levels of non-normality, and the advantage of rBGR becomes more prominent as the non-normality increases. For MCC, rBGR is slightly inferior than BGR and RegGMM under the normal distribution. However, rBGR is favored when the non-normality level increases. Both TNR and sign-MCC show excellent edge selection performance ($> 90\%$) for all three methods, with minimal differences ($< 10\%$) across the three non-normality levels. In summary, modeling the non-normality through random scales in rBGR result in equivalent (under normal distribution) or better performances in all metric for edge selection compared to the other methods without accounting for non-normality.

C.4 Additional Results for Proteomic Networks under Immunogenic Heterogeneity

C.4.1 Pre-processing Procedures and Convergence Diagnostics

For proteomics data, we first removed phosphorylation proteins and focus on proteins in 12 important cancer-related pathways (apoptosis, breast hormone signaling, breast reactive, cell cycle, core reactive, DNA damage response, EMT, PI3K/AKT, RAS/MAPK, RTK, TSC/mTOR and hormone receptor) (Ha et al., 2018). After centering proteomic data, we obtain 41 proteins from both OV and LUAD with 241 patients and 360 patients for OV and LUAD, respectively. For covariates, we obtained expression data from immune cells and treated the mRNA expression as the immune cell abundance. We averaged mRNA expression for the genes listed for seven immune cells (B cell, T cell, macrophages, monocytes, neutrophils, natural killer cells and plasma cell) and three pathways (proliferation, interferon and translation) (Nirmal et al., 2018). We further took the log transformation and standardized on the averaged expression data. For this analysis, we chose T cells and two important components of myeloid-derived suppressor cells (MDSC), monocytes and neutrophils, for both OV and LUAD for two reasons. First, both T cells and MDSC are essential in both OV (Luo et al., 2021; Yang et al., 2020) and LUAD (Spella and Stathopoulos, 2021; Wang et al., 2022). The existing biology also suggests the importance of macrophage and natural killer cells (NK cells), but since we observed a high correlation among T cells, macrophages (OV: 0.71 and LUAD: 0.80) and NK cells (OV: 0.77 and LUAD: 0.49) we did not include the marcophages and NK cells in this analysis. We ran rBGR on OV and LUAD with $20,000$ iterations and discarded first $19,000$ iterations. We adapted the symmetrization of (C.8) for covariate coefficients and (C.10) for edges. We examine the convergence of the algorithm through both the Geweke statistics and the likelihood trace plot shown in Figure C.2. Specifically, we

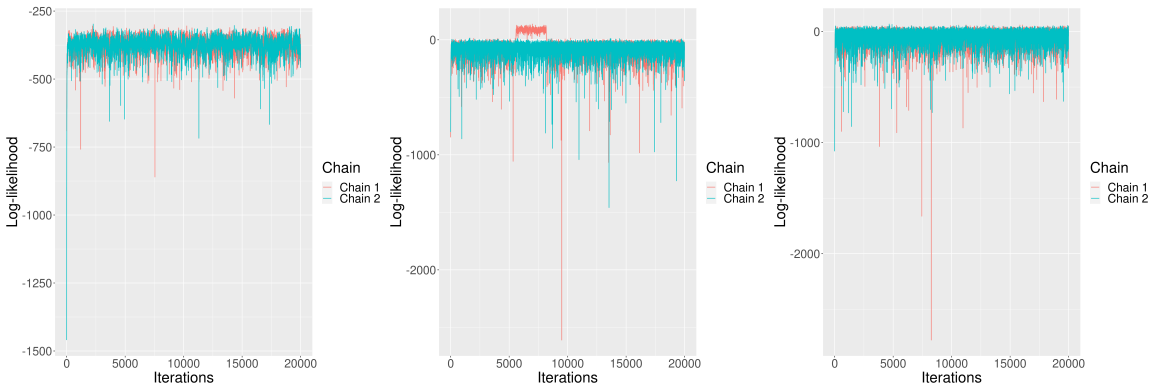


Figure C.2: Convergence diagnostics for using rBGR algorithm on lung cancer. Three randomly chosen nodes are initiated with two different chains. Both chains converge to a similar level of log-likelihood after the burn-in period of the first 19,000 iterations.

ensure the convergence of the algorithm by requiring the p-values of $\alpha_{j,k,h}$ from the Geweke statistics are all insignificant after Bonferroni correction (Armstrong, 2014). In Figure C.2, we randomly pick three proteins and run the algorithm with two chains of different initialization. Both chains converge to a similar level of log-likelihood after the burn-in period of the first 19,000 iterations, indicating the convergence of the algorithm.

C.4.2 Patient-Specific Networks for Ovarian Cancer

In this Section, we present the patient-specific network for ovarian cancer (see Figure C.3). Similar to the Main Paper of Chapter IV, we vary the abundance of one immune component with the rest two components fixed and focus on the edges that change the sign when the immune component abundance increase. In OV, we observe that only the edge of E-Cadherin-Fibronectin changes the sign the neutrophils abundance increases. Specifically, the this edge is positively correlated to the neutrophil abundance. When neutrophil abundance is higher, E-Cadherin-Fibronectin is positive; vice-versa, E-Cadherin-Fibronectin is negative when neutrophil is scarce. Recently, neutrophils have been shown to induce the expression of fibronectin through the epithelial–mesenchymal transition pathway, and the same pathway also represses

