TOPICS IN MODELING OF MULTIVARIATE MIXED DATA TYPES AND HIGHLY MULTIVARIATE SPATIAL DATA

by

Debangan Dey

A dissertation submitted to Johns Hopkins University

in conformity with the requirements for the degree of

Doctor of Philosophy

Baltimore, Maryland

April, 2022

Abstract

In public health, surveillance constitutes systematic data collection to analyze, interpret and implement public policies. Notable examples of surveillance include periodic large health surveys (e.g. National Health and Nutrition Examination Survey) and environmental surveillance through measuring pollutants and meteorological data at multiple monitoring sites. With technological advancements, we can record multiple varieties of data at each time point or spatial location. Unfortunately, the existing statistical literature is limited to modeling such complex multivariate data due to either lack of generalizability, scalability, or computational efficiencies. This dissertation focuses on building global, scalable, and efficient methods to bridge those gaps in the literature. This work focuses explicitly on three contexts: (1) using semi-parametric Gaussian copulas to build joint models of multivariate mixed type of data (binary/ordinal/truncated/continuous) that can define mutually consistent regression models for any type of outcome, (2) develop a consistent and robust estimator of the ubiquitous measure of classification accuracy: Area Under the Curve (AUC) under complex survey designs and connect it to a latent R-square analogous to linear models, and (3) propose a class of "Graphical Gaussian Processes" that can efficiently model highly multivariate

spatial data where tens or hundreds of variables are observed at each spatial location.

Thesis Committee

Primary Readers

Dr. Vadim Zipunnikov (Advisor)
Associate Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Dr. Abhirup Datta (Co-advisor)
Associate Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Dr. Kathleen Merikangas
Adjunct Professor
Department of Mental Health
Johns Hopkins Bloomberg School of Public Health

Dr. Patrick Finan
Associate Professor
Department of Psychiatry and Behavioral Sciences
Johns Hopkins Schoool of Medicine

Alternate Readers

Dr. Mei-Cheng Wang
Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Dr. Adam Spira
Professor
Department of Mental Health
Johns Hopkins Bloomberg School of Public Health

Dr. Peter Zandi
Professor
Department of Psychiatry and Behavioral Sciences
Johns Hopkins Schoool of Medicine

Acknowledgments

As this dissertation marks the end of my PhD journey, I want to take this moment to thank the people without whom this work would have been impossible.

First, I want to thank my advisor Dr. Vadim Zipunnikov, for his valuable support and guidance for the last 7 years. Thanks to Vadim for introducing me to the world of Biostatistics during my internship in 2015 and showing me the amount of public health impact we could make with statistical methods. This experience kept me motivated to pursue my PhD later on, and Vadim has been an incredible mentor throughout. Thanks to Vadim for helping me grow as a better researcher, writer, and communicator. I also want to thank Vadim for showing me the importance of a healthy work-life balance. And lastly, thanks to him for always keeping the conversations entertaining with his impeccable sense of humor.

I want to thank my co-advisor Dr. Abhirup Datta, for his outstanding mentorship and support throughout my PhD years. I thank Abhirup for inspiring me to innovative thinking and teaching me the necessary tools to become a successful statistician. I am grateful to him for our countless productive and intuitive discussions and for pushing me through the challenging PhD life.

I want to thank Dr. Kathleen Merikangas for mentoring me throughout my predoctoral fellowship at NIMH. Thanks to Kathleen that, I have learned how statistics can answer a lot of scientific questions about mental health. She has been hugely accommodating during my fellowship and was always available to answer any questions. I am really grateful to know her as a person for the last two and a half years and the amount of positive impact she left in my PhD journey and beyond.

I want to thank my thesis readers Dr. Patrick Finan and alternate readers Dr. Mei-Cheng Wang, Dr. Adam Spira, and Dr. Peter Zandi, for accepting my invitation to be on my committee. Thanks to them for their valuable feedback and for enriching this dissertation. I also want to thank Dr. Lisa Reider for being on my oral examination committee and providing her helpful comments on my research.

I want to thank Dr. Sudipto Banerjee for being instrumental to my research on spatial statistics and helping me through my post-PhD application process.

I want to thank Prosenjit, Parichoy, Diptavo, Rahul, and Arkajyoti (my seniors from Indian Statistical Institute in the department) for making me feel at home in Baltimore. In addition, I want to thank Marta, Junrui, and Andrew for being the most helpful office-mates and friends, for the times we spent when going to the office were still normal. I want to thank Mary Joy for always providing me a solution to all the stupid questions I had about administrative issues.

I want to thank the Johns Hopkins Pickup Soccer Group, which made me survive the stress of the PhD program. I am really thankful for the opportunity

to regularly play the sport I love most and make plenty of friends from the field.

I want to thank my Baltimore friends Sayan, Soumyo, Arunima, Prosenjit, Neha, Anindya, Rahul, Diptavo, and Sayantan for the countless hangouts, exploring restaurants, playing cricket/board games, and more importantly, for recreating the home away from home.

I want to thank my ISI friends Samriddha, Sohom, Dhrubajyoti, Rohan, Soham, Nabarun, Debarghya, Suyash, Arnab, and Sayan for the various trips we enjoyed around the world. Thanks for always being there to travel whenever I needed a fresh breath of air.

Lastly, I want to thank my family for providing constant motivation, even being a thousand miles away. Especially, I want to thank my parents for their persistent effort and commitment towards me, which have made me a successful PhD graduate.

Dedication

For my love of soccer, the beautiful game.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In public health, surveillance is a continuous collection of health-related data to analyze, interpret and implement public health policy and management. One domain of such surveillance is periodic large national health surveys to monitor the health status of a population. A specific example would be National Health and Nutritional Examination Survey, an annual survey collecting data through interviews and physical examinations from a sample of participants that is representative of the noninstitutionalized civil US population. A different type of surveillance is environmental surveillance that typically collects pollutant and meteorological data from spatial monitoring sites. With the advent of modern technologies, these surveillances can contain multiple varieties of data at each time-point and/or spatial location, thus resulting in complex multivariate datasets. This data revolution prompts a dire need to build global, scalable, and efficient statistical frameworks to analyse the collected data and extract important scientici insights. This thesis will focus on methodological development for modelling multivariate mixed type data and high-dimensional special data collected by health and meterological

surveys, respectively.

Large-scale health surveys such as National Health and Nutrition Examination Survey (NHANES) is a rich source of information that may help to understand complex inter-relationship between human health behaviours such as physical activity, sleep, dietary preferences, smoking, drinking, and others and various cardiometabolic biomarkers of health, multiple comorbidities, and health deficits. Most of the participant-specific information is recorded via many binary, ordinal, truncated, continuous, and categorical variables. Therefore, to gain truly novel insights in understanding complex interactions between all those variables, there is a critical need for flexible analytical frameworks to perform joint and conditional modelling of mixed data types.

Chapter 2 proposes Semiparametric Gaussian Copula Regression Modelling (SGCRM) that allows to model a joint dependence structure between observed continuous, truncated, ordinal, and binary variables and to construct conditional models with these four data types as outcomes with a guarantee that the models are mutually consistent among each other. SGCRM assumes a semiparametric Gaussian copula Liu, Lafferty, and Wasserman, 2009; Liu et al., 2012 mechanism that generates observed variables by monotonically transforming marginals of latent multivarite normal random variable and, then, dichotimizing/truncating those transformed variables. SGCRM estimates the correlation matrix of the latent normal variables through an inversion of "bridges" between Kendall's Tau rank correlations of observed mixed data type variables and latent Gaussian correlations. We propose computationally

efficient methods to predict latent variables and to do imputation of missing data. We establish the asymptotic normality of estimators and provide a computationally efficient way to calculate their asymptotic variance. Using NHANES 2003-06 data, we illustrate SGCRM and compare it with the traditional conditional regression models including simple linear regression, truncated Gaussian regression, ordinal probit, and probit regressions.

In Chapter 3, we bridge another gap in large sample survey literature. Area Under the Curve (AUC) is arguably the most popular measure of classification accuracy. But, in complex surveys, we need pairwise survey weights for participants to calculate consistent estimators of AUC. Unfortunately, the pairwise participant weights are often unavailable in the survey data. Moreover, AUC lacks a general interpretation unlike metrics like R^2 which denotes the variance explained in traditional linear models. To solve these issues, we use a semiparametric framework to introduce a latent scale-invariant R^2, a novel measure of variation explained for an observed binary outcome and an observed continuous predictor, and then directly link the latent R^2 to AUC. This enables a mutually consistent simultaneous use of AUC as a measure of classification accuracy and the latent R^2 as a scale-invariant measure of explained variation. Specifically, we employ Semiparametric Gaussian Copula (SGC) to model a joint dependence between observed binary outcome and observed continuous predictor via the correlation of latent standard normal random variables. Under SGC, we show how, both population-level AUC and latent scale-invariant R^2, defined as a squared latent correlation,

can be estimated using any of the four rank statistics calculated on binary-continuous pairs: Wilcoxon rank-sum, Kendall's Tau, Spearman's Rho, and Quadrant rank correlations. We then focus on three implications and applications: i) we explicitly show that under SGC, the population-level AUC and the population-level latent R^2 are related via a monotone function that depends on the population-level prevalence rate, ii) we propose Quadrant rank correlation as a robust semiparametric version of AUC; iii) we demonstrate how, under complex-survey designs, Wilcoxon rank sum statistics and Spearman and Quadrant rank correlations provide asymptotically consistent estimators of the population-level AUC using only single-participant survey weights. We illustrate these applications using binary outcome of five-year mortality and continuous predictors including Albumin, Systolic Blood Pressure, and accelerometry-derived measures of total volume of physical activity collected in 2003-2006 NHANES cohorts.

In the next chapter, we deal with data collected in environmental surveillance where sensors measure multiple variables across multiple sites. Gaussian Processes (GP) is a widely popular tool for researchers in analyzing such geospatial data, owing to their convenient formulation using Gaussian likelihood, high prediction accuracy, and publicly available software. Modeling multiple variables separately as a univariate GP can only inform us about spatial dependence. However, these marginal analyses will not be able to learn about the interdependencies among these variables. Suppose two variables are strongly correlated, and we only observe one of them at a location. In that case, we can make a more informed prediction of the missing variable

borrowing the strengths of both intervariable and spatial dependence. Hence, the focus of spatial analysis is increasingly shifting to a multivariate paradigm.

While building a multivariate GP, we want to retain the interpretability of the marginal properties of each spatial surface. Except for the multivariate Matérn GP (Apanasovich, Genton, and Sun, 2012; Gneiting, Kleiber, and Schlather, 2010), most other multivariate GPs fail to retain this property. We also want our models to be scalable in terms number of variables (q). Unfortunately, most multivariate covariance functions proposed in the literature involve a parameter set whose dimensionality is quadratic in the number of variables (q). Thus, even for a modest number of locations (n), these methods need optimization in high-dimensional space and suffer from the curse of dimensionality. These difficulties have restricted most illustrations of multivariate GPs to bi- or tri-variate applications. In Chapter 4, we introduce models for highly multivariate geospatial data where tens or hundreds of variables are often measured at each spatial location.

We focus on multivariate Matérn for its appealing interpretability and resort to graphical modeling techniques to reduce computational complexity. First, using the notion of process-level conditional independence, we introduce a graph between variable processes. Then, we propose a class of multivariate "Graphical Gaussian Processes" using a general construction called "stitching" that retains marginal distributions, ensures process-level conditional independence among variables, and approximately retains the cross-covariances for the variables corresponding to the graph edges. For

decomposable graphs, our approach offers huge computational gain and reduces the parameter space to cliques and separators of the variable graph. We tailor our approach so that we can learn about the variable graph simultaneously with the process parameters. We prove key theoretical results and demonstrate the utility in an application to air-pollution modelling.

Overall, this dissertation focuses on building global, scalable, and efficient statistical methods to solve the modeling issues for mixed-type and multivariate spatial data abundant in modern-day epidemiological and environmental studies. While there will be essential impacts resulting from the proposed methodology in public health, the methods developed in this dissertation will also enrich the state-of-the-art statistical literature effectuating further research in the analysis of even more complex data coming from health and spatial surveillance.

References

Liu, Han, John Lafferty, and Larry Wasserman (2009). "The nonparanormal: Semiparametric estimation of high dimensional undirected graphs". In: *Journal of Machine Learning Research* 10.Oct, pp. 2295–2328.

Liu, Han, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman (2012). "High-dimensional semiparametric Gaussian copula graphical models". In: *The Annals of Statistics* 40.4, pp. 2293–2326.

Apanasovich, Tatiyana V, Marc G Genton, and Ying Sun (2012). "A valid Matérn class of cross-covariance functions for multivariate random fields with any number of components". In: *Journal of the American Statistical Association* 107.497, pp. 180–193.

Gneiting, Tilmann, William Kleiber, and Martin Schlather (2010). "Matérn cross-covariance functions for multivariate random fields". In: *Journal of the American Statistical Association* 105.491, pp. 1167–1177.

Chapter 2

Semiparametric Gaussian Copula Regression Modelling for Mixed Data Types (SGCRM)

2.1 Introduction

Clinical and epidemiological studies as well as health surveys collect a large number of health outcomes as well as physiological and clinical measures. This information is typically encoded via a collection of continuous, truncated, ordinal, and binary variables. As a main motivating example, we consider National Health and Nutrition Examination Survey (NHANES), a cross-sectional, nationally representative survey that assesses demographic, dietary and health-related questions that can be used to better understand trends in health and nutrition. For example, self-reported current health status (HSD010, on scale 1-5, 1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor) is an example of the ordinal variable. Self-reported mobility problem (NAME, 0/1 = no/yes for mobility difficulty) and follow-up mortality status (mortstat, 0/1 = alive/deceased at the follow-up) are examples of binary variables. In

2003-2006 waves, NHANES measured physical activity on more than 10000 participants using accelerometers worn by the participants for seven days (REF). At participant level, accelerometry-measured activity is often summarized using two variables: Total Activity Count (TAC), a continuous measure of the total volume of physical activity, and a time spent doing Vigorous Physical Activity (VPA), which is typically seen as a truncated variable, as many participants do zero minutes of high-intensity physical activity. Figure S2 show scatterplot matrix of these five variables. Many NHANES studies used these five variables as outcomes in linear regression models (TAC), truncated regression (VPA), ordinal regression (health status), and logistic regression models (mobility difficulty and mortality status). Although, the results of those models have been summarized in many reviews (REF), those models and potentially their results are not necessarily mutually consistent. To better understand complex interrelationship between human health behaviors and various health outcomes using studies like NHANES, we need to, first, learn how to better understand complex interdependencies between mixed data types variables by developing flexible frameworks for their joint modelling.

Due to a lack of standard joint models for multivariate mixed data types, conditional modelling is frequently used instead. Conditional models focus on fixing one of the variables as an outcome and modeling the mean of this outcome as a function of other variables. When outcome is binary, the traditional way is to model the conditional mean of the outcome as a function of a linear combination of other variables, logistic and probit regressions are two most popular models. If the outcome is ordinal, it is essential to capture the

directionality or the increase in order of that outcome. The most popularly used ordinal regression model is the cumulative ordinal regression model McCullagh, 1980. It assumes that the truncation of underlying latent continuous variable rise to the observed categories. The coefficients in a cumulative model are easily interpretable in terms of transformed odds. However, when we consider different coefficients for different levels of the outcome, the estimation becomes difficult due to the imposed ordering constraint between estimating coefficients. When it comes to truncated outcome, the econometrics literature first introduced the most popular Tobit models (Tobin, 1958; Heckman, 1976; Hausman and Wise, 1977), which assumes the outcome to originate from the truncated observation of a latent normal variable. Then the conditional mean of the latent variable is modeled as a linear function of the predictors. The Tobit model was generalized as the Hurdle model (Cragg, 1971) where we fit conditional models for both the truncated and non-truncated outcome separately at the cost of introducing more parameters. Both these models require likelihood based numerical estimation approaches and can suffer from convergence issues.

The conditional models discussed above can only model the conditional mean functions of the outcome distribution given other variables. Hence, they can only draw inferences from a subset of available information. On the other hand, modeling the joint distribution of the data can give us a more comprehensive view of data. To address this issue, researchers have developed a class of factorization models that assume the marginal distributions of a set of variables and the conditional distribution of the rest of the variables given

the rest. A popular example is General Location models (GLOMs) (Olkin and Tate, 1961), which enforce conditional normality for continuous variables and arbitrary distribution for discrete components. Another example is conditional grouped conditional models (CGCMs) (Anderson and Pemberton, 1985; Leon and Carriégre, 2007; De Leon, 2005) that assumes that the discrete variables are derived by truncating a latent multivariate continuous distribution. These models employs polychoric and polyserial correlations to estimate joint covariance structure. Even though, factorization models provide a convenient way for specifying mixed distributions, they induce a hierarchy in the data that depend on the direction of conditioning. Different factorizations for the same set of variables can lead to different interpretations for the estimated parameters and different inferences for associations.

Another popular direction employs copulas (Song, Li, and Yuan, 2009) that define Vector Generalized Linear Models (VGLMs) for modeling mixed data type outcomes. This approach requires embedding the marginal distributions (univariate Generalized Linear Models) into the joint distribution function via a Gaussian copula. Gaussian copulas is a popular choice to couple marginal distributions because of their analytical tractability and flexibility. Jiryaie et al., 2016 introduced Gaussian copula distributions (GCD) that take a latent variable approach to embed discrete variables using the Gaussian copula. However, CGCMs, GLOMs, VGLMs, and GCDs require likelihood-based inference and are computationally intensive for high dimensions. Pairwise likelihood-based approaches (De Leon, 2005; Jiryaie et al., 2016) reduce the computational burden but can make worse classification than

the full likelihood-based approach (Jiryaie et al., 2016). It is more desirable to have a joint modeling framework that treats the mixed data types variables symmetrically and is scalable to high dimensions.

Recently, semiparametric models have seen a wide adaptation for joint modelling of multivariate mixed type data. Wang and Hua, 2014 used likelihood-based inference and Cai and Zhang, 2015 developed a rank-based approach to estimate a joint semiparametric Gaussian copula for continuous variables. Fan, Xue, and Zou, 2016 extended the rank-based approaches to perform quantile regression on continuous variables. Rank-based estimation of the covariance of the semi-parametric Gaussian copula family (Liu, Lafferty, and Wasserman, 2009; Liu et al., 2012) has been particularly attractive because of the fast and robust estimation procedure. Therefore, we have seen multiple extensions of the use of latent semi-parametric Gaussian copula to model mixed types data. Fan et al., 2017 developed the estimation in the case of of binary and continuous variables. Yoon, Carroll, and Gaynanova, 2018 extended the approach to include truncated variables, and Quan, Booth, and Wells, 2018 has additionally extended it to include ternary variables (ordinal variables with three categories) and general ordinal-continuous pairs of variables. Feng and Ning, 2019 represented an ordinal variable via multiple dummy binary variables and took a weighted correlation approach to recover the latent correlation for ordinal pairs with more than three categories. Whereas, Zhang et al., 2018 arrived at an incorrect bridging function trying to tackle the general ordinal case. Huang, Müller, and Gaynanova, 2021 provided an R package for speeding the computation of latent correlations between pairs of binary,

ternary, truncated and continuous variables using a numerical interpolation approach. But, this fast algorithm also requires the knowledge of the original analytic function. Unfortunately, there's no existing approach that can handle general ordinal variable within this framework. Our first contribution closes this gap by providing bridging formulas for the general case of an ordinal variable.

Our next contribution is a semi-parametric Gaussian copula joint modeling framework that treats mixed variables symmetrically and is scalable to high-dimensions. The main advantages of the proposed framework is as follows: i) joint modeling for mixed data type, ii) mutually consistent conditional modeling that is alternative to a wide range of conditional models including linear regression, logistic regression, logistic-ordinal, truncated outcome, iii) latent representations provide natural normalization/scaling of mixed data types, iv) the approach is semi-parametric and likelihood free, so it only requires estimating pair-wise Kendall's Tau correlations, v) interpretation of regression coefficients is familiar and based on R^2 interpretation of predicted variability on latent space, vi) the approach allows to define R^2 for all four types of outcomes and in addition, allows to quantify added-value (using latent-R^2 scale) as in $R^2_l(y|x,z) = R^2_l(y|x) + R^2_l(y|(z|x))$, vii) the approach is robust and computationally fast, viii) allows to do missing data imputation.

The rest of the paper is organized as follows. Chapter 2 reviews semiparametric Gaussian Copula models and states a novel result on incorporating ordinal case. Chapter 3 introduces Semiparametric Gaussian Copula Regression Model and states main asymptotical results for regression parameters.