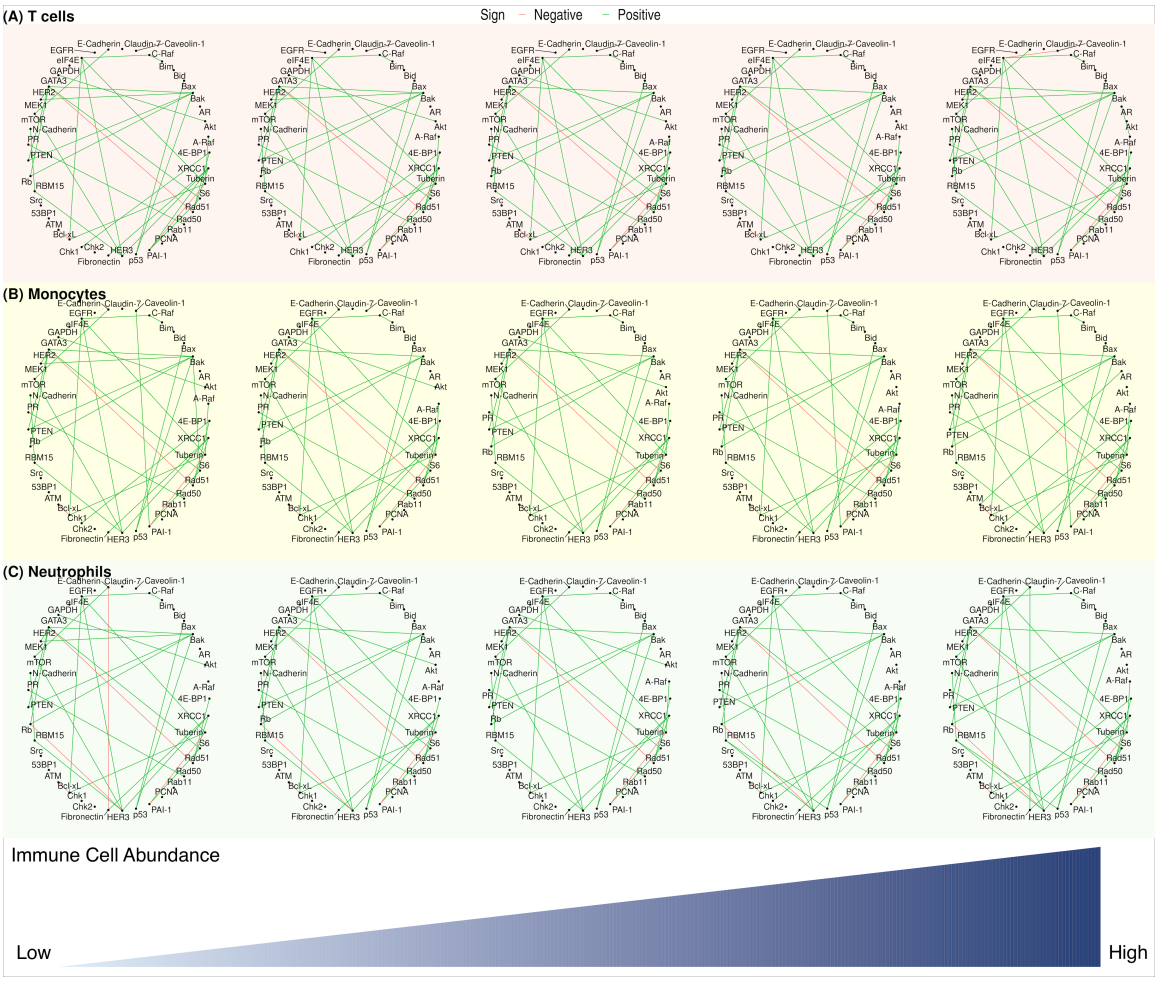


Figure C.**3:** Networks of OV under five different percentiles immune component of (A) T cells, (B) monocytes and (C) neutrophils with the rest two components fixed at mean zero. The estimated network for varying immune components are shown from the left to right for 5, 25, 50 ,75, and 95-th percentiles. Edges are identified with signs (green: positive and red: negative) when the ePPs are bigger than $c_1 = 0.5$.

the expression of E-Cadherin, resulting in the tumor growth (Martins-Cardoso et al., 2020).

BIBLIOGRAPHY

Abdolahi, S., Z. Ghazvinian, S. Muhammadnejad, M. Saleh, H. Asadzadeh Aghdaei, and K. Baghaei (2022), Patient-derived xenograft (PDX) models, applications and challenges in cancer research, *J Transl Med*, *20*(1), 206.

Agrawal, R., U. Roy, and C. Uhler (2020), Covariance Matrix Estimation under Total Positivity for Portfolio Selection*, *Journal of Financial Econometrics*, *20*(2), 367–389, doi:10.1093/jjfinec/nbaa018.

Airoldi, E. M. (2007), Getting started in probabilistic graphical models, *PLoS Comput Biol*, *3*(12), e252.

Akbani, R., et al. (2014), A pan-cancer proteomic perspective on The Cancer Genome Atlas, *Nat Commun*, *5*, 3887.

Aldous, D. (1996), Probability distributions on cladograms, in *Random Discrete Structures*, edited by D. Aldous and R. Pemantle, pp. 1–18, Springer New York, New York, NY.

Antonsson, B., F. Conti, A. Ciavatta, S. Montessuit, S. Lewis, I. Martinou, et al. (1997), Inhibition of Bax channel-forming activity by Bcl-2, *Science*, *277*(5324), 370–372.

Armstrong, R. A. (2014), When to use the Bonferroni correction, *Ophthalmic Physiol Opt*, *34*(5), 502–508.

Babic, S., L. Gelbgras, M. Hallin, and C. Ley (2021a), Optimal tests for elliptical symmetry: Specified and unspecified location, *Bernoulli*, *27*(4), 2189 – 2216, doi:10.3150/20-BEJ1305.

Babic, S., C. Ley, and M. Palangetic (2021b), The r journal: Elliptical symmetry tests in r, *The R Journal*, *13*, 661–672, doi:10.32614/RJ-2021-078, https://doi.org/10.32614/RJ-2021-078.

Baladandayuthapani, V., R. Talluri, Y. Ji, K. R. Coombes, Y. Lu, B. T. Hennessy, M. A. Davies, and B. K. Mallick (2014), Bayesian Sparse Graphical Models for Classification with Application to Protein Expression Data, *Ann Appl Stat*, *8*(3), 1443–1468.

Balko, J. M., et al. (2012), The receptor tyrosine kinase ErbB3 maintains the balance between luminal and basal breast epithelium, *Proc Natl Acad Sci U S A*, *109*(1), 221–226.

Bayat Mokhtari, R., T. S. Homayouni, N. Baluch, E. Morgatskaya, S. Kumar, B. Das, and H. Yeger (2017), Combination therapy in combating cancer, *Oncotarget*, *8*(23), 38,022–38,043.

Beaumont, M. A., W. Zhang, and D. J. Balding (2002), Approximate Bayesian computation in population genetics, *Genetics*, *162*(4), 2025–2035.

Berestycki, J., N. Berestycki, and J. Schweinsberg (2007), Beta-coalescents and continuous stable random trees, *Ann. Probab.*, *35*(5), 1835–1887, doi:10.1214/009117906000001114.

Bertorelle, G., A. Benazzo, and S. Mona (2010), ABC as a flexible framework to estimate demography over space and time: some cons, many pros, *Mol Ecol*, *19*(13), 2609–2625.

Bertotti, A., et al. (2011), A molecularly annotated platform of patient-derived xenografts ("xenopatients") identifies HER2 as an effective therapeutic target in cetuximab-resistant colorectal cancer, *Cancer Discov*, *1*(6), 508–523.

Bhadra, A., A. Rao, and V. Baladandayuthapani (2018), Inferring network structure in non-normal and mixed discrete-continuous genomic data, *Biometrics*, *74*(1), 185–195.

Bhateja, P., M. Chiu, G. Wildey, M. B. Lipka, P. Fu, M. C. L. Yang, et al. (2019), Retinoblastoma mutation predicts poor outcomes in advanced non small cell lung cancer, *Cancer Med*, *8*(4), 1459–1466.

Bhimani, J., K. Ball, and J. Stebbing (2020), Patient-derived xenograft models-the future of personalised cancer treatment, *Br J Cancer*, *122*(5), 601–602.

Biau, G., F. Cérou, and A. Guyader (2015), New insights into approximate Bayesian computation, *Ann. Inst. H. Poincaré Probab. Statist.*, *51*(1), 376–403, doi:10.1214/13-AIHP590.

Billera, L. J., S. P. Holmes, and K. Vogtmann (2001), Geometry of the space of phylogenetic trees, *Advances in Applied Mathematics*, *27*(4), 733–767, doi:10.1006/aama.2001.0759.

Blum, M. G. (2010), Approximate Bayesian computation: A nonparametric perspective, *Journal of the American Statistical Association*, *105*(491), 1178–1187, doi:10.1198/jasa.2010.tm09448.

Blum, M. G. B., M. A. Nunes, D. Prangle, and S. A. Sisson (2013), A Comparative Review of Dimension Reduction Methods in Approximate Bayesian Computation, *Statistical Science*, *28*(2), 189 – 208, doi:10.1214/12-STS406.

Bohnacker, T., et al. (2017), Deconvolution of Buparlisib's mechanism of action defines specific PI3K and tubulin inhibitors for therapeutic intervention, *Nat Commun*, *8*, 14,683.

Bonelli, M. A., et al. (2017), Combined inhibition of CDK4/6 and PI3K/AKT/mTOR pathways induces a synergistic anti-tumor effect in malignant pleural mesothelioma cells, *Neoplasia*, *19*(8), 637–648.

Brandts, J., and A. Cihangir (2016), Geometric aspects of the symmetric inverse M-matrix problem, *Linear Algebra and Its Applications*, *506*, 33–81, doi:10.1016/j.laa.2016.05.015.

Bravo, H. C., S. Wright, K. H. Eng, S. Keles, and G. Wahba (2009), Estimating tree-structured covariance matrices via mixed-integer programming, *J Mach Learn Res*, *5*, 41–48.

Bridson, M. R., and A. Haefilger (1999), *Metric Spaces of Non-Positive Curvature*, Springer Berlin, Heidelberg.

Burns, P. (2011), *The R Inferno*, Lulu.com.

Cardona, G., A. Mir, F. Rosselló, L. Rotger, and D. Sánchez (2013), Cophenetic metrics for phylogenetic trees, after Sokal and Rohlf, *BMC Bioinformatics*, *14*, 3.

Casella, G., and R. Berger (2001), *Statistical Inference*, Duxbury Resource Center.