Chapter 4 describes key advantages of SGCRM and covers two methodological applications of SGCRM - prediction of latent variables and imputation of missing observations. Chapter 5 studies the performance of SGCRM via a few simulation scenarios. Chapter 6 illustrates SGCRM in NHANES data. Finally, Chapter 7 concludes with a summary and a discussion.

2.2 Gaussian copula Model

Classical Gaussian model assumptions have been popular due to their computational simplicity. However, these assumptions can be too restrictive. As an alternative, Liu, Lafferty, and Wasserman, 2009 proposed non-Paranormal distribution (NPN) which can be seen as a semiparametric Gaussian Copula model.

Definition 2.2.1. (Non-paranormal distribution) A random vector $Z = (Z_1, \ldots, Z_p)' \sim NPN_p(0, \Sigma, f)$ if there exist monotone transformation functions $f = (f_1, ..., f_p)$ such that $L = (L_1, \ldots, L_p) = f(Z) = (f_1(Z_1), \ldots, f_p(Z_p)) \sim N(0, \Sigma)$ where $\Sigma_{jj} = 1$ for $1 \leqslant j \leqslant p$.

The last assumption on Σ is made to ensure the identifiability of the distribution as shown in Liu, Lafferty, and Wasserman, 2009.

Fan et al., 2017 introduced latent non-paranormal distribution which extended non-paranormal distribution to jointly model binary and continuous data. Yoon, Carroll, and Gaynanova, 2018 introduced truncated variables using latent non-paranormal variables and Quan, Booth, and Wells, 2018 extended the distribution to ternary-continuous pairs and ternary-ternary

pairs. We generalize these and demonstrate how general ordinal case can be treated. We define Generalized Latent Non-paranormal (GLNPN) distribution which covers four mixed data types including continuous, truncated, (general) ordinal (k ordered categories), binary variables.

Definition 2.2.2. (Generalized latent non-paranormal distribution) Suppose we observe a random vector $X = (X_c, X_t, X_o, X_b)'$, where X_c is p_c-dimensional continuous variable, X_t is p_t-dimensional truncated, X_o is p_o-dimensional ordinal (j-th ordinal variable has levels $\{0, 1, \cdots, l_j - 1\}$), and X_b is p_b-dimensional binary variable, and $p = p_c + p_t + p_o + p_b$. We assume that there exist latent variables $Z = (Z_c, Z_t, Z_o, Z_b)'$ such that

$$
\begin{aligned}
X_{cj} &= Z_{cj}, 1 \leqslant j \leqslant p_c \\
X_{tj} &= Z_{tj} I(Z_{tj} > \delta_{tj}), 1 \leqslant j \leqslant p_t \\
X_{oj} &= \sum_{k=0}^{l_j - 1} k I(\delta_{ojk} \leqslant Z_{oj} < \delta_{oj(k+1)}), 1 \leqslant j \leqslant p_o; \delta_{oj0} = -\infty, \delta_{ojl_j} = \infty \\
X_{bj} &= I(Z_{bj} > \delta_{bj}), 1 \leqslant j \leqslant p_b
\end{aligned}
\tag{2.1}
$$

If $Z = (Z_c, Z_t, Z_o, Z_b)' \sim NPN(0, \Sigma, f)$, we denote that $X = (X_c, X_t, X_o, X_b)' \sim \text{GLNPN}_p(0, \Sigma, f, \delta)$, where $\delta = \{\delta_{tj}; j = 1, \cdots, p_t\} \cup (\cup_{j=1}^{p_o} \{\delta_{oj(k+1)}; k = 0, \cdots, l_j\}) \cup \{\delta_{bj}; j = 1, \cdots, p_b\}$, i.e., is the set containing cutoffs for truncated, ordinal and binary variables.

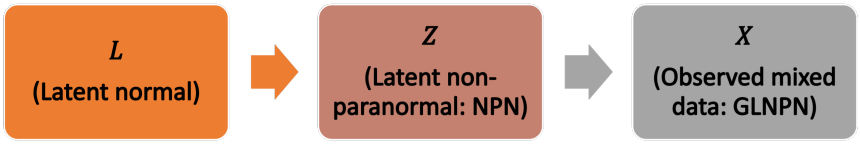


Figure 2.1: The data generation flow of GLNPN distribution

To shorten notations, we will refer to observed continuous, truncated, ordinal, binary variables generated according to GLNPN distribution as CTOB-GLNPN variables. Figure 2.1 shows the flowchart of the data generation mechanism for the observed GLNPN variables. Figure 2.2 shows an example of four observed CTOB-GLNPN variables generated via monotone-transformation-then-truncation of latent bivariate normal variables.

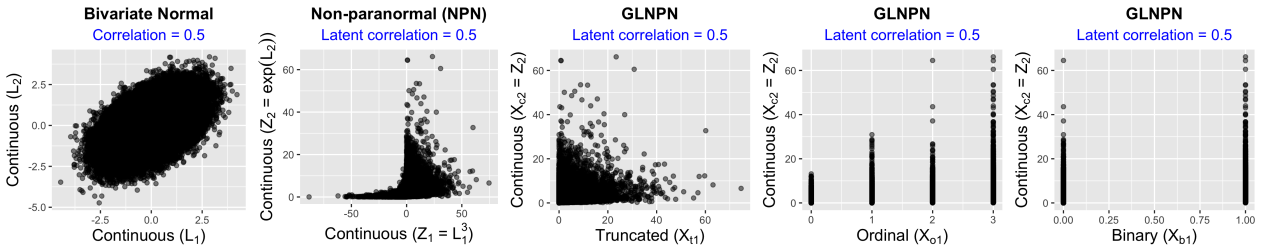


Figure 2.2: From left to right: (i) a scatterplot of bivariate standard normal variables with correlation of 0.5, (ii) a continuous-continuous pair, (iii) a truncated-continuous pair, iv) an ordinal-continuous pair, (v) a binary-continuous pair

Cut-off parameters of the GLNPN distribution suffers from identifiability issues (see details at Fan et al., 2017). The joint probability mass function of the discrete component or the density of the truncated component only depends on the set of transformed cutoffs: $\boldsymbol{\Delta} = f(\boldsymbol{\delta}) = \{f_j(\delta_{tj}); j = 1, \cdots, p_t\} \cup (\cup_{j=1}^{p_o}\{f_j(\delta_{oj(k+1)}); k = 0, \cdots, l_j\})$
$\cup \{f_j(\delta_{bj}); j = 1, \cdots, p_b\} = \{\Delta_{tj}; j = 1, \cdots, p_t\} \cup (\cup_{j=1}^{p_o}\{\Delta_{oj(k+1)}; k = 0, \cdots, l_j\})$
$\cup \{\Delta_{bj}; j = 1, \cdots, p_b\}$. To emphasize that, we will generally refer to the GLNPN distribution as $\text{GLNPN}_p(0, \Sigma, f, \boldsymbol{\Delta})$.

As a result of the identifiablity constraints for the cutoffs, the binary and ordinal components of GLNPN distribution are marginally equivalent to the latent Gaussian distribution for binary and ordinal variables. This comes as no surprise as the discrete components does not have enough information

to identify the marginal transformations. However, when we model the discrete component jointly with continuous and truncated variables, the class of GLNPN distributions becomes much larger than the class of latent Gaussian distributions. The marginal transformations from continuous and truncated variables make the joint distribution of mixed variables more flexible and potentially can give a substantial advantage to better explain the association between mixed type of variables.

2.2.1 Estimation of Correlation Matrix

2.2.1.1 Bridging functions

A few authors (including Fan et al., 2017; Yoon, Carroll, and Gaynanova, 2018; Quan, Booth, and Wells, 2018) have considered Kendall's τ rank correlation to estimate latent correlation matrix Σ across several settings. We can calculate a sample Kendall's tau between j-th and k-th variable as follows:

$$\hat{\tau}_{jk} = \frac{2}{n(n-1)} \sum_{1 \leqslant i < i' < n} sgn\{(X_{ij} - X_{i'j})(X_{ik} - X_{i'k})\} \tag{2.2}$$

The construction of a sample Kendall's τ reveals that it is invariant under a monotone transformation. Now, for two independent copies X_i, X_i' of the random vector X, the population-level Kendall's τ is defined as

$$\tau_{jk} = E[sgn\{(X_{ij} - X_{i'j})(X_{ik} - X_{i'k})\}] \tag{2.3}$$

The population Kendall's Tau ($\tau_{jk} = E(\hat{\tau}_{jk})$) is typically related to the latent correlation Σ_{jk} through a one-to-one bridging function F which for non-continuous components will depend on cutoffs - $\tau_{jk} = F(\Sigma_{jk})$. The estimated

Type	Continuous	Truncated	Ordinal*	Binary
Continuous	Liu, Lafferty, and Wasserman, 2009	Yoon, Carroll, and Gaynanova, 2018	Quan, Booth, and Wells, 2018	Fan et al., 2017
Truncated	Yoon, Carroll, and Gaynanova, 2018	Yoon, Carroll, and Gaynanova, 2018	Theorem 2.2.1	Yoon, Carroll, and Gaynanova, 2018
Ordinal	Quan, Booth, and Wells, 2018	Theorem 2.2.1	Theorem 2.2.1	Theorem 2.2.1
Binary	Fan et al., 2017	Yoon, Carroll, and Gaynanova, 2018	Theorem 2.2.1	Fan et al., 2017

Table 2.1: The reference of bridging functions for all possible pairs of variables.
**Ordinal cases for only three categories were derived in Quan, Booth, and Wells, 2018*

latent correlation is then obtained as $\hat{\Sigma}_{jk} = F^{-1}(\hat{\tau}_{jk})$.

Fan et al., 2017 calculated the bridging function for a pair of binary and continuous variables, Yoon, Carroll, and Gaynanova, 2018 showed how to deal with truncated variables in addition to continuous and binary. Quan, Booth, and Wells, 2018 provided formulas for bridging functions for ternary variables and for general ordinal-continuous pairs of variables. Feng and Ning, 2019 broke an ordinal variable into multiple dummy binary variables and took a weighted correlation approach to recover the latent correlation for ordinal pairs with more than three categories, whereas, Zhang et al., 2018 arrived at an incorrect bridging function trying to tackle the general ordinal case. We summarize the references to the correct bridging functions for all possible pairs of variables in Table 2.1. Our contribution here is to derive bridging functions for the general ordinal variable with arbitrary number of ordinal levels. The results is summarized in in Theorem 2.2.1.

The following theorem provides the bridging function for any arbitrary pair of variables -

Theorem 2.2.1. *Let X_j, X_k be two GLNPN variables, then the population Kendall's Tau is related to the latent correlation as follows: $\tau_{jk} = F(\Sigma_{jk})$, where F can depend on the cutoff Δ_j, Δ_k, which denotes the cutoff scalar or vector and corresponds to non-continuous components of vectors X_j and X_k. The bridging functions corresponding*

to all pairs of variables are as follows-

$$
\begin{aligned}
F_{\mathrm{cc}}(\rho) &= \frac{2}{\pi}\sin^{-1}(\rho) \\
F_{\mathrm{bb}}(\rho;\Delta_j,\Delta_k) &= 2\left\{\Phi_2(\Delta_j,\Delta_k;\rho)-\Phi(\Delta_j)\Phi(\Delta_k)\right\} \\
F_{\mathrm{cb}}(\rho;\Delta_j) &= 4\Phi_2(\Delta_j,0;\rho/\sqrt{2})-2\Phi(\Delta_j) \\
F_{\mathrm{tb}}(\rho;\Delta_j,\Delta_k) &= 2\{1-\Phi(\Delta_j)\}\Phi(\Delta_k)-2\Phi_3\left(-\Delta_j,\Delta_k,0;S_{3a}(\rho)\right)-2\Phi_3\left(-\Delta_j,\Delta_k,0;S_{3b}(\rho)\right) \\
F_{\mathrm{ct}}(\rho;\Delta_j) &= -2\Phi_2(-\Delta_j,0;1/\sqrt{2})+4\Phi_3\left(-\Delta_j,0,0;S_3(r)\right) \\
F_{\mathrm{tt}}(\rho;\Delta_j,\Delta_k) &= -2\Phi_4(-\Delta_j,-\Delta_k,0,0;S_{4a}(\rho))+2\Phi_4(-\Delta_j,-\Delta_k,0,0;S_{4b}(\rho)) \\
F_{\mathrm{co}}(\rho;\Delta_j) &= \sum_{r=1}^{l_j-1}(4\Phi_3(\Delta_{jr},\Delta_{j(r+1)},0)-2\Phi(\Delta_{jr})\Phi(\Delta_{j(r+1)})) \\
F_{\mathrm{oo}}(\rho;\Delta_j,\Delta_k) &= 2(\sum_{r=1}^{l_j-1}\sum_{s=1}^{l_k-1}[\tilde{\Phi}_2((\Delta_{jr},\Delta_{ks}),(\Delta_{j(r+1)},\Delta_{k(s+1)};\rho)\tilde{\Phi}_2((-\infty,-\infty),(\Delta_{jr},\Delta_{ks});\rho)- \\
&\quad \tilde{\Phi}_2((\Delta_{jr},\Delta_{k(s-1)}),(\Delta_{j(r+1)},\Delta_{ks};\rho)\tilde{\Phi}_2((-\infty,-\infty),(\Delta_{jr},-\Delta_{ks});-\rho)]) \\
F_{\mathrm{ob}}(\rho;\Delta_j,\Delta_k) &= 2(\sum_{r=1}^{l_j-1}[\tilde{\Phi}_2((\Delta_{jr},-\infty),(\Delta_{j(r+1)},-\Delta_k);-\rho)\tilde{\Phi}_2((-\infty,-\infty),(\Delta_{jr},\Delta_k);\rho)- \\
&\quad \tilde{\Phi}_2((\Delta_{jr},-\infty),(\Delta_{j(r+1)},\Delta_k);\rho)\tilde{\Phi}_2((-\infty,-\infty),(\Delta_{jr},-\Delta_k);-\rho)]) \\
F_{\mathrm{to}}(\rho;\Delta_j,\Delta_k) &= 2(\sum_{r=1}^{l_k-1}[\tilde{\Phi}_2((\Delta_{kr},-\infty),(\Delta_{k(r+1)},-\Delta_j);-\rho)\tilde{\Phi}_2((-\infty,-\infty),(\Delta_{kr},\Delta_j);\rho)+ \\
&\quad \tilde{\Phi}_4((\Delta_{kr},-\infty,-\infty,-\infty),(\Delta_{k(r+1)},\Delta_{kr},-\Delta_j,0);S_{5a}(\rho))- \\
&\quad \tilde{\Phi}_2((\Delta_{k(r-1)},-\infty),(\Delta_{kr},-\Delta_j);-\rho)\tilde{\Phi}_2((-\infty,-\infty),(-\Delta_{kr},\Delta_j);-\rho)- \\
&\quad \tilde{\Phi}_4((\Delta_{k(r-1)},-\infty,-\infty,-\infty),(\Delta_{kr},-\Delta_{kr},-\Delta_j,0);S_{5b}(\rho))])
\end{aligned}
\tag{2.4}
$$

with

$$S_{3a}(\rho) = \begin{pmatrix} 1 & -\rho & 1/\sqrt{2} \\ -\rho & 1 & -\rho/\sqrt{2} \\ 1/\sqrt{2} & -\rho/\sqrt{2} & 1 \end{pmatrix}, \quad S_{3b}(\rho) = \begin{pmatrix} 1 & 0 & -1/\sqrt{2} \\ 0 & 1 & -\rho/\sqrt{2} \\ -1/\sqrt{2} & -\rho/\sqrt{2} & 1 \end{pmatrix},$$

$$S_{3}(\rho) = \begin{pmatrix} 1 & 1/\sqrt{2} & \rho/\sqrt{2} \\ 1/\sqrt{2} & 1 & \rho \\ \rho/\sqrt{2} & \rho & 1 \end{pmatrix}, \quad S_{4a}(\rho) = \begin{pmatrix} 1 & 0 & 1/\sqrt{2} & -\rho/\sqrt{2} \\ 0 & 1 & -\rho/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -\rho/\sqrt{2} & 1 & -\rho \\ -\rho/\sqrt{2} & 1/\sqrt{2} & -\rho & 1 \end{pmatrix}$$

$$S_{4b}(\rho) = \begin{pmatrix} 1 & \rho & 1/\sqrt{2} & \rho/\sqrt{2} \\ \rho & 1 & \rho/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & \rho/\sqrt{2} & 1 & \rho \\ \rho/\sqrt{2} & 1/\sqrt{2} & \rho & 1 \end{pmatrix}, \quad S_{5a}(\rho) = \begin{pmatrix} 1 & 0 & 0 & -\frac{\rho}{\sqrt{2}} \\ 0 & 1 & -\rho & \frac{\rho}{\sqrt{2}} \\ 0 & -\rho & 1 & -\frac{1}{\sqrt{2}} \\ -\frac{\rho}{\sqrt{2}} & \frac{\rho}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 1 \end{pmatrix}$$

$$S_{5b}(\rho) = \begin{pmatrix} 1 & 0 & 0 & -\frac{\rho}{\sqrt{2}} \\ 0 & 1 & \rho & -\frac{\rho}{\sqrt{2}} \\ 0 & \rho & 1 & -\frac{1}{\sqrt{2}} \\ -\frac{\rho}{\sqrt{2}} & -\frac{\rho}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 1 \end{pmatrix}$$

where, Φ denotes the cdf of univariate standard normal, $\Phi_d(\ldots, S)$ denotes the cdf of d-variate standard normal with correlation matrix S, $\tilde{\Phi}_2((a,b),(c,d),\rho)$ denotes the probability of the rectangle $\{(u,v) : a < u < b, c < v < d\}$ for a standard bivariate normal with correlation ρ and $\tilde{\Phi}_4(a,b,c,d,S) = P(Z_1 < a, Z_2 < b, Z_3 < c, Z_4 < d)$ denotes the distribution function of a standard quadrivariate normal $Z = (Z_1, Z_2, Z_3, Z_4)$ with correlation matrix S.

Proof. The derivation of $F_{cc}, F_{bb}, F_{cb}, F_{tb}, F_{ct}, F_{tt}, F_{co}$ has been previously done in literature as reported Table 2.1. We provide novel derivations of F_{oo}, F_{ob}, F_{to} in Supplement S1. To the best of our knowledge, this theorem is the first result deriving analytical forms of pairwise bridging functions for ordinal-ordinal,

ordinal-binary and ordinal-continuous pair for ordinal variables with arbitrary levels. □

Theorem 2.2.1 shows that the bridging function depends on the cutoffs for binary, ordinal and truncated variables. Hence, we need to estimate these cutoffs. From the observed data, we estimate the cutoffs through the method of moments as follows:

$$\begin{aligned}
\text{Binary: } \widehat{\Delta}_j &= \Phi^{-1}\left(\frac{\sum_{i=1}^{n} I(X_{ij}=0)}{n}\right) \\
\text{Ordinal: } \widehat{\Delta}_{jr} &= \Phi^{-1}\left(\frac{\sum_{i=1}^{n} I(X_{ij}<=(r-1))}{n}\right), r=1,\ldots,l_j-1 \\
\text{Truncated: } \widehat{\Delta}_j &= \Phi^{-1}\left(\frac{\sum_{i=1}^{n} I(X_{ij}=0)}{n}\right)
\end{aligned} \tag{2.5}$$

Now, we can plug-in estimated cutoffs in the bridging functions from (2.4), so the bridging functions now only depend on latent correlations. After bridging, the correlation matrix formed by the bridged estimates $\hat{\Sigma} = (\hat{\Sigma}_{jk})$ is not guaranteed to be positive semi-definite. So, we need to perform an extra step and for the estimated matrix find the nearest positive-definite correlation matrix (Higham, 2002). In Algorithm 1, we lay out all steps our our estimation procedure for all four mixed data types.

Algorithm 1 GLNPN estimation algorithm

Input: Observed data, $X_i = (X_{ic}, X_{it}, X_{io}, X_{ib}), i = 1, \ldots, n$

Phase 1 – Estimating cutoffs

for j in $\{t, o, b\}$ **do**
    Estimate the set of cutoffs $\widehat{\Delta}_j$ from (2.5) and store them
end for

Phase 2 – Inverting bridging functions

for j in $\{c, t, o, b\}$ **do**
    for $k \neq j$ **do**
        Calculate sample Kendall's Tau: $\widehat{\tau}_{jk}$
        Get the appropriate bridging function F_{jk} and plug-in the estimated cutoffs
        Obtain $\widehat{\Sigma}_{jk} = F_{jk}^{-1}(\widehat{\tau}_{jk}) = \text{argmin}_{\rho \in (-1,1)}(F_{jk}(\rho) - \widehat{\tau}_{jk})^2$
    end for
end for

Phase 3 – Getting nearest PD correlation matrix

Get the initial estimate of the latent correlation matrix Σ as follows:

$$\widehat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{cc} & \hat{\Sigma}_{ct} & \hat{\Sigma}_{co} & \hat{\Sigma}_{cb} \\ \hat{\Sigma}_{tc} & \hat{\Sigma}_{tt} & \hat{\Sigma}_{to} & \hat{\Sigma}_{tb} \\ \hat{\Sigma}_{oc} & \hat{\Sigma}_{ot} & \hat{\Sigma}_{oo} & \hat{\Sigma}_{ob} \\ \hat{\Sigma}_{bc} & \hat{\Sigma}_{bt} & \hat{\Sigma}_{bo} & \hat{\Sigma}_{bb} \end{pmatrix}$$

Use *nearPD* (Higham, 2002) function in R to find the nearest positive definite correlation matrix of $\widehat{\Sigma}$ as our final estimate.