Chakraborty, M., V. Baladandayuthapani, A. Bhadra, and M. J. Ha (2021), Bayesian robust learning in chain graph models for integrative pharmacogenomics, doi:10.48550/ARXIV.2111.11529.

Cheng, S. S., G. J. Yang, W. Wang, C. H. Leung, and D. L. Ma (2020), The design and development of covalent protein-protein interaction inhibitors for cancer treatment, *J Hematol Oncol*, *13*(1), 26.

Chierchia, G., and B. Perret (2020), Ultrametric fitting by gradient descent, *Journal of Statistical Mechanics: Theory and Experiment*, *2020*(12), doi:10.1088/1742-5468/abc62d.

Chung, H. C., I. Gaynanova, and Y. Ni (2022), Phylogenetically informed Bayesian truncated copula graphical models for microbial association networks, *The Annals of Applied Statistics*, *16*(4), 2437–2457.

Clohessy, J. G., and P. P. Pandolfi (2015), Mouse hospital and co-clinical trial project–from bench to bedside, *Nat Rev Clin Oncol*, *12*(8), 491–498.

Cohen, H., R. Ben-Hamo, M. Gidoni, I. Yitzhaki, R. Kozol, A. Zilberberg, and S. Efroni (2014), Shift in GATA3 functions, and GATA3 mutations, control progression and clinical presentation in breast cancer, *Breast Cancer Res*, *16*(6), 464.

Conciatori, F., C. Bazzichetto, I. Falcone, L. Ciuffreda, G. Ferretti, S. Vari, et al. (2020), PTEN Function at the Interface between Cancer and Tumor Microenvironment: Implications for Response to Immunotherapy, *Int J Mol Sci*, *21*(15).

Cook, S. R., A. Gelman, and D. B. Rubin (2006), Validation of software for Bayesian models using posterior quantiles, *Journal of Computational and Graphical Statistics*, *15*(3), 675–692, doi:10.1198/106186006X136976.

Dagogo-Jack, I., and A. T. Shaw (2018), Tumour heterogeneity and resistance to cancer therapies, *Nat Rev Clin Oncol*, *15*(2), 81–94.

Danaher, P., P. Wang, and D. M. Witten (2014), The joint graphical lasso for inverse covariance estimation across multiple classes, *J R Stat Soc Series B Stat Methodol*, *76*(2), 373–397.

Dellacherie, C., S. Martinez, and J. San Martin (2014), *Inverse M-Matrices and Ultrametric Matrices*, Springer Cham.

Dobra, A., and A. Lenkoski (2011), Copula Gaussian graphical models and their application to modeling functional disability data, *The Annals of Applied Statistics*, *5*(2A), 969 – 993, doi:10.1214/10-AOAS397.

Dobrolecki, L. E., et al. (2016), Patient-derived xenograft (PDX) models in basic and translational breast cancer research, *Cancer Metastasis Rev*, *35*(4), 547–573.

Doornik, J. A., and H. Hansen (2008), An omnibus test for univariate and multivariate normality*, *Oxford Bulletin of Economics and Statistics*, *70*(s1), 927–939, doi: https://doi.org/10.1111/j.1468-0084.2008.00537.x.

Dummer, R., et al. (2018a), Encorafenib plus binimetinib versus vemurafenib or encorafenib in patients with BRAF-mutant melanoma (COLUMBUS): a multicentre, open-label, randomised phase 3 trial, *Lancet Oncol*, *19*(5), 603–615.

Dummer, R., et al. (2018b), Overall survival in patients with BRAF-mutant melanoma receiving encorafenib plus binimetinib versus vemurafenib or encorafenib (COLUMBUS): a multicentre, open-label, randomised, phase 3 trial, *Lancet Oncol*, *19*(10), 1315–1327.

Evrard, Y. A., A. Srivastava, J. Randjelovic, J. H. Doroshow, D. A. Dean, J. S. Morris, and J. H. Chuang (2020), Systematic Establishment of Robustness and Standards in Patient-Derived Xenograft Experiments and Analysis, *Cancer Res*, *80*(11), 2286–2297.

Fallat, S., S. Lauritzen, K. Sadeghi, C. Uhler, N. Wermuth, and P. Zwiernik (2017), Total positivity in Markov structures, *Annals of Statistics*, *45*(3), 1152–1184, doi: 10.1214/16-AOS1478.

Ferlay, J., M. Ervik, F. Lam, M. Colombet, L. Mery, M. Piñeros, A. Znaor, I. Soerjomataram, and F. Bray (2020), *Global Cancer Observatory: Cancer Today*, Lyon, France: International Agency for Research on Cancer, available from: `https://gco.iarc.fr/today`, accessed 05.28.2021.

Finegold, M., and M. Drton (2011), Robust graphical modeling of gene networks using classical and alternative t-distributions, *The Annals of Applied Statistics*, *5*(2A), 1057 – 1080, doi:10.1214/10-AOAS410.

Finegold, M., and M. Drton (2014), Robust Bayesian Graphical Modeling Using Dirichlet t-Distributions, *Bayesian Analysis*, *9*(3), 521 – 550, doi:10.1214/13-BA856.

Galon, J., and D. Bruni (2019), Approaches to treat immune hot, altered and cold tumours with combination immunotherapies, *Nat Rev Drug Discov*, *18*(3), 197–218.

Gao, H., et al. (2015), High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response, *Nat. Med.*, *21*(11), 1318–1325.

Geel, R. V., et al. (2014), Phase I study of the selective BRAFV600 inhibitor encorafenib (LGX818) combined with cetuximab and with or without the α-specific PI3K inhibitor BYL719 in patients with advanced BRAF-mutant colorectal cancer., *Journal of Clinical Oncology*, *32*(15_suppl), 3514–3514, doi:10.1200/jco.2014.32.15_suppl.3514.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2013), *Bayesian Data Analysis*, 3rd ed. ed., Chapman and Hall/CRC.

Geweke, J. (1992), Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments, in *In Bayesian Statistics*, pp. 169–193, University Press.

Geyer, C. J. (2011), Introduction to Markov chain Monte Carlo, in *Handbook of Markov chain Monte Carlo*, edited by S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, chap. 1, pp. 3–48, Chapman and Hall/CRC.

Giulino-Roth, L., et al. (2017), Inhibition of Hsp90 Suppresses PI3K/AKT/mTOR Signaling and Has Antitumor Activity in Burkitt Lymphoma, *Mol Cancer Ther*, *16*(9), 1779–1790.

Grant, R., A. Combs, and D. Acosta (2010), Experimental models for the investigation of toxicological mechanisms, in *Comprehensive Toxicology (Second Edition)*, edited by C. A. McQueen, second edition ed., pp. 203–224, Elsevier, Oxford, doi: https://doi.org/10.1016/B978-0-08-046884-6.00110-X.

Groisberg, R., and V. Subbiah (2021), Combination therapies for precision oncology: the ultimate whack-a-mole game, *Clin Cancer Res*, *27*(10), 2672–2674.

Gunning, P. W., U. Ghoshdastider, S. Whitaker, D. Popp, and R. C. Robinson (2015), The evolution of compositionally and functionally distinct actin filaments, *J Cell Sci*, *128*(11), 2009–2019.

Ha, M. J., S. Banerjee, R. Akbani, H. Liang, G. B. Mills, K.-A. Do, and V. Baladandayuthapani (2018), Personalized integrated network modeling of the cancer proteome atlas, *Scientific Reports*, *8*(1), 14,924, doi:10.1038/s41598-018-32682-x.

Haider, K., S. Rahaman, M. S. Yar, and A. Kamal (2019), Tubulin inhibitors as novel anticancer agents: an overview on patents (2013-2018), *Expert Opin Ther Pat*, *29*(8), 623–641.

Heaukulani, C., D. A. Knowles, and Z. Ghahramani (2014), Beta diffusion trees, in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, p. II–1809–II–1817, JMLR.org.

Hidalgo, M., et al. (2014), Patient-derived xenograft models: an emerging platform for translational cancer research, *Cancer Discov*, *4*(9), 998–1013.

hrer, J., et al. (2023), Statistical analysis of high-dimensional biomedical data: a gentle introduction to analytical goals, common approaches and challenges, *BMC Med*, *21*(1), 182.

Janku, F. (2014), Tumor heterogeneity in the clinic: is it a real problem?, *Ther Adv Med Oncol*, *6*(2), 43–51.

Joyce, J. A., and D. T. Fearon (2015), T cell exclusion, immune privilege, and the tumor microenvironment, *Science*, *348*(6230), 74–80.

Karlin, S., and Y. Rinott (1983), M-matrices as covariance matrices of multinormal distributions, *Linear Algebra and its Applications*, *52-53*, 419–438, doi:https://doi.org/10.1016/0024-3795(83)80027-5.

Knowles, D. A., and Z. Ghahramani (2015), Pitman-Yor diffusion trees for Bayesian hierarchical clustering, *IEEE Trans Pattern Anal Mach Intell*, *37*(2), 271–289.

Knowles, D. A., J. V. Gael, and Z. Ghahramani (2011), Message passing algorithms for dirichlet diffusion trees, *International Conference on Machine Learning (ICML)*.

Knuth, D. E. (1976), Big omicron and big omega and big theta, *ACM Sigact News*, *8*(2), 18–24.

Koga, Y., and A. Ochiai (2019), Systematic Review of Patient-Derived Xenograft Models for Preclinical Studies of Anti-Cancer Drugs in Solid Tumors, *Cells*, *8*(5).

Konopleva, M., G. Martinelli, N. Daver, C. Papayannidis, A. Wei, B. Higgins, M. Ott, J. Mascarenhas, and M. Andreeff (2020), MDM2 inhibition: an important step forward in cancer therapy, *Leukemia*, *34*(11), 2858–2874.

Krumbach, R., J. Schüler, M. Hofmann, T. Giesemann, H. H. Fiebig, and T. Beckers (2011), Primary resistance to cetuximab in a panel of patient-derived tumour xenograft models: activation of MET as one mechanism for drug resistance, *Eur J Cancer*, *47*(8), 1231–1243.

Kurtzeborn, K., H. N. Kwon, and S. Kuure (2019), MAPK/ERK Signaling in regulation of renal differentiation, *Int J Mol Sci*, *20*(7).

Lai, Y., X. Wei, S. Lin, L. Qin, L. Cheng, and P. Li (2017), Current status and perspectives of patient-derived xenograft models in cancer research, *J Hematol Oncol*, *10*(1), 106.

Lapointe, F.-J., and P. Legendre (1991), The generation of random ultrametric matrices representing dendrograms, *Journal of Classification*, *8*(2), 177–200, doi:10.1007/BF02616238.

Lauritzen, S., C. Uhler, and P. Zwiernik (2019), Maximum likelihood estimation in Gaussian models under total positivity, *Annals of Statistics*, *47*(4), 1835–1863, doi:10.1214/17-AOS1668.