2.3 Semiparametric Gaussian Copula Regression Model

In this section, we introduce Semiparametric Gaussian Copula Regression Model (SGCRM) and compare it with the traditional regression framework.

A classical regression model for a continuous outcome Y_i is typically written as

$$Y_i = \sum_{j=1}^{p} X_{ji}\beta_j + \epsilon_i, i = 1, \cdots, n \tag{2.6}$$

The simplest for understanding case is when both the outcome and all covariates are standard normal random variables. In that case, the simple linear regression conceptually assumes that both outcomes and predictors are on the same additive scale and tries to explain the variability of an outcome via variability of predictors. Various transformations of outcome/predictors can be used to deal with possible deviations from normality and symmetry. When outcome is not continuous alternative models such as probit, truncated regression, and other probit-like models have been proposed. However, they often loose the interpretability appeal of a simple linear regression model.

Semiparametric Gaussian Copula Regression for Mixed Data (SGCRM) can be seen as an alternative to the simple linear regression that deals with mixed types of outcomes and predictors by operating and connecting underlying continuous normal latent variables that generate observed mixed types variables. In this section, we introduce SGCRM and establish key asymptotical results for the estimates of the regression parameters. We then discuss the main advantages of SGCRM.

First, we define Semiparametric Gaussian Copula Regression for Mixed Data as follows.

$$\begin{cases} \text{Observed variables:} & (Y_1, \mathbf{X}_1), \ldots, (Y_n, \mathbf{X}_n) \overset{i.i.d}{\sim} GLNPN_{p+1}(0, \Sigma, (f_Y, f_X), \boldsymbol{\Delta}) \\ \text{Latent variables:} & (Z_1^Y, \mathbf{Z}_1^X), \ldots, (Z_n^Y, \mathbf{Z}_n^X) \quad \overset{i.i.d}{\sim} NPN_{p+1}(0, \Sigma, (f_X, f_Y)) \\ \text{SGCRM for latent variables:} & f_Y(Z_i^Y) = \sum_{k \in \{c,t,o,b\}} \sum_{j=1}^{p_k} f_X(Z_{kji}^X)\beta_{kj} + \epsilon_i, i = 1, \ldots, n. \end{cases} \tag{2.7}$$

Essentially, SGCRM is a simple linear regression for the outcome $f_Y(Z_i^Y)$ and predictors $f_X(Z_i^X)$, which, according to GLNPN, are jointly normal: $(f_Y(Z_i^Y), f_X(Z_i^X)) \overset{i.i.d}{\sim} N_{p+1}(0, \Sigma)$ with the correlation matrix Σ assuming the following partition:

$$\begin{bmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{YX} & \Sigma_{XX} \end{bmatrix}.$$

In SGCRM, we also assume that ϵ_i are i.i.d. from $N(0, 1 - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY})$.

It immediately follows that the regression coefficient $\boldsymbol{\beta} = \Sigma_{XX}^{-1}\Sigma_{XY}$. To estimate $\boldsymbol{\beta}$, we propose to use the estimate of Σ obtained via bridging as described in Section 2.2.1. Let $\hat{\Sigma}_n$ be the estimated latent correlation matrix for the model (2.7). Then, the estimates of the regression coefficient is given by $\hat{\beta}_n = \hat{\Sigma}_{n_{XX}}^{-1}\hat{\Sigma}_{n_{XY}}$.

In the next theorem, we derive asymptotic properties of both the estimator of latent correlation matrix and the regression parameter of SGCRM model. To formulate the theorem, we will need the following notations: let $vec(A)$ and $vecl(A)$ denote the vectorized matrix A and vector of lower-triangular elements of matrix A, respectively. Thus, $vecl(\hat{\Sigma}_n)$ and $vec(\hat{\Sigma}_n)$ are vectors of length $\frac{p(p-1)}{2}$ and p^2, respectively.

Theorem 2.3.1. *Suppose the following assumptions (Eicker, 1963) hold true: (i)*

the rank of $P_n = \hat{\Sigma}_{nXX}$ *is* p *and (ii)* $\frac{n\lambda_{min}(P_n)}{(\lambda_{max}(P_n))^2} \to 0$ *as* $n \to \infty$ *where* $\lambda_{min}(\cdot)$ *and* $\lambda_{max}(\cdot)$ *denote the smallest and largest eigenvalues of a matrix respectively. Then,* $\sqrt{n}(vecl(\hat{\Sigma}_n) - vecl(\Sigma))$ *is asymptotically normal with mean-vector* 0 *and a variance-covariance matrix* V_Σ. $\sqrt{n}(\hat{\beta}_n - \beta)$ *is asymptotically normal with mean* 0 *and a variance-covariance matrix* V_β.

Proof. Here, we layout the key ideas of the proof. First, using asymptotics of U-statistics in Hoeffding, 1992 and El Maache and Lepage, 2003, we establish the asymptotic normality of the Kendall's Tau estimates. Since the latent correlations are deterministic function (inverse bridging function) of the Kendall's Tau correlations, we use Delta method to obtain the asymptotic normality of the latent correlations. Next, we project the latent correlations onto a space of independent parameters (Archakova and Hansen, 2018), so that we can apply Delta method to obtain the asymptotic normality of the SGCRM regression coefficient. The regularity assumptions ensure the stability of the transformation $\hat{\beta}_n = \hat{\Sigma}_{n_{XX}}^{-1}\hat{\Sigma}_{n_{XY}}$ of the latent correlation matrix, so that we can apply Delta method. The detailed proof and analytical expressions of V_β and V_Σ are provided in Supplement S1. □

As part of the derivations, we solve a non-trivial computational problem by developing an efficient way of computing of the asymptotic covaraince of Kendall's Tau matrix. Our approach requires $O(n^2)$ FLOPs compared to the $O(n^4)$ FLOPs using naive brute-force approach. This reduction in computational burden enables us to calculate the asymptotic variance for moderate-to-large n.

2.3.1 Advantages of SGCRM

In this section, we present main differences between SGCRM model over traditional regression models developed for mixed type outcomes. These two approaches are contrasted in Table 2.2.

Aspect	**Traditional models (Observed space)**	**SGCRM (Latent space)**
Conditional associations	Use simple linear regression and probit-like regressions (probit, truncated, and ordinal probit). *Goodness of fit measure*: AUC or deviance (depending on defined model) Can be used to test for conditional independence for only Gaussian variables	Global model defines mutually consistent conditional models for all outcomes. (See Section 2.3) *Goodness of fit measure* Latent R-square Can be used to define a test for conditional independence for mixed type of variables.
Estimation	Requires likelihood computation, can be computationally infeasible for certain models. Non-robust but efficient.	Method of moments approach makes the estimation computationally efficient. Kendall's Tau rank correlation maintains the perfect balance between robust and efficiency.
Scaling	Need to manually normalize the variables to take into account heterogeneous scales. Maybe impossible for mixed types.	Inherent model assumptions take care of the scaling naturally.
Distributional assumptions	Parametric; convenient but limited.	Semi-parametric; allowing us to explain more general associations.

Aspect	Traditional models (Observed space)	SGCRM (Latent space)
Missing data imputation	Imputation by mean or restricted to complete cases.	Using latent correlation to impute missing data under missing-at-random assumption.
Interpretation	Can be interpreted on absolute scale and simplified. The signs of coefficients will denote the direction of association.	Can be interpreted on quantile scale. The signs of coefficients will denote the direction of association.
Prediction	Using model construction.	Using latent correlation and conditional expectation to construct the best linear unbiased predictors.

Table 2.2: Comparison between traditional approaches to model mixed data and Semi-parametric Gaussian Copula Regression Modeling

2.4 Methodological Applications of SGCRM

2.4.1 Latent variable predictions

Although, latent variables are not needed to estimate the regression parameter of SGCRM, other applications of SGCRM may require latent variables. To address this, we follow the ideas from Best Linear Unbiased Predictor (BLUPs) in mixed effect modelling and use a similar conditional expectation approach to find best predictors of latent variables conditionally on observed variables. Note that at this point we do not make a distinction between an outcome and predictors. We also drop sub-index i, as we only use participant-specific observed variables when we predict their latent representations. We introduce additional notations:

- $L = (L_c, L_t, L_o, L_b) = f(Z_c, Z_t, Z_o, Z_b)$, where f is a vector of coordinate wise monotone transformations as described in the definition of GLNPN;
- $L_{-c} = (L_t, L_o, L_b)$ and similarly for all other combinations of sub-indexes;
- ct denotes the union of continuous and truncated indices. $L_{-ct} = (L_o, L_b)$;
- $\Sigma_{a,a}$ indicates the sub-matrix of Σ with indices running over the set a;
- $\Sigma_{a,-a}$ denotes the rows of Σ; indexed by the set a but without the columns indexed by a and $\Sigma_{-a,a} = \Sigma'_{a,-a}$;
- $\Sigma_{-a,-a}$ indicates the sub-matrix of Σ with indices not in the set a.

To calculate $E(L|X) = E(L_c, L_t, L_o, L_b|X_c, X_o, X_t, X_b)$, we will consider two cases: Case 1 when $X_t = 0$ and Case 2 when $X_t > 0$.

Case $X_t = 0$**:**. We can observe that for continuous variables $L_c = f_c(X_c)$ and the values of X_t, X_o, X_b will restrict each coordinate of $L_{-c} = (L_t, L_o, L_b)$ to be in a certain interval based on the cutoffs.

That is, under our model assumptions $\{X_t = x_t, X_b = x_b, X_o = x_o\} \iff \{L_{-c} \in B\}$, where $B = \{\times_{\lambda \notin c} B_\lambda\}$ and $\times$ indicates Cartesian product and B_λ denotes an interval in $\mathbb{R}$ for the corresponding co-ordinate.

By using the fact that

$$L_{-c|c} = L_{-c}|L_c \sim N(\Sigma_{-c,c}\Sigma_{c,c}^{-1}L_c, \Sigma_{-c,-c} - \Sigma_{-c,c}\Sigma_{c,c}^{-1}\Sigma_{c,-c})$$

we can derive the following

$$
\begin{aligned}
E(L_c|X_c, X_t, X_o, X_b) &= f_c(X_c) \\
E((L_t, L_o, L_b)|X_c, X_t, X_o, X_b) &= E(L_{-c}|X_c, X_t, X_o, X_b) \\
&= E(L_{-c}|L_c, L_{-c} \in B_{-c})) \\
&= E(L_{-c|c}|L_{-c|c} \in B_{-c})
\end{aligned}
\tag{2.8}
$$

The last quantity in equation(2.8) is exactly the expectation of a multivariate normal random variable truncated in the set B. Thus, we get

$$
E(L|X_c, X_t, X_o, X_b) = (f_c(X_c), E(L_{-c|c}|L_{-c|c} \in B_{-c})) \tag{2.9}
$$

Case $X_t > 0$**:** Observe that $L_c = f_c(X_c), L_t = f_t(X_t)$ and the values of X_o, X_b will restrict each co-ordinate of $L_{-ct} = (L_o, L_b)$ to be in a certain interval based on the cutoffs. Under our model assumptions

$$
\{X_b = x_b, X_c = x_c\} \iff \{L_{-ct} \in B_{-ct}\}
$$

, where $B_{-ct} = \{\times_{\lambda \notin ct} B_\lambda\}$ and $\times$ indicates Cartesian product and B_λ denotes an interval in $\mathbb{R}$ for the corresponding co-ordinate. We also use the fact that

$$
L_{-ct|ct} = L_{-ct}|L_{ct} \sim N(\Sigma_{-ct,ct}\Sigma_{ct,ct}^{-1}L_{ct}, \Sigma_{-ct,-ct} - \Sigma_{-ct,ct}\Sigma_{ct,ct}^{-1}\Sigma_{ct,-ct})
$$

Using information above, we can derive the following results -

$$\begin{aligned} E(L_{ct}|X_c, X_t, X_o, X_b) &= f_{ct}(X_{ct}) \\ E((L_o, L_b)|X_c, X_t, X_o, X_b) &= E(L_{-ct}|X_c, X_t, X_o, X_b) \\ &= E(L_{-ct}|L_{ct}, L_{-ct} \in B_{-ct})) \\ &= E(L_{-ct|ct}|L_{-ct|ct} \in B_{-ct}) \end{aligned} \tag{2.10}$$

The last quantity in equation (2.10) is exactly the expectation of a multivariate normal random variable truncated in the set B. Thus, we get

$$E(L|X_c, X_t, X_o, X_b) = (f_c(X_c), f_t(X_t), E(L_{-ct|ct}|L_{-ct|ct} \in B_{-ct})) \tag{2.11}$$

To get exact values from equations (2.9) and (2.11), we need to know three things: (a) the functions f_c (over entire domain), f_t (only for non-zero values), (b) the sets B_{-c}, B_{-ct}, and (c) a way to calculate the expectation of truncated multivariate normal random variable. Below, we show how to derive these three.

(a) We illustrate this step by considering a single continuous and a single truncated variable. First, we get an empirical CDF estimates as follows

$$\begin{aligned} F_{cn}(x) &= \frac{1}{n+1} \sum_{i=1}^{n} I(X_{ci} \leqslant x), x \in \mathbb{R} \\ F_{tn}(x) &= \frac{1}{n+1} \sum_{i=1}^{n} I(X_{ti} \leqslant x), x > 0 \end{aligned} \tag{2.12}$$

Then, we use equation (2.12) to construct the estimator of monotone

transformations as follows

$$\begin{aligned} \hat{f}_c(x) &= \Phi^{-1}(F_{cn}(x)) \\ \hat{f}_t(x) &= \Phi^{-1}(F_{tn}(x)), \end{aligned} \tag{2.13}$$

where Φ is the standard normal CDF. This follows the approach for continuous variables discussed in Section 4 of Liu, Lafferty, and Wasserman, 2009.

(b) We plugin the method of moments estimates for cutoffs from Section 2.2.1 to get $\hat{B}_{-c}$ and $\hat{B}_{-ct}$.

(c) To calculate the first moment of a truncated multivariate normal distribution, we use ideas from Wilhelm and Manjunath, 2010. They proposed a recursive formula to calculate the moment generating function of a truncated multivariate normal distribution and then get the first derivative at 0 to calculate the desired expectation. We use their method implemented in R software package *tmvtnorm* (Wilhelm and G, 2015).

It is important to note the prediction of latent variables described above can be done subject-by-subject in an embarrassingly parallel way to reduce computational burden.

2.4.2 Missing data imputation

GLNPN framework provides a readily available way to perform imputation of missing mixed data observations using the same techniques as for prediction of latent variables.

Suppose we have missing observations for a particular subject. We split

the full vector X into observed and missing parts as $X = (X_O, X_M)$, where O denotes observed and M denotes missing parts and subject-specific index i has been omitted for notational simplicity. First we predict $E[L_M|X_O]$ and then obtain the prediction of X_M using an appropriate transformation-then-truncation step applied to $E[L_M|X_O]$. Remember that

$$L_{M|O} = L_M|L_O \sim N(\Sigma_{M,O}\Sigma_{O,O}^{-1}L_O, \Sigma_{M,M} - \Sigma_{M,O}\Sigma_{O,O}^{-1}\Sigma_{O,M}) \tag{2.14}$$

As X_O is $\sigma(L_O)$-measurable random variable, where $\sigma(L_O)$ denotes the σ-algebra generated by L_O, we can use the tower property of conditional expectations to get the following identity

$$E[L_M|X_O] = E[E[L_M|L_O]|X_O] = E[\Sigma_{M,O}\Sigma_{O,O}^{-1}L_O|X_O] = \Sigma_{M,O}\Sigma_{O,O}^{-1}E[L_O|X_O] \tag{2.15}$$

Finally, we can calculate $E[L_O|X_O]$ from the equation above using the same steps as in previous section.

2.5 Simulation

We conduct a series of simulation experiments to evaluate the performance of our approach. The data generation algorithm for the simulation experiments is presented below.

1. Generate a random correlation matrix Σ using the random partial correlation method in Joe, 2006. We calculate the condition number of Σ and if the number is below 10, we proceed to Step 2. The additional step of checking the condition number is to ensure the stability of matrix

inversion and our regression estimates.

2. We generate $n = 1000$ replicates of 8-variate latent normal variable from the following model

$$(L_{i1}, L_{i2}, L_{i3}, L_{i4}, L_{i5}, L_{i6}, L_{i7}, L_{i8}) \sim N(0, \Sigma), \quad i = 1, 2, \cdots, n$$

3. We then apply the cutoffs from Table 2.3 to generated latent variables from previous step to obtain observed binary (X_1, X_3), continuous (X_2), ordinal (X_4, X_5, X_6, X_7) and truncated (X_8) variables. We consider ordinal variables with 3 categories (X_4, X_5) and 4 categories (X_6, X_7). We vary the entropy of our binary and ordinal variables. The entropy of a discrete random variable is defined as $\sum_i p_i log(p_i)$, where p_i is the probability of the i-th distinct value. The entropy indicates the average level of information contained in the variable's possible outcomes. Varying entropy enables us to consider the performance of our approach across different levels of information.

4. We perform Steps $1 - 3$ for 200 different seeds to replicate our experiment 200 times.

For regression modeling, we treat X_1 (the binary variable with high entropy) as our outcome. We use methods described in Section 2.2.1 to estimate the latent correlation matrix and regression coefficients for each instance of the simulated data. We also calculate the asymptotic confidence intervals of our estimates from Theorem 2.3.1. Finally, for every instance of a simulated correlation matrix, we perform 500 replicates of our experiment to get an

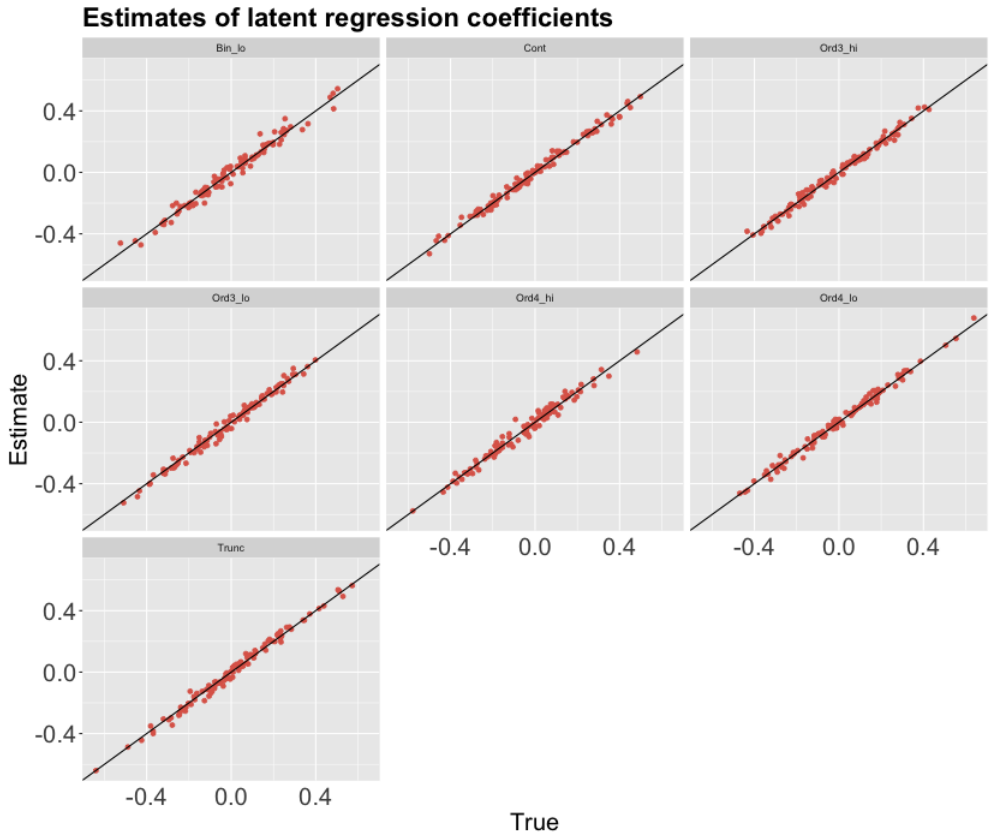


Figure 2.3: The estimates of latent regression coefficients over different simulation scenarios. The black line denotes $y = x$ line

empirical distribution of our estimated parameters. We calculate coverage of these 500 estimates against the asymptotic confidence intervals to gauge the accuracy of the asymptotic intervals.