Lauritzen, S. L. (1996), *Graphical Models*, New York : Oxford University Press.

Li, M., L. Li, and J. Kang (2023+), Bayesian inference of spatially varying correlations via thresholded gaussian processes.

Li, R., X. Zou, T. Zhu, H. Xu, X. Li, and L. Zhu (2020), Destruction of neutrophil extracellular traps promotes the apoptosis and inhibits the invasion of gastric cancer cells by regulating the expression of bcl-2, bax and nf-ϰb, *OncoTargets and Therapy*, *13*, 5271–5281, doi:10.2147/OTT.S227331.

Liu, H., X. Chen, J. Lafferty, and L. Wasserman (2010), Graph-valued regression, in *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2*, NIPS'10, p. 1423–1431, Curran Associates Inc., Red Hook, NY, USA.

Liu, H., F. Han, M. Yuan, J. Lafferty, and L. Wasserman (2012), High-dimensional semiparametric Gaussian copula graphical models, *The Annals of Statistics*, *40*(4), 2293 – 2326, doi:10.1214/12-AOS1037.

Liu, J. S. (1994), The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem, *Journal of the American Statistical Association*, *89*(427), 958–966.

Lu, H., Q. Zhou, J. He, Z. Jiang, C. Peng, R. Tong, and J. Shi (2020), Recent advances in the development of protein-protein interactions modulators: mechanisms and clinical trials, *Signal Transduct Target Ther*, *5*(1), 213.

Luo, X., J. Xu, J. Yu, and P. Yi (2021), Shaping Immune Responses in the Tumor Microenvironment of Ovarian Cancer, *Front Immunol*, *12*, 692,360.

Mahale, S., S. B. Bharate, S. Manda, P. Joshi, P. R. Jenkins, R. A. Vishwakarma, and B. Chaudhuri (2015), Antitumour potential of BPT: a dual inhibitor of CDK4 and tubulin polymerization, *Cell Death Dis*, *6*, e1743.

Martins-Cardoso, K., V. H. Almeida, K. M. Bagri, M. I. D. Rossi, C. S. Mermelstein, S. nig, and R. Q. Monteiro (2020), Neutrophil Extracellular Traps (NETs) Promote Pro-Metastatic Phenotype in Human Breast Cancer Cells through Epithelial-Mesenchymal Transition, *Cancers (Basel)*, *12*(6).

Mathai, A. (1980), Moments of the trace of a noncentral Wishart matrix, *Communications in Statistics - Theory and Methods*, *9*(8), 795–801, doi:10.1080/03610928008827921.

McCullagh, P. (2006), Structured covariance matrices in multivariate regression models, *Tech. rep.*, Department of Statistics, University of Chicago.

McCullagh, P., J. Pitman, and M. Winkel (2008), Gibbs fragmentation trees, *Bernoulli*, *14*(4), 988 – 1002, doi:10.3150/08-BEJ134.

Meinshausen, N., and P. Bühlmann (2006), High-dimensional graphs and variable selection with the Lasso, *The Annals of Statistics*, *34*(3), 1436 – 1462, doi:10.1214/009053606000000281.

Mezard, M., and A. Montanari (2009), *Information, Physics, and Computation*, Oxford University Press, Inc., USA.

Mieldzioc, A., M. Mokrzycka, and A. Sawikowska (2021), Identification of block-structured covariance matrix on an example of metabolomic data, *Separations*, *8*(11), doi:10.3390/separations8110205.

Miller, E., M. Owen, and J. S. Provan (2015), Polyhedral computational geometry for averaging metric phylogenetic trees, *Advances in Applied Mathematics*, *68*, 51–91, doi:10.1016/j.aam.2015.04.002.

Mulgrave, J. J., and S. Ghosal (2022), Regression-Based Bayesian Estimation and Structure Learning for Nonparanormal Graphical Models, *Stat Anal Data Min*, *15*(5), 611–629.

Murtagh, F., and P. Legendre (2014), Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion?, *Journal of Classification*, *31*(3), 274–295, doi:10.1007/s00357-014-9161-z.

Nabben, R., and R. S. Varga (1994), A Linear Algebra Proof that the Inverse of a Strictly Ultrametric Matrix is a Strictly Diagonally Dominant Stieltjes Matrix, *SIAM Journal on Matrix Analysis and Applications*, *15*(1), 107–113, doi:10.1137/s0895479892228237.

Narayan, R. S., et al. (2020), A cancer drug atlas enables synergistic targeting of independent drug vulnerabilities, *Nat Commun*, *11*(1), 2935.

Neal, R. (2003), Density Modeling and Clustering Using Dirichlet Diffusion Trees, *Bayesian Statistics*, *7*, 619–629.

Ni, Y., F. C. Stingo, and V. Baladandayuthapani (2019), Bayesian graphical regression, *Journal of the American Statistical Association*, *114*(525), 184–197, doi:10.1080/01621459.2017.1389739.

Ni, Y., V. Baladandayuthapani, M. Vannucci, and F. C. Stingo (2022a), Bayesian graphical models for modern biological applications, *Statistical Methods & Applications*, *31*(2), 197–225, doi:10.1007/s10260-021-00572-8.

Ni, Y., F. C. Stingo, and V. Baladandayuthapani (2022b), Bayesian covariate-dependent gaussian graphical models with varying structure, *Journal of Machine Learning Research*, *23*(242), 1–29.