Figure 2.3 shows the estimates of latent regression coefficients against the true values. We observe that across different combinations, the estimated and true parameters are very well aligned along the diagonal line. We also report the coverage of our proposed asymptotic confidence interval for regression coefficients (Fig. S1). The median coverage (across 100 seeds) of SGCRM regression coefficients is slightly below the expected 95% line. We expect this

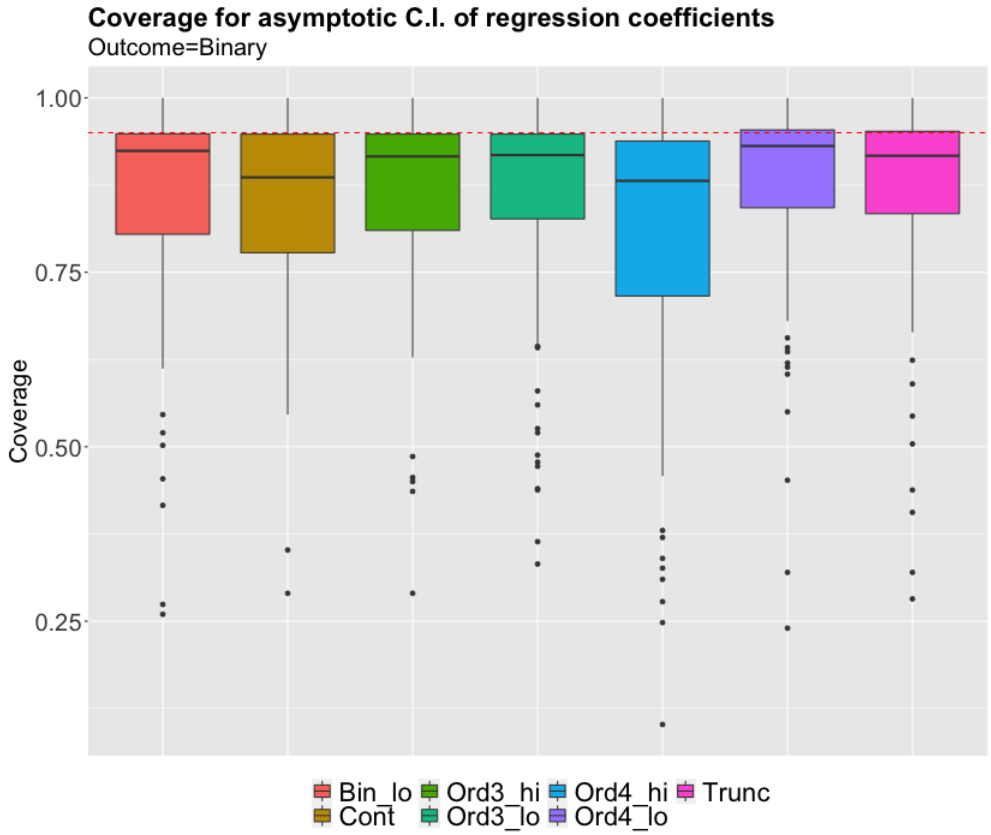


Figure 2.4: The coverage of 95% asymptotic confidence interval for SGCRM regression coefficients. The red dotted line corresponds to the 0.95 coverage.

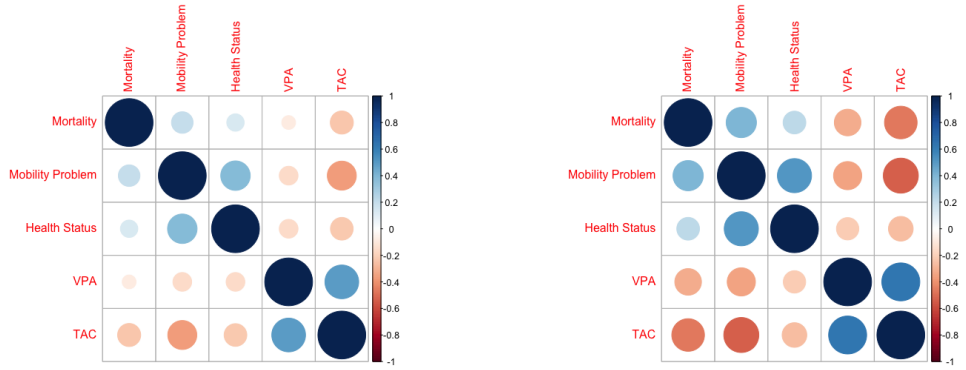
undercoverage as we do not consider the estimated cutoffs' uncertainty in calculating the asymptotic variances. Accounting for additional uncertainty from the use of plug-in cutoff estimators would make the calculations analytically complex with a small practical gain.

2.6 NHANES 2003-2006

Our method is illustrated in National Health and Nutrition Examination Survey (NHANES) 2003 – 2006. We focus on five variables discussed in

Table 2.3: Mixed GLNPN variables

Variable number	Variable type	Cutpoint(s)
X_1	Binary (high entropy)	0.3
X_2	Continuous	Not applicable
X_3	Binary (low entropy)	1
X_4	Ordinal (3 categories and high entropy)	$(-0.1, 0.6)$
X_5	Ordinal (3 categories and low entropy)	$(-1, 1)$
X_6	Ordinal (4 categories and high entropy)	$(-0.7, 0.1, 0.6)$
X_7	Ordinal (4 categories and low entropy)	$(-0.3, 0.1, 0.2)$
X_8	Truncated	0



(a) Pearson correlation matrix **(b)** GLNPN latent correlation matrix

Figure 2.5: The estimated 5×5 correlation matrices of our variables from NHANES $2003 - 04$ and $2005 - 06$

Introduction: TAC, VPA, mortality, Health Status, and Mobility Problem. For the analysis, we excluded participants who (1) have missing mortality information or alive with follow-up less than 5 years, (2) are younger than 50 years old or aged 85 and older, (3) have missing any of the above-mentioned variables of interest, (4) have died due to accident, and (5) had fewer than 3 valid accelerometry days (a valid day is defined as a day with at least 10 hours of wear time) (Leroux et al., 2019). The final analytical sample consisted of 3069 subjects with 313 deaths within 5 years.

Figure 2.5 compares estimated Pearson and latent correlation matrices (numerical values are in Tables S1 and S2). We observe that the latent correlation matrix detects stronger correlation between variables compared to naively interpreting Pearson correlations for mixed types variables. For example, the correlation between Mortality and Mobility Problem increases from 0.2 to 0.39, the correlation between Mortality and VPA increases from -0.08 to -0.30.

Table 2.4: Comparison of simple linear model and SGCRM results for continuous outcome

$TAC \sim MobilityProblem + HealthStatus$				
	Simple linear regression		SGCRM	
	Covariate	Coefficients	Covariate	Coefficients
1	Mobility Problem (1)	-0.342 (-0.378, -0.306)	Mobility Problem	-0.543 (-0.594, -0.492)
2	Health Status (2)	-0.073 (-0.134, -0.013)	Health Status	0.011 (-0.036, 0.059)
3	Health Status (3)	-0.152 (-0.211, -0.094)		NA
4	Health Status (4)	-0.153 (-0.217, -0.09)		NA
5	Health Status (5)	-0.181 (-0.271, -0.09)		NA

Table 2.5: Comparison of truncated Gaussian regression and SGCRM results for truncated outcome

$VPA \sim MobilityProblem + HealthStatus$				
	Truncated Gaussian regression		SGCRM	
	Covariate	Coefficients	Covariate	Coefficients
1	Mobility Problem (1)	-629.875 (-753.338, -506.412)	Mobility Problem	-0.32 (-0.376, -0.265)
2	Health Status (2)	-63.201 (-88.362, -38.041)	Health Status	-0.055 (-0.108, -0.004)
3	Health Status (3)	-194.491 (-239.269, -149.714)		NA
4	Health Status (4)	-195.522 (-248.435, -142.608)		NA
5	Health Status (5)	-285.281 (-403.648, -166.914)		NA

Table 2.6: Comparison of probit ordinal regression and SGCRM results for ordinal outcome

$HealthStatus \sim MobilityProblem + VPA$				
	Probit ordinal regression		SGCRM	
	Covariate	Coefficients	Covariate	Coefficients
1	Mobility Problem (1)	0.874 (0.79, 0.959)	Mobility Problem	0.494 (0.45, 0.537)
2	I(VPA == 0)	0.12 (0.04, 0.199)		NA
3	VPA	-0.017 (-0.024, -0.011)	VPA	-0.047 (-0.091, -0.003)

Table 2.7: Comparison of probit regression and SGCRM results for binary outcome

$Mortality \sim MobilityProblem + HealthStatus + TAC$				
	Probit regression		SGCRM	
	Covariate	Coefficients	Covariate	Coefficients
1	Mobility Problem (1)	0.335 (0.192, 0.478)	Mobility Problem	0.195 (0.097, 0.297)
2	Health Status (2)	0.086 (-0.212, 0.402)	Health Status	0.033 (-0.042, 0.103)
3	Health Status (3)	0.244 (-0.038, 0.547)		NA
4	Health Status (4)	0.195 (-0.102, 0.511)		NA
5	Health Status (5)	0.455 (0.094, 0.826)		NA
6	TAC	$-3.952 \times 10^{-6}(-4.752 \times 10^{-6}, -3.174 * 10^{-6})$	TAC	-0.352 (-0.427, -0.272)

After estimation of the GLNPN latent correlation matrix, we next fit four mutually consistent conditional SGCRM models by treating one of the mixed types variables as outcome and some of the others as predictors. Specifically, we will consider one outcome for each type: TAC (continuous), VPA (truncated), Health Status (ordinal), and Mortality (binary) will be outcomes. We compare SGCRM models with the traditional counterparts such as simple linear regression, truncated regression, ordinal probit and probit regressions in Tables 2.4, 2.5, 2.6, 2.7, respectively. Both SGCRM and traditional estimates are reported with 95% confidence intervals. We want to compare the direction and significance of conditional associations captured by SGCRM and traditional models.

We start with a continuous outcome, TAC. Table 2.4 contrasts the results of the two models. We observe that Mobility Problem has a significant negative effect on TAC in both SGCRM ($-0.543(-0.594, -0.492)$) and the linear model ($-0.342(-0.378, -0.306)$). Furthermore, different levels of reported health status have a significant negative effect on TAC in the simple linear model, but when the ordinal categories of Health Status are represented via corresponding GLNPN latent variable we do not observe significant association with TAC in SGCRM model. This is one obvious disadvantage of our approach for ordinal

variables. If the effect is not present across all levels, when collapsing all levels may potentially loose partial significance, as we observe here. We will discuss this more in Discussion.

Next, we treat truncated variable VPA as the outcome in SGCRM and compare SGCRM model vs Gaussian truncated regression model. The results are shown in Table 2.5. The direction and the significance of associations estimated by SGCRM are in agreement with those estimated by truncated regression. However, regression coefficients from truncated Gaussian regression model are not scaled and, hence, their magnitude cannot be compared. In contrast, SGCRM coefficients are normalized and can be compared across covariates. For example, we observe that estimated effect of mobility problem is much higher that that of health status: $-0.32(-0.376, -0.265)$ vs $-0.055(-0.108, -0.004)$.

Next, we model Health Status as an ordinal outcome with Mobility Problem and VPA as two covariates. Because of VPA is a trunctated variable, we represent VPA via two components in the traditional model: (1) an indicator variable of VPA being equal to 0, and (2) the VPA value itself. Note that representation is not needed in SGCRM. The results are shown in Table 2.6. VPA is negatively associated with a higher value (worse) Health Status both in probit ordinal model with the regression coefficient of $-0.017(-0.024, -0.011)$ and in SGCRM model with the regression coefficient of $-0.047(-0.091, -0.003)$. Mobility problem is significant and is positively associated with a higher value (worse) Health Status in both models. Again, SGCRM allows to compare the magnitude of estimated effects of Mobility Problem and VPA with a much

higher effect of the former.

Finally, we look at the 5-year mortality as a binary outcome with mobility problem, Health Status, and TAC as covariates. The results are shown in Table 2.7. We see that the highest (worst) Health Status level is signifact in probit regression, but when the ordinal categories of Health Status are represented via corresponding GLNPN latent variable we do not observe significant association with mortality in SGCRM model. Again, as we discussed above this is a limitation of SGCRM. We again get interpretable regression coefficients and we see that the effect of TAC is almost twice higher than of mobility problem.

It is also important to note that all four SGCRM models are mutually consistent in contrast to the set of four traditional regressions: simple linear, truncated, probit ordinal, and probit.

In the final step, we calculate predictions of latent variables using methods described in Section 2.4.1. Figure 2.6 shows the distribution of the predicted latent variables for the five variables. The figure also shows the interrelation between those variables on the off-diagonal blocks of the scatterplot matrix. We see that the distribution of predicted latent variables approximates the assumed latent normal distribution, but with multiple modes originated from the discontinuities of the distributions of observed variables. It is interesting to note that scatterplots of predicted latent variables for mortality vs. mobility problem, mortality vs TAC, and TAC vs mobility problem reveal linear patterns (with some discontinuities around the cutoffs). This particular observation would be harder, if possible, to make by visually exploring scatterplots of observed counterparts.

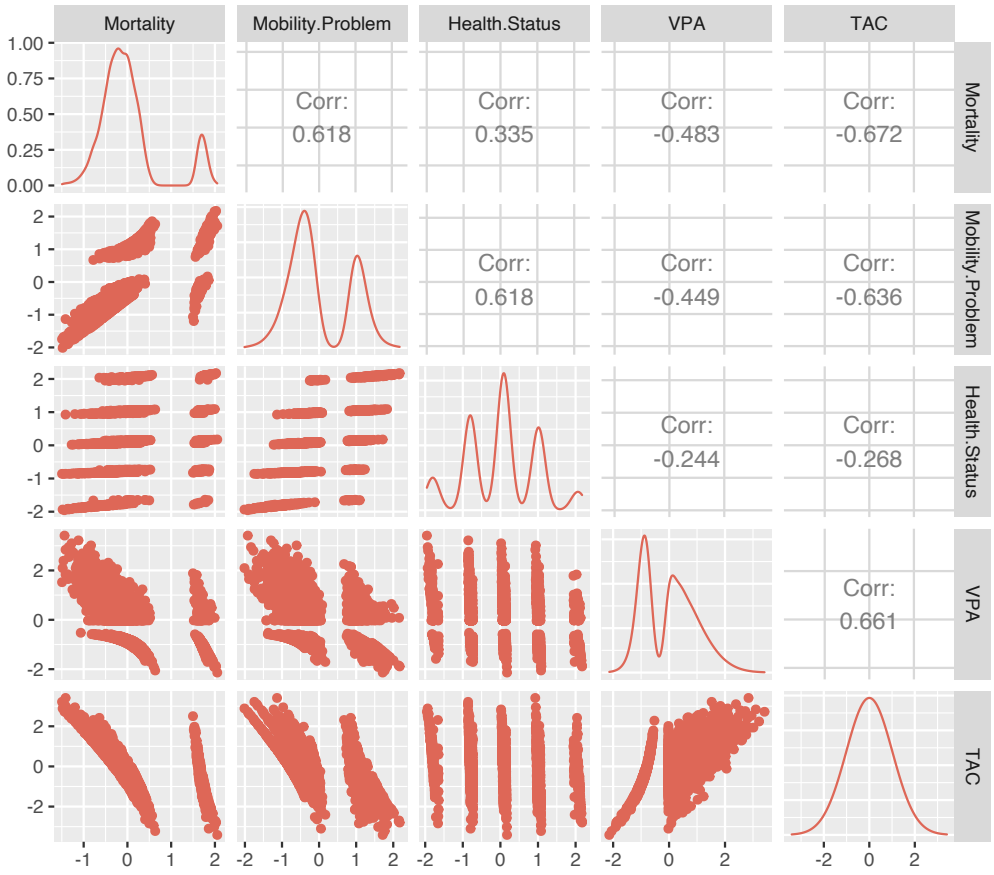


Figure 2.6: Predictions of latent variables in NHANES

2.7 Discussion

The main contribution of this paper is a joint modeling approach for mixed data types that builds on semiparametric Gaussian copula. The approach is scale-free, robust, and fast. Adapted to perform linear regression on the latent space, SGCRM provides mutually consistent conditional regression models as a unifying alternative to a range of popular conditional regression models such as simple linear regression and probit-like regressions including truncated regression, ordinal probit regression, probit and others. Our likelihood-free

approach is more computationally efficient than likelihood-based joint copula models. Finally, embedding the variables using a semiparametric Gaussian copula automatically normalizes the scale of all latent variables that results in standardized and more interpretable regression coefficients. The approach allows to define R^2 for all four types of outcomes. Finally, the approach allows to perform missing data imputation.

In NHANES application, we kept our models simple to carefully illustrate the approach and the interpretability of the results. In terms of computational complexity, our method needs to estimate $O(p^2)$ correlation parameters and the calculation of Kendall's Tau takes only $O(nlogn)$ FLOPs for each estimation. Hence, with the quadratic complexity in p our approach scales very well with respect of increasing p. Moreover, we propose a computationally efficient way of calculating the asymptotic variance-covariance matrix of the parameters in $O(n^2)$ FLOPs compared to the brute force approach of $O(n^4)$.

In terms of limitations, SGCRM is less flexible in dealing with multiple ordinal levels. For example, we can estimate two different effects for an ordinal variable with three categories. In comparison, our model works on the latent scale for the ordinal variable and assumes a uniform magnitude and direction of the effect. Compared to the traditional models, SGCRM also loses interpretability of original scales of covariates, because SGCRM coefficients are only interpretable at a latent scale.

As future work, it would be important to develop quantile scale interpretation of SGCRM regression results. SGCRM can also be adapted to handle survival outcomes. Moreover, it would be interesting to adapt GNPLN to

deal with functional and multi-level/longitudinal mixed data. As one of the methodological applications, we provided an algorithm for imputing missing observations and it would be very interesting to compare that approach with existing ones. Finally, predicted latent variables can be used within distance-based clustering approaches under mixed data settings as well as for dimension reduction of multivariate mixed data.

S1 Proofs

Proof of Theorem 2.2.1: Let X_j, X_k be ordinal with levels $\{0, 1, \cdots, l_j - 1\}$ and $\{0, 1, \cdots, l_k - 1\}$ respectively and (L_j, L_k) is the corresponding latent standard bivariate normal with correlation σ_{jk}. Then for two independent observations i, i' -

$$\begin{aligned} P(X_{ij} > X_{i'j}, X_{ik} > X_{i'k}) &= \sum_{r=1}^{l_j-1} \sum_{s=1}^{l_s-1} [P(X_{ij} = r, X_{ik} = s) P(X_{i'j} < r, X_{i'k} < s)] \\ &= \sum_{r=1}^{l_j-1} \sum_{s=1}^{l_s-1} [P(\Delta_{jr} \leqslant L_{ij} < \Delta_{(j+1)r}, \Delta_{ks} \leqslant L_{ik} < \Delta_{(k+1)s}) P(L_{i'j} < \Delta_{jr}, L_{i'k} < \Delta_{ks})] \\ &= \sum_{r=1}^{l_j-1} \sum_{s=1}^{l_k-1} [\tilde{\Phi}_2((\Delta_{jr}, \Delta_{ks}), (\Delta_{j(r+1)}, \Delta_{k(s+1)}; \Sigma_{jk}) \tilde{\Phi}_2((-\infty, -\infty), (\Delta_{jr}, \Delta_{ks}); \Sigma_{jk})] \end{aligned} \tag{S1}$$

Similarly,

$$
\begin{aligned}
P(X_{ij} > X_{i'j}, X_{ik} < X_{i'k}) &= \sum_{r=1}^{l_j-1} \sum_{s=1}^{l_s-1} [P(X_{ij} = r, X_{ik} = s-1) P(X_{i'j} < r, X_{i'k} > (s-1))] \\
&= \sum_{r=1}^{l_j-1} \sum_{s=1}^{l_s-1} [P(\Delta_{jr} \leqslant L_{ij} < \Delta_{(j+1)r}, \Delta_{(k-1)s} \leqslant L_{ik} < \Delta_{ks}) P(L_{i'j} < \Delta_{jr}, L_{i'k} > \Delta_{ks})] \\
&= \sum_{r=1}^{l_j-1} \sum_{s=1}^{l_s-1} [P(\Delta_{jr} \leqslant L_{ij} < \Delta_{(j+1)r}, \Delta_{(k-1)s} \leqslant L_{ik} < \Delta_{ks}) P(L_{i'j} < \Delta_{jr}, -L_{i'k} < -\Delta_{ks})] \\
&= \sum_{r=1}^{l_j-1} \sum_{s=1}^{l_k-1} [\tilde{\Phi}_2((\Delta_{jr}, \Delta_{k(s-1)}), (\Delta_{j(r+1)}, \Delta_{ks}; \Sigma_{jk}) \tilde{\Phi}_2((-\infty, -\infty), (\Delta_{jr}, -\Delta_{ks}); -\Sigma_{jk})]
\end{aligned}
\tag{S2}
$$

By symmetry, the population Kendall's Tau, τ_{jk} for X_j, X_k can be written as follows -

$$
\begin{aligned}
\tau_{jk} &= 2(P(X_{ij} > X_{i'j}, X_{ik} > X_{i'k}) - P(X_{ij} > X_{i'j}, X_{ik} < X_{i'k})) \\
&= 2(\sum_{r=1}^{l_j-1} \sum_{s=1}^{l_k-1} [\tilde{\Phi}_2((\Delta_{jr}, \Delta_{ks}), (\Delta_{j(r+1)}, \Delta_{k(s+1)}; \Sigma_{jk}) \tilde{\Phi}_2((-\infty, -\infty), (\Delta_{jr}, \Delta_{ks}); \Sigma_{jk}) \\
&\quad - \tilde{\Phi}_2((\Delta_{jr}, \Delta_{k(s-1)}), (\Delta_{j(r+1)}, \Delta_{ks}; \Sigma_{jk}) \tilde{\Phi}_2((-\infty, -\infty), (\Delta_{jr}, -\Delta_{ks}); -\Sigma_{jk})])
\end{aligned}
\tag{S3}
$$

Now, reducing the above calculations for $l_k = 2$, will yield the bridging function between a general ordinal and binary pairs.

Now, suppose we have a truncated variable X_m with cutoff Δ_m and corresponding latent normal variable L_m, then redoing the above calculations will

look like -

$$
\begin{aligned}
P(X_{ij} > X_{i'j}, X_{im} > X_{i'm}) &= \sum_{r=1}^{l_j-1}[P(X_{ij} = r, X_{im} > 0)P(X_{i'j} < r, X_{i'm} = 0) \\
&+ P(X_{ij} = r, X_{i'j} < r, X_{i'm} > 0, X_{im} - X_{i'm} > 0)] \\
&= \sum_{r=1}^{l_j-1}[P(\Delta_{jr} \leqslant L_{ij} < \Delta_{(j+1)r}, L_{im} > \Delta_m)P(L_{i'j} < \Delta_{jr}, L_{i'm} \leqslant \Delta_m) \\
&+ P(\Delta_{jr} \leqslant L_{ij} < \Delta_{(j+1)r}, L_{i'j} < \Delta_{jr}, L_{i'm} > \Delta_m, , \frac{L_{im} - L_{i'm}}{2} > 0)] \\
&= \sum_{r=1}^{l_j-1}[P(\Delta_{jr} \leqslant L_{ij} < \Delta_{(j+1)r}, -L_{im} < -\Delta_m)P(L_{i'j} < \Delta_{jr}, L_{i'm} \leqslant \Delta_m) \\
&+ P(\Delta_{jr} \leqslant L_{ij} < \Delta_{(j+1)r}, L_{i'j} < \Delta_{jr}, -L_{i'm} < -\Delta_m, , \frac{L_{i'm} - L_{im}}{\sqrt{2}} < 0)] \\
&= \sum_{r=1}^{l_j-1}[\tilde{\Phi}_2((\Delta_{jr}, -\infty), (\Delta_{j(r+1)}, -\Delta_m); -\rho)\tilde{\Phi}_2((-\infty, -\infty), (\Delta_{jr}, \Delta_m); \rho) \\
&+ \tilde{\Phi}_4((\Delta_j r, -\infty, -\infty, -\infty), (\Delta_{j(r+1)}, \Delta_{jr}, -\Delta_m, 0); S_{5a}(\rho))] \\
P(X_{ij} < X_{i'j}, X_{im} > X_{i'm}) &= \sum_{r=1}^{l_j-1}[P(X_{ij} = (r-1), X_{im} > 0)P(X_{i'j} > (r-1), X_{i'm} = 0) \\
&+ P(X_{ij} = r, X_{i'j} < r, X_{i'm} > 0, X_{im} - X_{i'm} > 0)] \\
&= \sum_{r=1}^{l_j-1}[P(\Delta_{j(r-1)} \leqslant L_{ij} < \Delta_{jr}, L_{im} > \Delta_m)P(L_{i'j} > \Delta_{jr}, L_{i'm} \leqslant \Delta_m) \\
&+ P(\Delta_{(j-1)r} \leqslant L_{ij} < \Delta_{jr}, L_{i'j} > \Delta_{jr}, L_{i'm} > \Delta_m, , \frac{L_{im} - L_{i'm}}{2} > 0)] \\
&= \sum_{r=1}^{l_j-1}[P(\Delta_{j(r-1)} \leqslant L_{ij} < \Delta_{jr}, -L_{im} < -\Delta_m)P(-L_{i'j} < -\Delta_{jr}, L_{i'm} \leqslant \Delta_m) \\
&+ P(\Delta_{(j-1)r} \leqslant L_{ij} < \Delta_{jr}, -L_{i'j} < -\Delta_{jr}, -L_{i'm} < -\Delta_m, , \frac{L_{i'm} - L_{im}}{2} < 0)] \\
&= \sum_{r=1}^{l_j-1}[\tilde{\Phi}_2((\Delta_{j(r-1)}, -\infty), (\Delta_{jr}, -\Delta_m); -\rho)\tilde{\Phi}_2((-\infty, -\infty), (-\Delta_{jr}, \Delta_m); -\rho) + \\
&\tilde{\Phi}_4((\Delta_{j(r-1)}, -\infty, -\infty, -\infty), (\Delta_{jr}, -\Delta_{jr}, -\Delta_m, 0); S_{5b}(\rho)]
\end{aligned}
\tag{S4}
$$

where, $S_{5a}(\rho) = cov(L_{ij}, L_{i'j}, -L_{i'm}, \frac{L_{i'm}-L_{im}}{\sqrt{2}}) = \begin{pmatrix} 1 & 0 & 0 & -\frac{\rho}{\sqrt{2}} \\ 0 & 1 & -\rho & \frac{\rho}{\sqrt{2}} \\ 0 & -\rho & 1 & -\frac{1}{\sqrt{2}} \\ -\frac{\rho}{\sqrt{2}} & \frac{\rho}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 1 \end{pmatrix}$

and

$S_{5b}(\rho) = cov(L_{ij}, -L_{i'j}, -L_{i'm}, \frac{L_{i'm}-L_{im}}{\sqrt{2}}) = \begin{pmatrix} 1 & 0 & 0 & -\frac{\rho}{\sqrt{2}} \\ 0 & 1 & \rho & -\frac{\rho}{\sqrt{2}} \\ 0 & \rho & 1 & -\frac{1}{\sqrt{2}} \\ -\frac{\rho}{\sqrt{2}} & -\frac{\rho}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 1 \end{pmatrix}$.

Hence, we get -

$F_{to}(\rho; \Delta_j, \Delta_m) = 2(\sum_{r=1}^{l_j-1}[\tilde{\Phi}_2((\Delta_{jr}, -\infty), (\Delta_{j(r+1)}, -\Delta_m); -\rho)\tilde{\Phi}_2((-\infty, -\infty), (\Delta_{jr}, \Delta_m); \rho)$
$+ \tilde{\Phi}_4((\Delta_j r, -\infty, -\infty, -\infty), (\Delta_{j(r+1)}, \Delta_{jr}, -\Delta_m, 0); S_{5a}(\rho)) -$
$\tilde{\Phi}_2((\Delta_{j(r-1)}, -\infty), (\Delta_{jr}, -\Delta_m); -\rho)\tilde{\Phi}_2((-\infty, -\infty), (-\Delta_{jr}, \Delta_m); -\rho) -$
$\tilde{\Phi}_4((\Delta_{j(r-1)}, -\infty, -\infty, -\infty), (\Delta_{jr}, -\Delta_{jr}, -\Delta_m, 0); S_{5b}(\rho)]$.

Proof of Theorem 2:

First, let's familiarize ourselves with some notations. For a $p \times p$ correlation matrix A, we can get singular-value decomposition of A as $A = Q\Lambda Q^T$, where Q is an orthonormal matrix and $\Lambda = diag(\lambda_1, \cdots, \lambda_p)$ are the eigenvalues of A. Let's define $logA = Qlog\Lambda Q^T$, where $log\Lambda = diag(log\lambda_1, \cdots, log\lambda_p)$.

First we need to state the results for the asymptotic variance of Kendall's Tau as calculated in Hoeffding, 1992 and El Maache and Lepage, 2003. Using the results of U-statistics asymptotics, we state the results in the Lemma S1.1 below.

Lemma S1.1. *Let K_n be the Kendall's Tau matrix estimated from the data, then*

$\sqrt{n}(vecl(K_n) - vecl(K))$ *is asymptotically normal with mean* 0 *and variance-covariance matrix* V_K*, where,* $K_{ij} = E(sgn((X_{i1} - X_{i2})(X_{j1} - X_{j2}))$ *and*

$$V_{K_{(ij),(kl)}} = 4 * (E(sgn((X_{i1} - X_{i2})(X_{j1} - X_{j2})(X_{k2} - X_{k3})(X_{l2} - X_{l3}))) - (vecl(K)vecl(K)^T)_{(ij),(kl)})$$

$\{(ij),(kl)\}$ *denotes the entries corresponding to the covariance of Kendall's tau between* (ij) *and* (kl)*-th pair of variables.*

Now, we can rewrite the expression $E(sgn((X_{i1-X_i2})(X_{j1} - X_{j2})(X_{k2} - X_{k3})(X_{l2} - X_{l3})))$ as follows -

$$\begin{aligned}
&E(sgn((X_{i1-X_i2})(X_{j1} - X_{j2})(X_{k2} - X_{k3})(X_{l2} - X_{l3}))) \\
&= E(E(sgn((X_{i1} - X_{i2})(X_{j1} - X_{j2})(X_{k2} - X_{k3})(X_{l2} - X_{l3}))|(X_{i2}, X_{j2}, X_{k2}, X_{l2}))) \\
&= E(E(sgn((X_{i1} - X_{i2})(X_{j1} - X_{j2}))|(X_{i2}, X_{j2}, X_{k2}, X_{l2})) * E(sgn((X_{k2} - X_{k3})(X_{l2} - X_{l3}))|(X_{i2}, X_{j2}, X_{k2}, X_{l2}))) \\
&= E(E(sgn((X_{i1} - X_{i2})(X_{j1} - X_{j2}))|(X_{i2}, X_{j2})) * E(sgn((X_{k2} - X_{k3})(X_{l2} - X_{l3}))|(X_{k2}, X_{l2}))) \\
&= E(H_{ij}(X_{i2}, X_{j2})H_{kl}(X_{k2}, X_{l2}))
\end{aligned} \tag{S5}$$

,where, $H_{ij}(x,y) = E(sgn((X_i - x)(X_j - y))$. We can estimate $H_{ij}(x,y)$ from sample as - $\hat{H}_{ij}(x,y) = \frac{1}{n}\sum_{m=1}^{n}(sgn((X_{im} - x)(X_{jm} - y))$

Hence, the quantity in (S5) can be estimated as -

$$\frac{1}{n}\sum_{m=1}^{n}\hat{H}_{ij}(X_{im}, X_{jm})\hat{H}_{kl}(X_{km}, X_{lm})$$

.

Evaluating each $\hat{H}_{ij}$ requires $O(n)$ FLOPs and taking products and summing them over takes $O(n)$ FLOPs resulting in $O(n^2)$ computational complexity. This way of computation is a significant improvement over calculating the quantity in (S5) blatantly which would have required $O(n^4)$ FLOPs. Hence, we

provide a novel efficient way of calculating asymptotic variance of Kendall's Tau which would have been infeasible for even moderate n.

Now we want to derive the asymptotic normality of $vecl(\Sigma_n)$ and $vec(\beta)$ using Delta method and the following result.

As shown in Archakova and Hansen, 2018, a correlation matrix A can be parametrized by $vecl(logA)$. There exists a bijective map $\gamma : \mathbb{C}_p \longrightarrow \mathbb{R}^{p(p-1)/2}$ which is defined by $\gamma(A) = vecl(logA)$, where $\mathbb{C}_p$ denotes the set of $p \times p$ correlation matrices. As described in Tracy and Jinadasa, 1988, a general technique of defining derivatives with respect to a structured matrix (such as a correlation matrix) is to first define a map from the matrix to the independent elements of the matrix and then extend the function under investigation to the set of general matrices. For example, let's take a function $h(A)$ of a correlation matrix A, then we will define the derivative as -

$$\frac{dvec(h(A))}{dvecl(\gamma(A))} = \frac{dvec(h(A))}{dvec(A)} \frac{dvec(A)}{dvecl(A)} \frac{dvecl(A)}{dvecl(\gamma(A))}$$

.

where, the first derivative $\frac{dvec(h(A))}{dvec(A)}$ is defined assuming h is a general map defined on unstructured matrices. We can use this result and chain rule to derive the following -

$$\begin{aligned} \frac{dvecl(\Sigma)}{dvecl(\gamma(K))} &= \frac{dvecl(\Sigma)}{dvecl(K)} \frac{dvecl(K)}{dvecl(\gamma(K))} = \mathbb{D}_g \Gamma \\ \frac{dvec(\beta)}{dvecl(\gamma(K))} &= \frac{dvec(\beta)}{dvec(\Sigma)} \frac{dvec(\Sigma)}{dvecl(\Sigma)} \frac{dvecl(\Sigma)}{dvecl(\gamma(K))} = \mathbb{D}_\beta H_p \mathbb{D}_g \Gamma \end{aligned} \tag{S6}$$

Here, H_p denotes duplication matrix of order p which transforms $vecl(A)$

to $vec(A)$ for any matrix A, $\mathbb{D}_g = diag(g'(K))$, where g' is the first order derivative of the indidividual bridging functions and then we apply it to K, $\mathbb{D}_\beta = ((\Sigma_{22}^{-1}, -(\Sigma_{21} \otimes I_p)(\Sigma_{22}^{-1} \otimes \Sigma_{22}^{-1}))\tilde{E}$, where $\tilde{E}$ transforms $vec(\Sigma)$ to $(vec(\Sigma_{21}), vec(\Sigma_{22}))$. Γ is calculated in Archakova and Hansen, 2018.

Now, to use the results in (S6), we first derive the asymptotic normality of $\sqrt{n}(vecl(\gamma(K_n) - \gamma(K))$ using results in Archakova and Hansen, 2018 and calculate the asymptotic covariance matrix as V_γ. Then, under the regularity assumptions, we apply delta method to get asymptotic covariance matrix of $\sqrt{n}(vecl(\hat{\Sigma}_n - vecl(\Sigma))$ as $V_\Sigma = (\mathbb{D}_g\Gamma)^T V_\gamma \mathbb{D}_g\Gamma$ and asymptotic covariance matrix of $\sqrt{n}(\hat{\beta}_n - \beta)$ as $V_\beta = (\mathbb{D}_\beta H_p \mathbb{D}_g\Gamma)^T V_\gamma \mathbb{D}_\beta H_p \mathbb{D}_g\Gamma$.

S2 Additional Plots and Tables

Table S1: Latent correlation matrix of 5 variables of interest

	Mortality	Mobility Problem	Health Status	VPA	TAC
Mortality	1.00	0.39	0.23	-0.30	-0.45
Mobility Problem	0.39	1.00	0.50	-0.35	-0.53
Health Status	0.23	0.50	1.00	-0.24	-0.27
VPA	-0.30	-0.35	-0.24	1.00	0.63
TAC	-0.45	-0.53	-0.27	0.63	1.00

Table S2: Pearson's correlation matrix of 5 variables of interest

	Mortality	Mobility Problem	Health Status	VPA	TAC
Mortality	1.00	0.2	0.13	-0.08	-0.23
Mobility Problem	0.2	1.00	0.38	-0.15	-0.37
Health Status	0.13	0.38	1.00	-0.15	-0.37
VPA	-0.08	-0.15	-0.15	1.00	0.5
TAC	-0.23	-0.37	-0.22	0.5	1.00

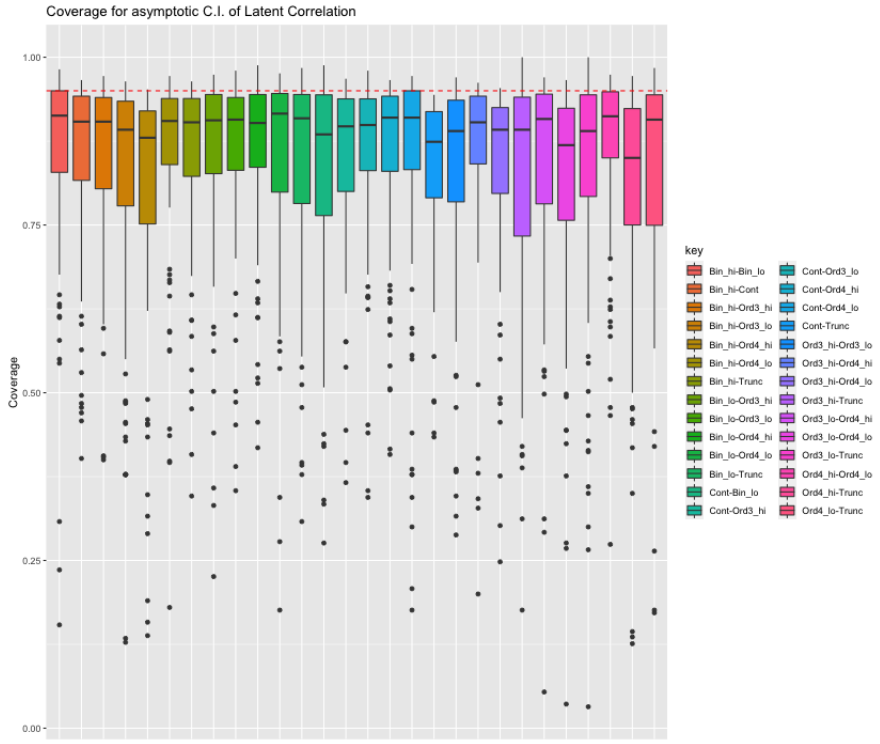


Figure S1: The coverage of the 95% asymptotic confidence interval for latent correlations. The red dotted line denotes 0.95 line

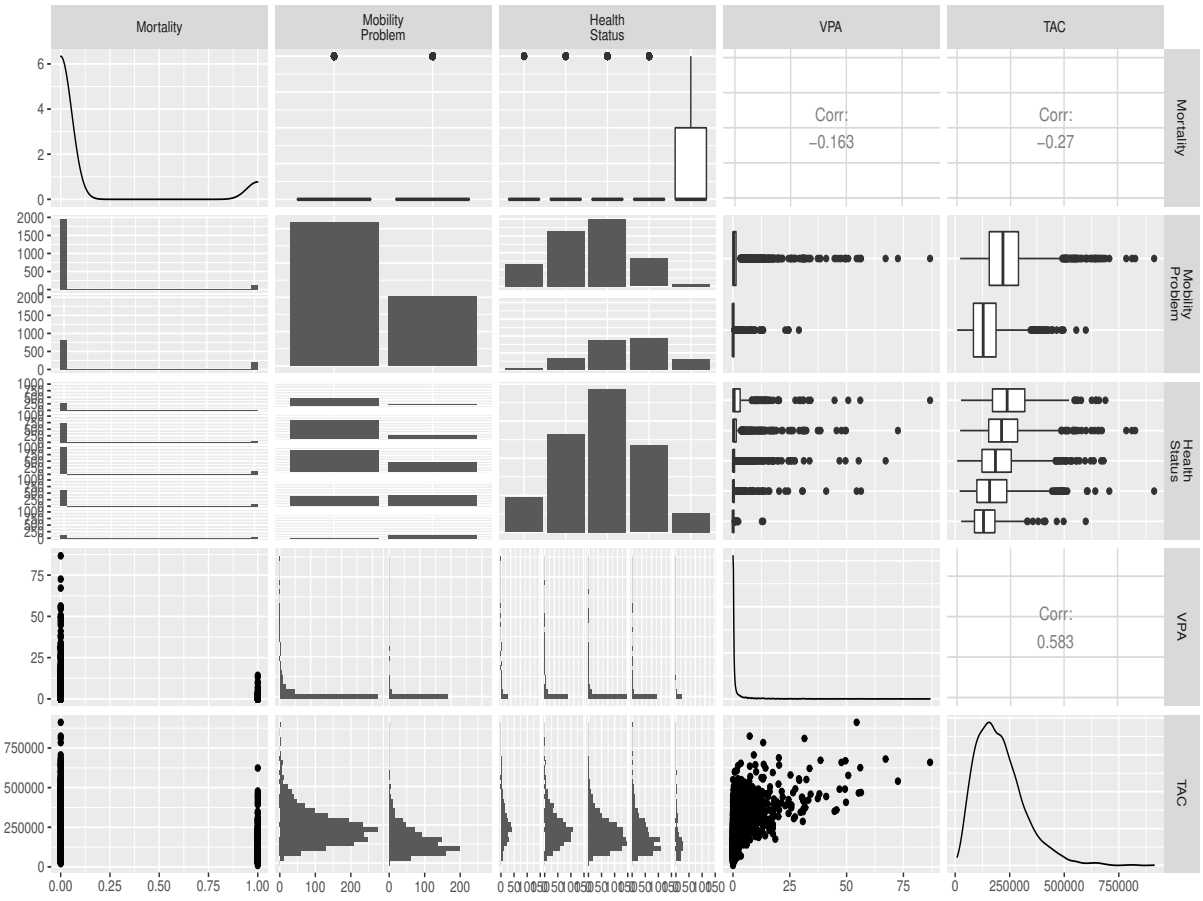


Figure S2: Exploratory analysis for our variables of interest from NHANES

References

McCullagh, Peter (1980). "Regression models for ordinal data". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 42.2, pp. 109–127.