Nirmal, A. J., T. Regan, B. B. Shih, D. A. Hume, A. H. Sims, and T. C. Freeman (2018), Immune Cell Gene Signatures for Profiling the Microenvironment of Solid Tumors, *Cancer Immunol Res*, *6*(11), 1388–1400.

Nunes, M., et al. (2015), Evaluating patient-derived colorectal cancer xenografts as preclinical models by comparison with patient clinical data, *Cancer Research*, *75*(8), 1560–1566, doi:10.1158/0008-5472.CAN-14-1590.

Nye, T. M. (2020), Random walks and Brownian motion on cubical complexes, *Stochastic Processes and their Applications*, *130*(4), 2185–2199, doi:10.1016/j.spa.2019.06.013.

Oh, D. Y., and Y. J. Bang (2020), HER2-targeted therapies - a role beyond breast cancer, *Nat Rev Clin Oncol*, *17*(1), 33–48.

Owen, M., and J. S. Provan (2011), A fast algorithm for computing geodesic distances in tree space, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *8*(1), 2–13, doi:10.1109/TCBB.2010.3.

Peng, J., P. Wang, N. Zhou, and J. Zhu (2009), Partial Correlation Estimation by Joint Sparse Regression Models, *J Am Stat Assoc*, *104*(486), 735–746.

Peterson, C. B., F. C. Stingo, and M. Vannucci (2015), Bayesian Inference of Multiple Gaussian Graphical Models, *J Am Stat Assoc*, *110*(509), 159–174.

Pillai, R. N., M. Behera, L. D. Berry, M. R. Rossi, M. G. Kris, B. E. Johnson, P. A. Bunn, S. S. Ramalingam, and F. R. Khuri (2017), HER2 mutations in lung adenocarcinomas: A report from the Lung Cancer Mutation Consortium, *Cancer*, *123*(21), 4099–4105.

Pitman, J. (2006), *Combinatorial stochastic processes*, vol. 1875, 1–247 pp., Springer, doi:10.1007/b11601500.

Pitt, M., D. Chan, and R. Kohn (2006), Efficient bayesian inference for gaussian copula regression models, *Biometrika*, *93*(3), 537–554.

Prangle, D., M. G. B. Blum, G. Popovic, and S. A. Sisson (2014), Diagnostic tools for approximate bayesian computation using the coverage property, *Australian & New Zealand Journal of Statistics*, *56*(4), 309–329, doi:10.1111/anzs.12087.

Rashid, N. U., et al. (2020), High-dimensional precision medicine from patient-derived xenografts, *Journal of the American Statistical Association*, *0*(0), 1–15, doi:10.1080/01621459.2020.1828091.

Repetto, M. V., M. J. Winters, A. Bush, W. Reiter, D. M. Hollenstein, G. Ammerer, P. M. Pryciak, and A. Colman-Lerner (2018), CDK and MAPK synergistically regulate signaling dynamics via a shared multi-site phosphorylation region on the scaffold protein Ste5, *Mol Cell*, *69*(6), 938–952.

Robert, C., et al. (2019), Five-year outcomes with dabrafenib plus trametinib in metastatic melanoma, *N Engl J Med*, *381*(7), 626–636.

Sawyers, C. L. (2013), Perspective: combined forces, *Nature*, *498*(7455), S7.

Scheipl, F., L. Fahrmeir, and T. Kneib (2012), Spike-and-slab priors for function selection in structured additive regression models, *Journal of the American Statistical Association*, *107*(500), 1518–1532, doi:10.1080/01621459.2012.737742.

Scott, S. L., A. W. Blocker, F. V. Bonassi, H. A. Chipman, E. I. George, and R. E. McCulloch (2016), Bayes and big data: The consensus monte carlo algorithm, *International Journal of Management Science and Engineering Management*, *11*(2), 78–88.

Sisson, S. A., Y. Fan, and M. Beaumont (2019), *Handbook of Approximate Bayesian Computation*, 1st ed. ed., Chapman and Hall/CRC.

Sokal, R. R., and F. J. Rohlf (1962), The comparison of dendrograms by objective methods, *Taxon*, *11*(2), 33–40.

Sonawane, A. R., S. T. Weiss, K. Glass, and A. Sharma (2019), Network Medicine in the Age of Biomedical Big Data, *Front Genet*, *10*, 294.

Spella, M., and G. T. Stathopoulos (2021), Immune Resistance in Lung Adenocarcinoma, *Cancers (Basel)*, *13*(3).

Steel, M. (2016), *Phylogeny Discrete and Random Processes in Evolution*, vol. 89, SIAM-Society for Industrial and Applied Mathematics.

Storey, J. D., and R. Tibshirani (2003), Statistical significance for genomewide studies, *Proceedings of the National Academy of Sciences of the United States of America*, *100*(16), 9440–9445, doi:10.1073/pnas.1530509100.

Sturmfels, B., C. Uhler, and P. Zwiernik (2021), Brownian Motion Tree Models Are Toric, *Kybernetika*, *56*(6), 1154–1175, doi:10.14736/kyb-2020-6-1154.

Sun, P. L., H. Sasano, and H. Gao (2017), Bcl-2 family in non-small cell lung cancer: its prognostic and therapeutic implications, *Pathol Int*, *67*(3), 121–130.

Sun, W., P. E. Sanderson, and W. Zheng (2016), Drug combination therapy increases successful drug repositioning, *Drug Discov Today*, *21*(7), 1189–1195.