Tobin, James (1958). "Estimation of relationships for limited dependent variables". In: *Econometrica: journal of the Econometric Society*, pp. 24–36.

Heckman, James J (1976). "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models". In: *Annals of economic and social measurement, volume 5, number 4*. NBER, pp. 475–492.

Hausman, Jerry A and David A Wise (1977). "Social experimentation, truncated distributions, and efficient estimation". In: *Econometrica: Journal of the Econometric Society*, pp. 919–938.

Cragg, John G (1971). "Some statistical models for limited dependent variables with application to the demand for durable goods". In: *Econometrica: Journal of the Econometric Society*, pp. 829–844.

Olkin, Ingram and Robert F Tate (1961). "Multivariate correlation models with mixed discrete and continuous variables". In: *The Annals of Mathematical Statistics*, pp. 448–465.

Anderson, John A and JD Pemberton (1985). "The grouped continuous model for multivariate ordered categorical variables and covariate adjustment". In: *Biometrics*, pp. 875–885.

Leon, Alexander R de and KC Carriégre (2007). "General mixed-data model: Extension of general location and grouped continuous models". In: *Canadian Journal of Statistics* 35.4, pp. 533–548.

De Leon, AR (2005). "Pairwise likelihood approach to grouped continuous model and its extension". In: *Statistics & probability letters* 75.1, pp. 49–57.

Song, Peter X-K, Mingyao Li, and Ying Yuan (2009). "Joint regression analysis of correlated data using Gaussian copulas". In: *Biometrics* 65.1, pp. 60–68.

Jiryaie, F, N Withanage, B Wu, and AR De Leon (2016). "Gaussian copula distributions for mixed data, with application in discrimination". In: *Journal of Statistical Computation and Simulation* 86.9, pp. 1643–1659.

Wang, William Yang and Zhenhao Hua (2014). "A semiparametric gaussian copula regression model for predicting financial risks from earnings calls". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1, pp. 1155–1165.

Cai, T Tony and Linjun Zhang (2015). "High-dimensional Gaussian copula regression: Adaptive estimation and statistical inference". In: *arXiv preprint arXiv:1512.02487*.

Fan, Jianqing, Lingzhou Xue, and Hui Zou (2016). "Multitask quantile regression under the transnormal model". In: *Journal of the American Statistical Association* 111.516, pp. 1726–1735.

Liu, Han, John Lafferty, and Larry Wasserman (2009). "The nonparanormal: Semiparametric estimation of high dimensional undirected graphs". In: *Journal of Machine Learning Research* 10.Oct, pp. 2295–2328.

Liu, Han, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman (2012). "High-dimensional semiparametric Gaussian copula graphical models". In: *The Annals of Statistics* 40.4, pp. 2293–2326.

Fan, Jianqing, Han Liu, Yang Ning, and Hui Zou (2017). "High dimensional semiparametric latent graphical model for mixed data". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.2, pp. 405–421.

Yoon, Grace, Raymond J Carroll, and Irina Gaynanova (2018). "Sparse semiparametric canonical correlation analysis for data of mixed types". In: *arXiv preprint arXiv:1807.05274*.

Quan, Xiaoyun, James G Booth, and Martin T Wells (2018). "Rank-based approach for estimating correlations in mixed ordinal data". In: *arXiv preprint arXiv:1809.06255*.

Feng, Huijie and Yang Ning (2019). "High-dimensional Mixed Graphical Model with Ordinal Data: Parameter Estimation and Statistical Inference". In: *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 654–663.

Zhang, Aiying, Jian Fang, Vince D Calhoun, and Yu-ping Wang (2018). "High dimensional latent Gaussian copula model for mixed data in imaging genetics". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, pp. 105–109.

Huang, Mingze, Christian L Müller, and Irina Gaynanova (2021). "latentcor: An R Package for estimating latent correlations from mixed data types". In: *arXiv preprint arXiv:2108.09180*.

Higham, Nicholas J (2002). "Computing the nearest correlation matrix—a problem from finance". In: *IMA journal of Numerical Analysis* 22.3, pp. 329–343.

Eicker, Friedhelm (1963). "Asymptotic normality and consistency of the least squares estimators for families of linear regressions". In: *The annals of mathematical statistics* 34.2, pp. 447–456.

Hoeffding, Wassily (1992). "A class of statistics with asymptotically normal distribution". In: *Breakthroughs in statistics*. Springer, pp. 308–334.

El Maache, Hamani and Yves Lepage (2003). "Spearman's rho and Kendall's tau for multivariate data sets". In: *Lecture Notes-Monograph Series*, pp. 113–130.

Archakova, Ilya and PR Hansen (2018). *A new parametrization of correlation matrices*. Tech. rep. Working Paper.

Wilhelm, Stefan and B. G. Manjunath (2010). "tmvtnorm : A Package for the Truncated Multivariate Normal Distribution". In.

Wilhelm, Stefan and Manjunath B G (2015). *tmvtnorm: Truncated Multivariate Normal and Student t Distribution*. URL: http://CRAN.R-project.org/package=tmvtnorm.

Joe, Harry (2006). "Generating random correlation matrices based on partial correlations". In: *Journal of Multivariate Analysis* 97.10, pp. 2177–2189.

Leroux, Andrew, Junrui Di, Ekaterina Smirnova, Elizabeth J Mcguffey, Quy Cao, Elham Bayatmokhtari, Lucia Tabacu, Vadim Zipunnikov, Jacek K Urbanek, and Ciprian Crainiceanu (2019). "Organizing and analyzing the activity data in NHANES". In: *Statistics in Biosciences*, pp. 1–26.

Tracy, DS and KG Jinadasa (1988). "Patterned matrix derivatives". In: *Canadian Journal of Statistics* 16.4, pp. 411–418.

Chapter 3

Connecting population-level AUC and latent scale-invariant R^2 via Semiparametric Gaussian Copula and rank correlations

3.1 Introduction

Classification offers a wide range of techniques from classical methods modelling underlying probabilities such as logistic regression and linear discriminant analysis to more recent methods including support vector machines, random forests, and neural networks (Kuhn and Johnson, 2013; Bishop, 2006). Various measures of classification accuracy have been proposed (Steyerberg et al., 2009; Steyerberg et al., 2010; Harrell Jr, 2015). Receiver Operating Characteristic curve (ROC) represents the accuracy of a classification model via a curve of trade-offs between false positive and true positive rates (Kuhn and Johnson, 2013; Saito and Rehmsmeier, 2015). The Area Under the ROC Curve (AUC) is often used to summarize the entire ROC curve and to compare

classification models in terms of their discrimination strength. AUC can be equivalently represented as the probability that a randomly chosen case has a larger value of the continuous predictor than a randomly chosen control. From this conditional distribution perspective, AUC can be viewed as a fully nonparametric measure of concordance between an observed binary outcome and an observed continuous predictor. This definition has been linked to non-parametric Kendall's Tau rank correlation and Wilcoxon rank-sum statistics (Kendall et al., 1987). The limitations of AUC as a single summary of classification accuracy has been widely discussed in literature (Lobo, Jiménez-Valverde, and Real, 2008; Harrell Jr, 2015; Steyerberg et al., 2009; Saito and Rehmsmeier, 2015).

In regression models with binary outcomes such as logistic and probit regressions, an alternative fully parametric perspective is taken by using goodness-of-fit measures. A big class of these measures are likelihood-based (Tutz, 2011; DeMaris, 2002; Schemper, 2003). Another class of these measures focuses on extending R^2, as a recognisable and intuitive goodness-of-fit measure for linear models with continuous outcomes. Specifically, two main alternative interpretations of R^2 as the proportion of variance explained and as a squared correlation have been adapted and extended to models with binary outcome by DeMaris, 2002; Schemper, 2003; Yazici, Alpu, and Yang, 2007.

In this paper, we will use Semiparametric Gaussian Copula (Fan et al., 2017) (SGC) to bridge the classification accuracy and goodness-of-fit perspectives. Specifically, we connect AUC as a measure of classification accuracy

and a novel latent scale-invariant R^2 as a measure of explained variation. Semiparametric Gaussian Copula (SGC) was used by Fan et al., 2017 to model a joint dependence between an observed binary outcome and an observed continuous predictor via the correlation of latent standard normal random variables. A major computational advantage of the approach in Fan et al., 2017 is the estimation procedure that does not require any likelihood maximization and estimates the latent correlation via a bridging procedure. The procedure links Kendall's Tau for binary-continuous pairs and the latent correlation via a known monotone "bridging" function that depends on the population-level prevalence rate of cases. The plug-in estimation is done using sample versions of Kendall's Tau and the prevalence rate. This bridging trick, also called inversion, is frequently used to estimate parameters of specific copula families via linking these parameters to Kendall's Tau or other rank correlations and inverting these links (Nelsen, 2007; Joe, 2014). Conceptually, the bridging argument in Fan et al., 2017 is similar to the to classical results bridging biserial correlation on binary-continuous pairs and tetrachonic correlation on binary-binary pairs to Pearson correlation of underlying continuous variables that generated binary variables via dichotomization (MacCallum et al., 2002). Because SGC uses Kendall's Tau to estimate the latent correlation, it becomes possible to connect AUC to the latent correlation that captures dependence between an observed binary outcome, generated via dichotomization of the underlying latent continuous variable, and an observed continuous predictor. In addition to classical expression of AUC via Kendall's Tau and Wilcoxon rank sum statistics, we will show that AUC is a linear (up to an absolute value, here and throughout the paper) function of Spearman rank correlation (Sidak,

Sen, and Hajek, 1999) for binary-continuous pairs. Under SGC, we will also explicitly link these three rank statistics as well as Quadrant rank correlation (Sidak, Sen, and Hajek, 1999) to the latent R^2, defined as a square of the latent correlation, via corresponding monotone "bridging" functions that we derive in this paper. Importantly, being semiparametric our approach results in the latent R^2 that is scale-invariant.

After building these connections, we will focus on two main applications. In our main application, we demonstrate how our framework addresses a problem of calculating AUC under complex survey designs. Specifically, we will show how Wilcoxon's rank-sum statistics as well as Spearman and Quadrant rank correlations can be used with single participant survey weights to construct asymptotically unbiased estimators of the population-level AUC. In the second application, we argue that Quadrant rank-correlation can be used as a robust semiparametric version of AUC. In extensive simulation studies, we show that AUC defined via Quadrant correlation is robust in scenarios with outliers in binary-continuous case. As shown in Croux and Dehon, 2010 for the continuous-continuous case, Quadrant rank correlation is more robust compared to Kendall's Tau and Spearman rank correlations, so our contribution can be seen as an extension of this for binary-continuous case under SGC assumption.

The rest of the paper is organized as follows. In Section 2, we discuss the four rank-statistics and their relationship with AUC. In Section 3, we introduce Semiparametric Gaussian Copula and derive bridging functions that connect the four rank statistics to the latent R^2. In Section 4, we discuss the main

applications of the framework. In Section 5, we provide extensive simulation results. In Section 6, we demonstrate the proposed framework on NHANES 2003-2006 cohorts. Discussion concludes with a summary, limitations, and future work.

3.2 AUC and Rank Statistics

In this section, we establish the links between AUC and the rank statistics. Let us first introduce notations. Throughout the paper, we will consider binary-continuous pairs of random variables (Y, X), where Y is an observed binary outcome and X is an observed continuous predictor. We will refer to M_Y, M_X as the population medians of Y and X, respectively. We denote by $F_Y(), F_X()$ the cumulative distribution functions of random variables Y and X, respectively. We will refer to $Y = 1$ as a case and define the population-level prevalence rate of cases as $p = P(Y = 1)$. Finally, we denote $X_1 = (X|Y = 1)$ and $X_0 = (X|Y = 0)$ to be random variables following the conditional distribution of the continuous predictor for cases, $Y = 1$, and controls, $Y = 0$, respectively. Using these notations, population-level AUC, denoted by A, can be defined as $A = \max\{P(X_1 > X_0), P(X_1 < X_0)\}$. It is easy to see that $P(X_1 > X_0) = 1 - P(X_1 < X_0)$, so $A \geqslant 0.5$.

We consider three rank correlations including Kendall's Tau, Spearman's Rho and Quadrant, also known as Blomqvist's Beta, as well as Wilcoxon's rank-sum statistic, used to nonparametrically test the equality of two distributions. We lay out the population-level definitions of these rank statistics in the case of binary-continuous pairs: 1) **Kendall's Tau**: $r_K = E((Y_i - Y_i')sgn(X_i - X_i'))$;

2) **Wilcoxon's rank-sum statistic**: $W = P(X \leqslant X_1) - P(X \leqslant X_0)$; 3) **Spearman correlation**: $r_S = 12E[F_Y(Y)F_X(X)] - 3$; 4) **Quadrant correlation**: $r_Q = E[sgn((Y - M_Y)(X - M_X))]$, where (Y_i, X_i) and $(Y_i^{'}, X_i^{'})$ are two independent copies following the same bivariate joint distribution and $sgn(x) = I\{x \neq 0\}(2I\{x > 0\} - 1)$ denotes the sign function. We next establish the relationship between these rank statistics and AUC. All derivations for these results are presented in Section S1.

3.2.1 Non-parametric relationships

Kendall's Tau captures concordance within a pair of bivariate observations and relates to AUC as $A = 0.5 + |r_K/(4p(1-p))|$.

Wilcoxon rank-sum statistic is a linear rank statistics. The use of W for calculating AUC under complex survey designs has been discussed by Professor Thomas Lumley in his blog *Bias and Inefficient* (https://notstatschat.rbind.io/2017/12/26/statistics-on-pairs/). Additionally, Lumley and Scott, 2013 demonstrated how survey-weighted rank test can be constructed using Wilcoxon rank-sum statistic to compare two distributions in a complex survey design. We followed his approach by fixing a minor error and obtaining population-level relationship $A = 0.5 + |W|$.

Finally, we show that Spearman's rank correlation and Wilcoxon rank-sum are linearly related in binary-continuous case. This is primarily due to the fact that the cumulative distribution function of a binary random variable takes exactly two values. Specifically, we show that Spearman's rank correlation and AUC are linearly related as follows $A = 0.5 +$

$\left|(r_S - (6p^2 - 6p + 3))/(12p^2(1-p))\right|$.

3.2.2 Robust semiparametric AUC via Quadrant rank correlation

In the case of continuous-continuous (X, Y) pairs, the calculation of Quadrant rank correlation requires counting the number of pairs in the first, second, third, and fourth quadrant of the (X, Y) two-dimensional plane. Because it involves only the sign of the distance of ranks from the median and not the ranks themselves, Quadrant rank correlation is highly robust with a breakdown point of 50%, but less efficient than Kendall's Tau and Spearman's Rho rank correlations (Croux and Dehon, 2010). Even though we related AUC linearly to the three rank statistics above, it is not possible to express Quadrant rank correlation linearly in terms of AUC without introducing additional assumptions on the joint distribution of (X, Y) pairs. In the next section, we introduce semiparametric Gaussian copula assumptions on the joint distribution of (X, Y) and will establish a non-linear relationship between Quadrant rank correlation and AUC. This relationship will also provide a more robust semiparametric estimate of AUC in the case of binary-continuous (Y, X) pairs.

3.3 Semiparametric Gaussian Copula

3.3.1 Introduction to the copula

The assumption of multivariate Gaussinity is arguably the most popular in multivariate statistical analysis. However, in many applications, this assumption is not realistic. To address this, Liu et al., 2012 proposed a Semi-parametric

Gaussian Copula (SGC) model. Below, we provide definitions from Liu et al., 2012.

Definition 3.3.1. We say that a pair of continuous random variables (Y, X) follows a **non-paranormal** distribution, if there exist monotone functions $f_Y(), f_X()$ such that $(U, V) = (f_Y(Y), f_X(X)) \sim N_2(0, 0, 1, 1, r)$.

For binary-continuous pairs, Fan et al., 2017 defined latent non-paranormal distribution as follows.

Definition 3.3.2. Suppose we have binary variable Y and continuous variable X. Then, if there exists a latent variable Z and monotone functions $f_Z(), f_X()$ such that $(Y, X) = (I\{f_Z(Z) > \Delta\}, X)$ and $(U, V) = (f_Z(Z), f_X(X)) \sim N_2(0, 0, 1, 1, r)$, then we say that the binary-continuous pair (Y, X) follows **latent non-paranormal distribution**.

The approach in Fan et al., 2017 estimates the latent correlation via mapping Kendall's tau using a one-to-one function $G_K()$, called "bridging" function, so that $r_K = G_K(r) = 4\Phi_2(\Delta, 0, \frac{r}{\sqrt{2}}) - 2\Phi(\Delta)$, where $\Phi_2(a, b, r)$ denotes the cumulative distribution function of a standard bivariate normal distribution with correlation r. Bridging function $G_K(r)$ is odd, so $G_K(-r) = -G_K(r)$. Figure S1 in Section S3. shows $G_K(r)$, $G_K^{'}(r)$, and $G_K^{''}(r)$ for different values of $p = 1 - \Phi(\Delta)$. Based on the sign of $G_K^{''}(r)$, we can see that $G_k(r)$ is convex for higher values of p and concave for lower values of p, and neither in between.

Thus, it becomes possible to connect AUC, via Kendall's Tau and SGC assumption, to the latent correlation that captures the dependence at data generating level between an observed binary outcome, conceptualized as a dichotomized continuous variable, and an observed continuous predictor.

3.3.2 Bridging the latent correlation and rank statistics

Next, we will show how Spearman and Quadrant rank correlations can also be used to estimate the latent correlation of SGC.

Lemma 3.3.1. *Under SGC, Spearman rank correlation, r_S, and Quadrant rank correlation, r_Q, can be mapped to the latent correlation r as follows:*

$$
\begin{aligned}
r_S &= G_S(r) = 12[\Phi_2(0, -\Delta, \frac{r}{\sqrt{2}}) + p\Phi_2(0, \Delta, -\frac{r}{\sqrt{2}})] - 3 \\
r_Q &= G_Q(r) = [\Phi_2(-\Delta, 0, r) - \Phi_2(-\Delta, 0, -r)]\mathbb{I}(M_Y = 0) + \\
&\quad [\Phi_2(\Delta, 0, r) - \Phi_2(\Delta, 0, -r)]\mathbb{I}(M_Y = 1).
\end{aligned} \tag{3.1}
$$

The detailed proof is provided in Section S1 (Equations (S7), (S8), (S9) and (S10)).

We will refer to $G_K(), G_S(), G_Q()$ as bridging functions and the subscripts K, S, Q will specify a specific rank correlation. Fan et al., 2017 showed that $G_K(r)$ is a strictly increasing function of r in $(-1, 1)$. We establish similar results for $G_S(r)$ and $G_Q(r)$ in below.

Lemma 3.3.2. *The bridging functions $G_S(r)$ and $G_K(r)$ are strictly increasing functions of r in $(-1, 1)$ and hence, the inverse functions exist.*

The proof is provided in the Section S1.

Thus, we can ensure that latent correlation estimators obtained by inverting bridging functions are well defined. It is important to remember a

few properties of the bridging functions. First, the bridging function is constructed assuming independence at the population level. Second, we only use bridging functions to bridge between population-level statistics. Third, the bridging functions do not depend on the sampling scheme which might bring dependence between subjects through sampling mechanism.

3.4 Applications

In this section, we consider main applications of the proposed framework.

Using the results from previous section, we now can connect the population-level AUC and rank statistics as follows.

$$\begin{aligned} A_K &= \frac{1}{2} + \left| \frac{r_K}{4p(1-p)} \right|, \\ A_W &= \frac{1}{2} + |W| \\ A_S &= \frac{1}{2} + \left| \frac{G_K(G_S^{-1}(r_S))}{4p(1-p)} \right| = \frac{1}{2} + \left| \frac{r_S - (6p^2 - 6p + 3)}{12p^2(1-p)} \right|, \\ A_Q &= \frac{1}{2} + \left| \frac{G_K(G_Q^{-1}(r_Q))}{4p(1-p)} \right|. \end{aligned} \tag{3.2}$$

At the population-level, all four ways are equivalent, i.e. $A = A_K = A_W = A_S = A_Q$. Of course, in sample we likely end up with different estimates of AUC. It is also important to note that the dependence of AUC and the latent correlation r involves the prevalence rate p.

3.4.1 Latent R^2 for univariate continuous predictor

Our approach provides a framework to introduce a novel goodness-of-fit statistics, the latent R^2, that we denote as R^2_l and define as a square of the latent correlation. Various goodness-of-fit measures have been previously proposed for models with binary outcome. Many of them focused on extending R^2, as a popular and intuitive measure in linear models with continuous outcome. Two main alternative interpretations of R^2 i) as the proportion of variance explained or ii) as squared correlation have been pursued and generalized to models with binary outcome DeMaris, 2002; Schemper, 2003; Yazici, Alpu, and Yang, 2007. Among the limitations of those proposals is that the range of the values can go outside of the usual $(0, 1)$ interval as well as the lack of invariance to the scale of the continuous predictors DeMaris, 2002; Schemper, 2003; Yazici, Alpu, and Yang, 2007.

Under SGC, we estimate the latent correlation using three different estimators corresponding to three rank correlations:

$$\begin{aligned} R^2_{lK} &= (G_K^{-1}(r_K))^2, \\ R^2_{lS} &= (G_S^{-1}(r_S))^2, \\ R^2_{lQ} &= (G_Q^{-1}(r_Q))^2. \end{aligned} \tag{3.3}$$

Again, at the population level, $R^2_l = R^2_{lK} = R^2_{lS} = R^2_{lQ}$.

Thus, R^2_l quantifies the proportion of variance explained by the continuous predictor in the latent normalized space and gives us back a familiar intuition available for linear models with a continuous outcome.

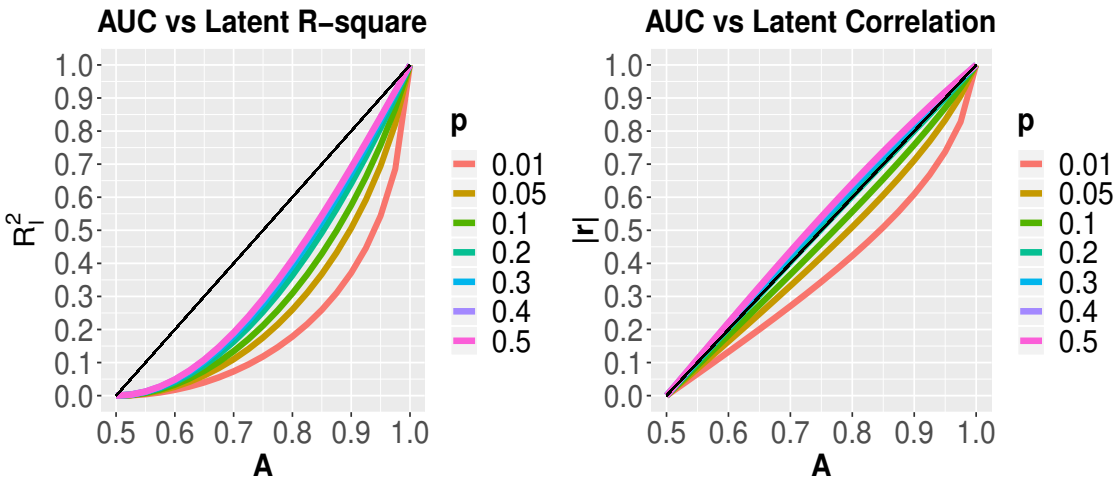


Figure 3.1: The relationships between AUC and the latent R_l^2 (left panel) and AUC and the absolute value of latent correlation (right panel) and its dependence on p.

Figure 3.1 shows the relationship between AUC and the latent R_l^2 in the left panel and the relationship between AUC and the absolute value of the latent correlation, $|r|$, in the right panel. As a reference, we include a linear line $|r| = 2A - 1$. We show these relationships for different values of the prevalence rate, p. The right panel shows that AUC and $|r|$ are almost identical for $p = 0.5$, but with p getting smaller, the same value of AUC corresponds to increasingly smaller values of $|r|$. For example, AUC of 0.8 corresponds to $|r|$ of 0.6, if $p = 0.5$, and $|r|$ of 0.4, if $p = 0.01$. The latent correlation tends to lie below the linear reference line for most of values of p. The same observations are true for the relationships between AUC and R_l^2, but with a much larger curvature due to the squared nature of R_l^2's scale compared to the linear scale of $|r|$.

3.4.2 Latent R^2 for multivariate continuous predictor

The definition of R^2_l can be readily extended to a case of a linear combination $X'\beta$ for a fixed pre-defined multivariate parameter β. Indeed, if a linear combination $X'\beta$ can be treated as a scalar continuous predictor and, if the SGC assumptions for $X'\beta$ are valid, the same argument connecting AUC and R^2_l can be applied to a binary outcome Y and a scalar continuous predictor $X'\beta$.

An alternative two-step procedure to define R^2_l for a multivariate continuous predictor can be done as follows. First, estimate the joint latent correlation matrix, Σ, of binary outcome Y and multivariate continuous predictor X. Second, define R^2_l as $R^2_l = \Sigma_{YX}\Sigma^{-1}_{XX}\Sigma_{XY}$, where Σ_{YX} denotes the latent correlation between the outcome and predictor and Σ_{XX} denotes the latent correlation matrix for the predictor. This approach generalizes R^2 for linear models with continuous outcome and continuous predictors. However, whether there is a specific relationship between AUC and R^2_l defined in this way remains to be investigated and is beyond the scope of this paper.

3.4.3 Complex survey designs

Many social and public health studies are conducted using data from surveys following complex designs. To properly estimate the population-level AUC in complex surveys, one would need to know the pairwise sampling weights, or pairwise probabilities of selection, for study participants. However, pairwise survey weights are often not available. To tackle this, researchers have used various approximations such as the product of individual participant weights

(Korn and Graubard, 2011) or have applied non-parametric approaches (Yao, Li, and Graubard, 2015). In this section, we show how rank statistics and single participant survey weights can be used to construct asymptotically unbiased estimators of the population-level AUC.

First, we introduce the definition of the survey-weighted AUC.

Definition 3.4.1. Suppose, $X_{11}, X_{12}, ..., X_{1m_1}$ and $X_{01}, X_{02}, ..., X_{0m_0}$ are i.i.d. samples from the distributions of X_1 and X_0, respectively. Let us assume that the total sample size is $n = m_0 + m_1$. Then, the survey-weighted AUC can be calculated as

$$\hat{A}_{wt} = \frac{1}{\sum_{i=1}^{m_0}\sum_{j=1}^{m_1}\frac{1}{w(i,j)}}\sum_{i=1}^{m_1}\sum_{j=1}^{m_0}\frac{1}{w(i,j)}h(X_{1i}, X_{0j}) = \hat{E}_w h(X_1, X_0), \tag{3.4}$$

where $h(x,y) = \mathbb{I}(x > y) + 0.5\mathbb{I}(x = y)$, $w(i)$ and $w(i,j)$ are single and pairwise participant weights, respectively.

3.4.3.1 AUC using single participant weights

The key idea of the proposal outlined below is to use single participant survey weights to estimate population-level rank statistics and then connect them to the population-level AUC using bridging functions. We can define the survey-weighted estimators (Horvitz-Thompson estimators) of Kendall's Tau, Wilcoxon rank-sum statistic, Spearman and Quadrant rank correlations as follows.

$$\hat{r}_K = \frac{1}{\sum_{i<j} \frac{1}{\hat{w}(i,j)}} \sum_{i<j} \frac{1}{\hat{w}(i,j)} [(Y_i - Y_j) sgn(X_i - X_j)]$$

$$\hat{W} = \frac{1}{\sum_{i:Y_i=1} \frac{1}{w(i)}} \sum_{i:Y_i=1} \frac{1}{w(i)} \hat{F}_X(X_i) - \frac{1}{\sum_{i:Y_i=0} \frac{1}{w(i)}} \sum_{i:Y_i=0} \frac{1}{w(i)} \hat{F}_X(X_i) \tag{3.5}$$

$$\hat{r}_S = 12 \frac{1}{\sum_{i=1}^{n} \frac{1}{w(i)}} \sum_{i=1}^{n} \frac{1}{w(i)} [\hat{F}_Y(Y_i) \hat{F}_X(X_i)] - 3$$

$$\hat{r}_Q = \frac{1}{\sum_{i=1}^{n} \frac{1}{w(i)}} \sum_{i=1}^{n} \frac{1}{w(i)} sgn((Y_i - \hat{M}_Y)(X_i - \hat{M}_X)).$$

The estimates of population-level medians, M_Y and M_X, and the population-level distribution functions, F_Y and F_X, are obtained using Horvitz-Thomposon estimators.

3.4.3.2 Asymptotic properties

Lumley and Scott, 2013 established the asymptotic properties of statistics of the form

$$\hat{T} = \frac{1}{\sum_{i:Y_i=1} \frac{1}{w(i)}} \sum_{i:Y_i=1} \frac{1}{w(i)} g(\hat{F}_X(X_i)) - \frac{1}{\sum_{i:Y_i=0} \frac{1}{w(i)}} \sum_{i:Y_i=0} \frac{1}{w(i)} g(\hat{F}_X(X_i))$$

under complex survey designs. The function $g()$ can follow any of the assumptions stated in Section S2. Theorem 1 in Lumley and Scott, 2013 has been used to show that $\sqrt{n}(\hat{W} - W)$ are asymptotically normal with mean zero. We adopt a similar approach to prove the asymptotical normality for r_S and r_Q.

Theorem 3.4.1. *Under assumptions A1 to A4 in Section S2,* $\sqrt{n}(\hat{r}_S - r_S)$ *and* $\sqrt{n}(\hat{r}_Q - r_Q)$ *are asymptotically normal with mean zero. Hence,* $(\hat{r}_S - r_S)$ *and* $(\hat{r}_Q - r_Q)$ *converge to zero in probability.*

The proof is provided in Section S2.

Kendall's Tau requires pairwise survey weights $\hat{w}(i,j)$. We consider three different estimates of Kendall's Tau: (1) *unweighted*, $\hat{r}_{Kuw}$, with $\hat{w}(i,j) = 1$; (2) *true weighted*, $\hat{r}_{Ktw}$, with $\hat{w}(i,j) = w(i,j)$; and (3) *product weighted*, $\hat{r}_{Kpw}$, with $\hat{w}(i,j) = w(i)w(j)$. Note that in most of practical settings, we do not know or have an access to true pairwise weights and can only calculate $\hat{r}_{Ktw}$ in simulations.

We can estimate the population-level AUC using Equations (3.2) and define corresponding estimates as $\hat{A}_{Kuw}$, $\hat{A}_{Ktw}$, $\hat{A}_{Kpw}$, $\hat{A}_W$, $\hat{A}_S$, and $\hat{A}_Q$.

Using Theorem 3.4.1 and applying the delta method to Equations 3.2 leads to the the following result.

Corollary 3.4.1.1. *Under the assumptions of Theorem 3.4.1,* $\sqrt{n}(\hat{A}_S - A_S)$ *and* $\sqrt{n}(\hat{A}_Q - r_Q)$ *are asymptotically normal with mean zero. Hence,* $(\hat{A}_S - A_S)$ *and* $(\hat{A}_Q - A_Q)$ *converge to zero in probability.*

Note that we can apply delta method because of the differentiability of the bridging functions as shown in the proof of Lemma 3.3.2 in Appendix A.

Thus, using single participant survey weights, we can define asymptotically unbiased estimators of the population-level AUC. Finally, we can invert bridging functions to get three different estimates of the latent R_l^2 as $\hat{R}_{lK}^2$, $\hat{R}_{lS}^2$, and $\hat{R}_{lQ}^2$, respectively.

3.5 Simulations

In this section, we perform extensive simulation studies to explore the behaviour of the proposed estimators of AUC. Under two-stage stratified cluster sampling, we consider several different combinations of strata informativeness and outlyingness in binary-continuous pairs.

Following Yao, Li, and Graubard, 2015, we set up a two-stage stratified cluster sampling as follows. Suppose, we have a finite population, N subjects, with H strata and each strata has K Primary Sampling Units (PSUs). The number of subjects in the gth PSU in the h-th strata are N_{hg}. Suppose, we pick n_{hg} subjects from each PSU hg and the total number of subjects selected from hth strata is n_h with the total number of subjects in the sample to be n. We sample subjects as follows: first, randomly select u out of U PSUs within each stratum and then randomly select a fixed number of subjects (n_h/u) from each of the PSU in h-th strata, for $h = 1, \cdots, H$. Based on this sampling scheme, we can easily calculate both the individual selection probability and pairwise selection probabilities as described in Section 2.3 of Yao, Li, and Graubard, 2015.

First, we define **strata informativeness** as follows. If n_h depends on the AUC in h-th strata, we call this informative sampling. If the sampling is informative, we oversample strata with higher AUCs. We rank strata according to AUC for that specific strata. Depending on the level of oversampling, we define three types of informativeness: 0-None ($n_h = n/H$ for all h) , 1-Moderate (top three: $n_h = n * 0.18$, middle four: $n_h = n * 0.07$, bottom three: $n_h = n * 0.06$), 2-Strong (top three: $n_h = n * 0.22$, middle four: $n_h = n * 0.07$, bottom three:

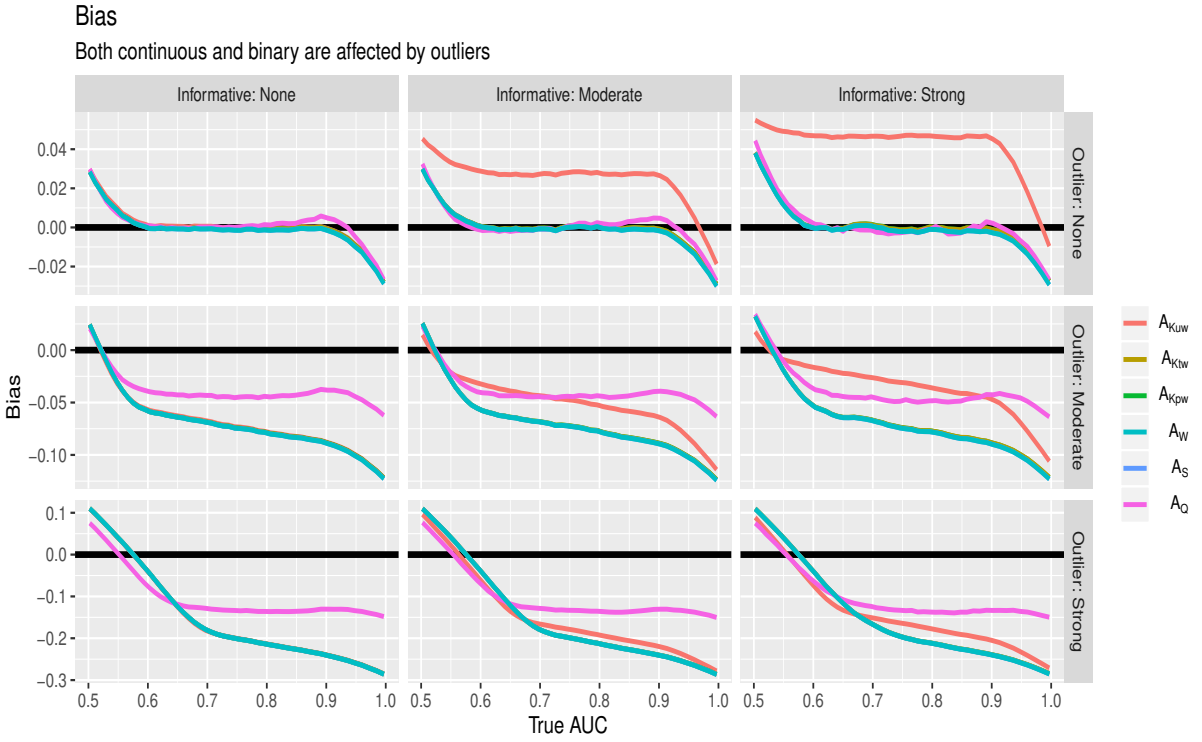


Figure 3.2: Simulation results: Bias of the AUC estimators under different scenario

$n_h = n * 0.02$). We define **outlyingness** as follows. Following Croux and Dehon, 2010, we introduce "outlying" observations at latent level by substituting $N_2(0, 0, 1, 1, r)$ with $N_2(4, -4, 0.01, 0.01, 0)$ or $N_2(-4, 4, 0.01, 0.01, 0)$ randomly selected with probability $\frac{1}{2}$. We consider three levels of outlyingness: 0-None (0% of outliers), 1-Moderate (5% of outliers), 2-Strong (15% of outliers).

We create 9 different scenarios corresponding to the 3×3 combinations of strata informativeness and outlyingness. We take $H = 10, U = 10, u = 2, n = 600, N = 60000$. For each scenario, we simulate data and calculate estimates as follows.

(a) **Schematics**

1. For a fixed latent correlation, r, we calculate true AUC, A, using bridging.

2. Take strata-specific AUCs, A_h, equally spaced within $(A - 0.1, A + 0.1)$. For example, if AUC A is 0.8 and there are 10 stratas, the strata specific AUCs are $0.7, 0.722, 0.744, 0.766, 0.788, 0.811, 0.833, 0.855, 0.877, 0.9$.

3. Depending on informativeness of the sampling scenario, calculate n_h based on A_h.

4. Calculate individual, $w(i)$, and pairwise, $w(i, j)$, survey weights for this particular scheme.

(b) **Within Strata**

1. For A_h, use bridging to calculate strata-specific latent correlation, r_h.

2. Generate $(U_i, V_i), i = 1, \cdots, N_h \sim N_2(0, 0, 1, 1, r_h)$.

(c) **Sampling and outliers**

1. Get a sample of size n using sampling scheme described in the beginning of the section.

2. Depending on the level of "outlyingness", randomly select 0%, 5% or 15% of "outliers" across all stratas by changing corresponding (U_i, V_i) from $N_2(0, 0, 1, 1, r_h)$ to a random draw from $N_2(4, -4, 0.01, 0.01, 0)$ or $N_2(-4, 4, 0.01, 0.01, 0)$ with probability $\frac{1}{2}$.

3. Transform $Y_i = I(U_i > \Delta)$ and keep $X_i = V_i$ for the final finite sample.

(d) **Estimation**

1. Estimate Δ as $\hat{\Delta} = \Phi^{-1}(1 - \frac{\sum_{i=1}^{n} \frac{Y_i}{w_i}}{\sum_{i=1}^{n} \frac{1}{w_i}})$

2. Estimate $\hat{r}_K$, using three different weighting schemes, $\hat{r}_S$ and $\hat{r}_Q$, using single participant weights, by inverting corresponding bridging functions described in Section 3.3.2.

3. Estimate A using Equations (3.2) to obtain $\hat{A}_{Kuw}$, $\hat{A}_{Ktw}$, $\hat{A}_{Kpw}$, $\hat{A}_W$, $\hat{A}_S$, and $\hat{A}_Q$.

We vary true latent correlation r between $(0.005, 0.995)$ taking 50 equally spaced points and generate data from a specific scenario 100 times. The results are used to report bias and mean squared errors (MSEs).

Figure 3.2 that shows bias across the nine scenarios. Note that each row has row-specific range for vertical axis, with the lowest range in the top row (0% of outliers) and the highest range in the bottom row (15% of outliers). The key findings are as follows. With increasing informativeness in sampling (from the most left to the most right column), bias of the unweighted estimator is relatively higher than that of the weighted versions. With increasing outlyingness (from the top to the bottom row), bias of A_Q is significantly lower than weighted and unweighted AUC obtained using the other rank correlations. Even in the most challenging scenario, shown in the bottom most right corner, A_Q exhibits maximum absolute bias of only 0.12 while other estimates have maximum absolute bias as large as 0.3.