Syed, V., K. Mukherjee, J. Lyons-Weiler, K. M. Lau, T. Mashima, T. Tsuruo, and S. M. Ho (2005), Identification of ATF-3, caveolin-1, DLC-1, and NM23-H2 as putative antitumorigenic, progesterone-regulated genes for ovarian cancer cells by gene profiling, *Oncogene*, *24*(10), 1774–1787.

Tentler, J. J., A. C. Tan, C. D. Weekes, A. Jimeno, S. Leong, T. M. Pitts, J. J. Arcaroli, W. A. Messersmith, and S. G. Eckhardt (2012), Patient-derived tumour xenografts as models for oncology drug development, *Nat Rev Clin Oncol*, *9*(6), 338–350.

Topp, M. D., et al. (2014), Molecular correlates of platinum response in human high-grade serous ovarian cancer patient-derived xenografts, *Molecular Oncology*, *8*(3), 656–668, doi:https://doi.org/10.1016/j.molonc.2014.01.008.

Turner, B. M., P. B. Sederberg, S. D. Brown, and M. Steyvers (2013), A method for efficiently sampling from distributions with correlated dimensions, *Psychol Methods*, *18*(3), 368–384.

van Geel, R. M. J. M., et al. (2017), A phase Ib dose-escalation study of encorafenib and cetuximab with or without alpelisib in metastatic BRAF-mutant colorectal cancer, *Cancer Discov*, *7*(6), 610–619.

Vora, S. R., et al. (2014), CDK 4/6 inhibitors sensitize PIK3CA mutant breast cancer to PI3K inhibitors, *Cancer Cell*, *26*(1), 136–149.

Wang, C., Q. Yu, T. Song, Z. Wang, L. Song, Y. Yang, et al. (2022), The heterogeneous immune landscape between lung adenocarcinoma and squamous carcinoma revealed by single-cell RNA sequencing, *Signal Transduct Target Ther*, *7*(1), 289.

Wang, Y. L., C. C. Lee, Y. C. Shen, P. L. Lin, W. R. Wu, Y. Z. Lin, et al. (2021), Evading immune surveillance via tyrosine phosphorylation of nuclear PCNA, *Cell Rep*, *36*(8), 109,537.

Weinstein, J. N., E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, et al. (2013), The Cancer Genome Atlas Pan-Cancer analysis project, *Nat Genet*, *45*(10), 1113–1120.

Whiteside, T. L. (2008), The tumor microenvironment and its role in promoting tumor growth, *Oncogene*, *27*(45), 5904–5912.

Wu, L., et al. (2021), ErbB3 is a critical regulator of cytoskeletal dynamics in brain microvascular endothelial cells: implications for vascular remodeling and blood-brain-barrier modulation, *J Cereb Blood Flow Metab*, p. 271678X20984976.

Yang, Y., Y. Yang, J. Yang, X. Zhao, and X. Wei (2020), Tumor Microenvironment in Ovarian Cancer: Function and Therapeutic Strategy, *Front Cell Dev Biol*, *8*, 758.

Yao, T.-H., Z. Wu, K. Bharath, J. Li, and V. Baladandayuthapan (2023), Probabilistic learning of treatment trees in cancer, *Annals of Applied Statistics*, p. Forthcoming.

Yoshida, G. J. (2020), Applications of patient-derived tumor xenograft models and tumor organoids, *Journal of Hematology & Oncology*, *13*(1), 4, doi:10.1186/s13045-019-0829-z.

Yu, H., J. Moharil, and R. H. Blair (2020), Bayesnetbp: An r package for probabilistic reasoning in bayesian networks, *Journal of Statistical Software*, *94*(3), 1–31, doi:10.18637/jss.v094.i03.

Yuan, Y., W. Wen, S. E. Yost, Q. Xing, J. Yan, E. S. Han, J. Mortimer, and J. H. Yim (2019), Combination therapy with BYL719 and LEE011 is synergistic and causes a greater suppression of p-S6 in triple negative breast cancer, *Sci Rep*, *9*(1), 7509.

Zhang, J., and Y. Li (2022), High-dimensional gaussian graphical regression models with covariates, *Journal of the American Statistical Association*, *0*(0), 1–13, doi:10.1080/01621459.2022.2034632.

Zhang, X., et al. (2013), A Renewable Tissue Resource of Phenotypically Stable, Biologically and Ethnically Diverse, Patient-Derived Human Breast Cancer Xenograft Models, *Cancer Research*, *73*(15), 4885–4897, doi:10.1158/0008-5472.CAN-12-4081.

Zhao, Y., H. Yu, and W. Hu (2014), The regulation of MDM2 oncogene and its impact on human cancers, *Acta Biochim Biophys Sin (Shanghai)*, *46*(3), 180–189.

Zhao, Y., et al. (2018), Development of a new patient-derived xenograft humanised mouse model to study human-specific tumour microenvironment and immunotherapy, *Gut*, *67*(10), 1845–1854.

Zhu, H., X. Zhu, L. Zheng, X. Hu, L. Sun, and X. Zhu (2017), The role of the androgen receptor in ovarian cancer carcinogenesis and its clinical implications, *Oncotarget*, *8*(17), 29,395–29,405.

Zorzi, M., and A. Ferrante (2012), On the estimation of structured covariance matrices, *Automatica*, *48*(9), 2145–2151, doi:https://doi.org/10.1016/j.automatica.2012.05.057.