These results provide a clear illustration that A_Q can be seen and used as a robust semiparametric version of AUC. In the case of higher sampling

informativeness and moderate presence of outliers, we observe that the unweighted AUC performs better than the weighted AUC. We argue that the intuition behind this is that in these kind of scenarios, the outlying observations might have higher weights and thus, unweighted estimates might be less biased than the weighted ones.

The similar observations are true for MSEs of the estimators, shown in Figure S2 in Section S3.

3.6 Classification of 5-year mortality in NHANES 2003-2006

The National Health and Nutrition Examination Survey (NHANES) is a biennial stratified multi-stage sample survey of non-institutionalized US population conducted by Center of Disease Control and Prevention, USA (https://www.cdc.gov/nchs/nhanes/index.htm). Using complex survey weighting techniques, the results obtained from NHANES can be considered as nationally representative. To illustrate our approach, we use accelerometry and laboratory data as well as linked mortality data for NHANES 2003-2006. We define a 5-year all cause mortality to be the observed binary outcome and use continuous predictors including age, albumin, systolic blood pressure and summaries of accelerometry-estimated physical activity which was a part of NHANES 2003-2006 protocol (Leroux et al., 2019).

Age (in years) is derived from the NHANES variable *RIDAGEYR*, defined as the age of a participant at the time of household screening. Albumin (in ug/mL) is the variable *URXUMA*. Systolic blood pressure (in mmHg) has

been calculated as the average of multiple readings of the systolic blood pressure (up to four sequential readings per participant) which we denote here as $BPXSY$. Following Leroux et al., 2019, we also include total activity count (TAC), time in minutes spent in moderate-to-vigorous activity (MVPA) and active-to-sedentary transition probability (ASTP). These accelerometry-derived measures have been shown to be significantly predictive of a 5-year follow-up mortality with classification performance comparable to and sometimes exceeding that of Age (Leroux et al., 2019; Smirnova et al., 2019). We excluded participants who (1) have missing mortality information or alive with follow-up less than 5 years, (2) are younger than 50 or aged 85 and older, (3) have missing any of predictor variables of interest, (4) have died due to accident, and (5) had fewer than 3 days of data with at least 10 hours of estimated accelerometry wear time (Leroux et al., 2019). Our final analytical sample consisted of 3069 subjects with 321 deaths within 5 years. This gives an unweighted estimate of prevalence rate as $p = 0.1$. Note that we use term "prevalence rate" while referring to p loosely here, just to stay consistent with the rest of the paper. We use *rnhanesdata* package (https://github.com/andrew-leroux/rnhanesdata) to process and recalculate the survey weights for our chosen subset of participants. Finally, the binary outcome is defined as a binarized 5-year follow-up mortality.

We use methods described in Section 3.4.3 and Equation 3.5 to calculate our proposed estimates and their standard errors and confidence intervals. Specifically, (i) we follow the approach by Yeo, Mantel, and Liu, 1999 and use bootstrap to get standard error and confidence intervals, (ii) we use

svyrepdesign function in **R** package *Survey* to create 100 bootstrap replicates of the sample design using survey weights, (iii) then we calculate AUC and R^2_l estimates from each of these replicates, and report bootstrap standard error and 95% confidence intervals in Tables 3.1 and 3.2.

The results are consistent with those in Leroux et al., 2019. For unweighted AUC, A_{Kuw}, if we rank AUCs from highest to lowest, the predictors are ordered as TAC, Age, MVPA, ASTP, Albumin, Systolic BP. For weighted versions of AUC, A_{Kpw}, A_W, and A_S, MVPA becomes more discriminative than Age. However, Age has been used for NHANES survey weights, so an influence on the results is expected. Interestingly, when using A_Q, MVPA becomes the predictor with highest AUC and TAC becomes the second highest. This is likely due many elderly participants having zero minutes of moderate-to-vigorous physical activity, where as, TAC provides a better discrimination between participants with MVPA zeroes (Varma et al., 2018). Hence, because Quadrant rank correlation only counts the quadrant concordance, MVPA gets this slight preference by Quadrant.

The results for R^2_l show that TAC, the total volume of physical activity, explains about 33%, 29%, 29%, and 23% variation in the binary outcome of mortality, when we use the unweighted and weighted Kendall's Tau, weighted Spearman's and Quadrant rank correlations, respectively. The lowest R^2_l is for Systolic BP, which explains less than one percent of the variation.

It is very important to re-iterate that the one-to-one bridging of AUCs and R^2_l depends on the prevalence rate. So, if we had a different p, we would get a different R^2_l.

	Variables	A_{Kuw}	Rank	A_{Kpw}	Rank	A_W	Rank	A_S	Rank	A_Q	Rank
1	TAC	0.75 (0.75, 0.75)	1	0.8 (0.75, 0.83)	1	0.8 (0.75, 0.83)	1	0.8 (0.75, 0.83)	1	0.77 (0.73, 0.8)	2
2	MVPA	0.73 (0.73, 0.73)	3	0.78 (0.74, 0.81)	2	0.78 (0.73, 0.81)	2	0.78 (0.74, 0.81)	2	0.78 (0.75, 0.82)	1
3	Age	0.74 (0.74, 0.74)	2	0.77 (0.72, 0.8)	3	0.76 (0.72, 0.8)	4	0.77 (0.72, 0.8)	3	0.74 (0.7, 0.77)	4
4	ASTP	0.73 (0.73, 0.73)	4	0.76 (0.73, 0.8)	4	0.76 (0.73, 0.81)	3	0.76 (0.73, 0.8)	4	0.74 (0.7, 0.78)	3
5	Albumin	0.65 (0.65, 0.65)	5	0.7 (0.66, 0.73)	5	0.7 (0.66, 0.73)	5	0.7 (0.66, 0.73)	5	0.68 (0.64, 0.71)	5
6	Systolic BP	0.54 (0.54, 0.54)	6	0.53 (0.5, 0.57)	6	0.53 (0.5, 0.57)	6	0.53 (0.5, 0.57)	6	0.5 (0.5, 0.57)	6

Table 3.1: AUC estimates and 95% bootstrap confidence intervals for continuous predictors in NHANES 2003-2006.

	Variables	R^2_{lKuw}	Rank	R^2_{lKpw}	Rank	R^2_{lS}	Rank	R^2_{lQ}	Rank
1	TAC	0.33 (0.28, 0.39)	1	0.29 (0.2, 0.37)	1	0.29 (0.2, 0.37)	1	0.23 (0.17, 0.29)	2
2	MVPA	0.27 (0.23, 0.35)	3	0.25 (0.18, 0.32)	2	0.25 (0.18, 0.32)	2	0.26 (0.19, 0.35)	1
3	Age	0.29 (0.24, 0.36)	2	0.23 (0.15, 0.3)	3	0.23 (0.15, 0.3)	3	0.19 (0.12, 0.24)	4
4	ASTP	0.26 (0.22, 0.3)	4	0.22 (0.16, 0.31)	4	0.22 (0.16, 0.31)	4	0.19 (0.13, 0.25)	3
5	Albumin	0.11 (0.1, 0.14)	5	0.13 (0.08, 0.17)	5	0.13 (0.08, 0.17)	5	0.1 (0.06, 0.15)	5
6	Systolic BP	0.01 (0.01, 0.01)	6	0 (0, 0.02)	6	0 (0, 0.02)	6	0 (0, 0.01)	6

Table 3.2: R^2_l estimates and 95% confidence intervals for continuous predictors in NHANES 2003-2006.

3.7 Discussion

We used Semiparametric Gaussian Copula to define latent R^2_l, a measure of variation explained for the case of observed binary outcome and observed continuous predictor. Conceptually, R^2_l can be considered as a parameter of the data generating process that does not depend on the prevalence rate p and have an intuitive scale-invariant interpretation. The scale-dependence was considered to be a major limitation of other previously proposed R^2-type measures (DeMaris, 2002; Schemper, 2003; Steyerberg et al., 2009). Under SGC, AUC and R^2_l are directly related and their mutually consistent interpretation can provide a more complete description of both discrimination and dependence, especially, under highly unbalanced cases (Saito and Rehmsmeier, 2015). This property also allows us to compare AUCs (via converting to corresponding R^2_l) across studies with different prevalence of the binary outcome of interest. We showed that if $p = 0.5$, AUC and the latent correlation are

almost linearly related. However, once p gets smaller, the two measures exhibit a significant nonlinear divergence. Hence, for a fixed AUC, variation explained is getting smaller while the prevalence rate getting smaller. This is similar to examples considered in Chapter 15 of Steyerberg et al., 2009 that compared AUC vs likelihood-based Nagelkerke's R^2 for different values of the prevalence rate. Note that our proposal established an exact relationship between AUC and scale-invariant semiparametric R_l^2.

We also demonstrated how four rank statistics and prevalence rate can be used to estimate both AUC and the latent R_l^2. We proved that our weighted AUC estimators defined through Spearman and Quadrant correlations are asymptotically normal and hence, consistent under reasonable complex survey design assumptions. We additionally showed that AUC is sensitive to outliers and proposed A_Q, AUC calculated via Quadrant rank correlation, as a robust semiparametric version of AUC. Finally, we demonstrated how AUC can be calculated using only single participant survey weights under complex survey designs. As of interesting note, we showed that Kendall's Tau and Spearman rank correlations are linearly related in a binary-continuous case, which is in contrast to the continuous-continuous case where they are only asymptotically equivalent and the latter is a linear projection of the former (Sidak, Sen, and Hajek, 1999).

There are a few limitations in the proposed framework. First, the main assumption is that binary-continuous pairs are generated according to a Semiparametric Gaussian Copula. Even though SGC is a flexible framework, it is desirable to develop data-driven methods to test this assumption and be able

to detect deviations from it. Another limitation, that remains to be addressed in the future work, is that we do not handle in any way the presence of ties that may occur in practice even for continuous variables.

This work can be extended in many interesting ways. First, latent continuous random variables generating observed binary outcomes could be of interest by themselves and methods development for calculating best predictors for these random variables would be welcome. Second, the SGC approach has been recently extended to include truncated variables in (Yoon, Carroll, and Gaynanova, 2018) and some specific cases of ordinal variables in (Quan, Booth, and Wells, 2018). This opens up an opportunity to extend R_l^2 to a wider class of mixed data types. Similarly to Croux and Dehon, 2010 and Nikitin, 1995, the future work should investigate and compare the asymptotic efficiency of the proposed estimators of AUC and R_l^2. It also would be interesting to compare R_l^2 to other previously proposed R^2-type measures. Finally, extending R_l^2 to the multivariate mixed data type predictors would be a natural next step.

Acknowledgment: We would like to thank Professor Thomas Lumley for a discussion of our approach and pointing to his blog post on the use of Wilcoxon rank-sum statistics under complex designs. We would like to thank Dr. Barry Graubard and Dr. Wenliang Yao for sharing their R code with us and providing useful insights. We also would like to thank Professor Ravi Varadhan and Dr. Stas Kolenikov for early discussions of estimating AUC under complex survey designs. Finally, we would like to thank Professor Irina

Gaynanova for her insightful discussions of the SGC model and its extensions.

Data availability statement: The data that supports the findings of this study are available in the supplementary material of this article, as well as, in Github (https://github.com/Ddey07/AUC_R2/).

S1 Proofs concerning relation between AUC and Rank Statistics

S1.1 Derivations of the relationships between rank statistics and AUC

In the next discussion, we repeatedly use the fact that for continuous random variable X, $sgn(X) = 2I(X > 0) - 1$ with probability one.

$$
\begin{aligned}
r_K &= E((Y_i - Y_i')sgn(X_i - X_i')) \\
&= E(sgn(X_i - X_i')|(Y_i - Y_i') = 1)P((Y_i - Y_i') = 1) \\
&- E(sgn(X_i - X_i')|(Y_i - Y_i') = -1)P((Y_i - Y_i') = -1) \\
&= p(1-p)(E(sgn(X_i - X_i')|(Y_i - Y_i') = 1) - E(sgn(X_i - X_i')|(Y_i - Y_i') = -1)) \\
&= p(1-p)(2P(X_i > X_i'|(Y_i - Y_i') = 1) - 2P(X_i > X_i'|(Y_i - Y_i') = -1)) \\
&= 2p(1-p)(P(X_1 > X_0) - P(X_1 < X_0)) \\
&\implies |r_K| = 2p(1-p)|P(X_1 > X_0) - P(X_1 < X_0)|
\end{aligned}
\tag{S1}
$$

From the definition of AUC, we know that, $|P(X_1 > X_0) - P(X_1 < X_0)| =$

$(2A-1)$. Combining this fact with Equation (S1), we get -

$$A = \frac{1}{2} + \left| \frac{r_K}{4p(1-p)} \right| \tag{S2}$$

$$\begin{aligned}
W &= P(X \leqslant X_1) - P(X \leqslant X_0) \\
&= P(X \leqslant X_1|Y=0)P(Y=0) + P(X \leqslant X_1|Y=1)P(Y=1) - \\
&P(X \leqslant X_0|Y=0)P(Y=0) - P(X \leqslant X_0|Y=1)P(Y=1) \\
&= P(X_1 > X_0)(1-p) + P(X_1' \leqslant X_1)p - \\
&P(X_0' \leqslant X_0)(1-p) - P(X_1 < X_0)p \\
&= P(X_1 > X_0) + \frac{1}{2}p - \frac{1}{2}(1-p) - p \\
&[X_1 \stackrel{d}{=} X_1', X_0 \stackrel{d}{=} X_0', \therefore P(X_1 \leqslant X_1') = P(X_0 \leqslant X_0') = \frac{1}{2}] \\
&= P(X_1 > X_0) - \frac{1}{2} \\
\implies |W| &= A - \frac{1}{2} \\
\implies A &= \frac{1}{2} + |W|
\end{aligned} \tag{S3}$$

$$
\begin{aligned}
r_S &= 12E[F_X(X)F_Y(Y)] - 3 \\
&= 12E[F_X(X')\{(1-p)\mathbb{I}(Y'=0) + \mathbb{I}(Y'=1)\}] - 3 \\
&= 12E[F_X(X)] - 12pE[F_X(X')\mathbb{I}(Y'=0)] - 3 \\
&= 3 - 12pE[P(X \leqslant X'|X')\mathbb{I}(Y'=0)] \\
&= 3 - 12pE[E[\mathbb{I}(X \leqslant X')\mathbb{I}(Y'=0)|X', Y']] \\
&= 3 - 12pP(X \leqslant X', Y'=0) \\
\implies P(X \leqslant X', Y'=0) &= \frac{3-r_S}{12p}
\end{aligned}
\tag{S4}
$$

$$
\begin{aligned}
W &= P(X \leqslant X_1) - P(X \leqslant X_0) \\
&= P(X \leqslant X'|Y'=1) - P(X \leqslant X'|Y'=0) \\
&= \frac{P(X \leqslant X', Y'=1)}{p} - \frac{P(X \leqslant X', Y'=0)}{1-p} \\
&= \frac{P(X \leqslant X', Y'=1) + P(X \leqslant X', Y'=0)}{p} - P(X \leqslant X', Y'=0)(\frac{1}{p} + \frac{1}{1-p}) \\
&= \frac{1}{2p} - \frac{P(X \leqslant X', Y'=0)}{p(1-p)}
\end{aligned}
\tag{S5}
$$

Combining equation (S4) and (S5), we get -

$$\begin{aligned} W &= \frac{1}{2p} - \frac{\frac{3-r_S}{12p}}{p(1-p)} \\ &= \frac{r_S - (6p^2 - 6p + 3)}{12p^2(1-p)} \\ \implies A &= \frac{1}{2} + \left| \frac{r_S - (6p^2 - 6p + 3)}{12p^2(1-p)} \right| \end{aligned} \tag{S6}$$

S1.2 Derivation of bridging functions and proofs of Lemmas 1 and 2.

We lay out the proof of Lemma 1 below. Remember that $F_Y(y) = (1-p)\mathbb{I}(y < 1) + \mathbb{I}(y >= 1)$, so, $F_Y(Y) = (1-p)\mathbb{I}(Y < 1) + \mathbb{I}(Y >= 1) = p\mathbb{I}(U <= \Delta) + \mathbb{I}(U > \Delta)$. Also, $F_X(X) = \Phi(V)$, $\Phi(V) \sim U(0,1)$ and $E(\Phi(V)) = 0.5$. Hence,

$$\begin{aligned} r_S &= 12E[F_Y(Y)F_X(X)] - 3 \\ &= 12E[P(X_2 < X_1|X_1)P(Y_3 \leqslant Y_1|Y_1)] - 3 \\ &= 12E[P(X_2 < X_1|X_1)P(Y_3 \leqslant Y_1|Y_1)] - 3 \\ &= 12E[E[\mathbb{I}(X_2 < X_1, Y_3 \leqslant Y_1)|X_1, Y_1]] - 3 \\ &= 12P(X_2 < X_1, Y_3 \leqslant Y_1) - 3 \\ &= 12\gamma - 3 \end{aligned} \tag{S7}$$

Using the established relationship between Spearman's rank correlation and Type 2 concordance, γ, our next step is to derive the relationship between

Type 2 concordance and the latent correlation, r.

$$\gamma = P(X_2 < X_1, Y_3 \leqslant Y_1)$$

$$= P(V_2 < V_1, U_3 <= \Delta, U_1 > \Delta] + P(V_2 < V_1, U_3 > \Delta, U_1 > \Delta) +$$

$$P(V_2 < V_1, U_3 <= \Delta, U_1 <= \Delta)$$

$$= (1-p)P(\frac{(V_2 - V_1)}{\sqrt{(2)}} < 0, U_1 > \Delta) + pP(\frac{(V_2 - V_1)}{\sqrt{(2)}} < 0, U_1 > \Delta) + \tag{S8}$$

$$(1-p)P(\frac{(V_2 - V_1)}{\sqrt{(2)}} < 0, U_1 <= \Delta)$$

$$= (1-p)\Phi_2(0, -\Delta, \frac{r}{\sqrt{2}}) + p\Phi_2(0, -\Delta, \frac{r}{\sqrt{2}}) + p\Phi_2(0, \Delta, -\frac{r}{\sqrt{2}})$$

$$= \Phi_2(0, -\Delta, \frac{r}{\sqrt{2}}) + p\Phi_2(0, \Delta, -\frac{r}{\sqrt{2}})$$

Using Equations (S7) and (S8), we can conclude that

$$r_S = 12[\Phi_2(0, -\Delta, \frac{r}{\sqrt{2}}) + p\Phi_2(0, \Delta, -\frac{r}{\sqrt{2}})] - 3 = G_S(r) \tag{S9}$$

We derive the bridging function for Quadrant rank correlation as follows.

$$
\begin{aligned}
r_Q &= E[sgn((Y - M_Y)(X - M_X))] \\
&= P(U > \Delta, V > 0) - P(U > \Delta, V < 0)\mathbb{I}(M_Y = 0) + \\
&P(U < \Delta, V < 0) - P(U < \Delta, V > 0)\mathbb{I}(M_Y = 1) \qquad \text{(S10)} \\
&= [\Phi_2(-\Delta, 0, r) - \Phi_2(-\Delta, 0, -r)]\mathbb{I}(M_Y = 0) + \\
&[\Phi_2(\Delta, 0, r) - \Phi_2(\Delta, 0, -r)]\mathbb{I}(M_Y = 1) = G_Q(r)
\end{aligned}
$$

Fan et al., 2017 proved the following result in the appendix and used it to prove monotonicity and hence invertibility of $G_K(r)$. The result is stated as follows.

Lemma S1.1. *For any fixed* Δ_1, Δ_2, $\Phi_2(\Delta_1, \Delta_2, r) = \int_{-\infty}^{\Delta_1} \Phi(\frac{\Delta_2 - rx}{\sqrt{1-r^2}})\phi(x)dx$, *where* $\phi(x)$ *is the standard normal density, and moreover,* $\Phi_2(\Delta_1, \Delta_2, r)$ *is a strictly increasing function of* r *in* $(-1, 1)$. *Hence, inverse exists for the function.*

Following these ideas that have been used to prove the monotonicity of the bridging function $G_K(r)$, we will now prove Lemma 2.

Proof. Without the loss of generality, we assume that $M_Y = 0$, and define $F_\Delta(r) = \Phi_2(-\Delta, 0, r) = \Phi_2(0, -\Delta, r)$. Then, using Lemma 1, $\frac{\partial F_\Delta(r)}{\partial r} = F'_\Delta(r) > 0$ for $r \in (-1, 1)$. Also, we can write

$$
\begin{aligned}
G_Q(r) &= \Phi_2(-\Delta, 0, r) - \Phi_2(-\Delta, 0, -r) = F_\Delta(r) - F_\Delta(-r) \\
\implies \frac{\partial G_Q(r)}{\partial r} &= F'_\Delta(r) + F'_\Delta(-r) > 0
\end{aligned} \qquad \text{(S11)}
$$

Also using the fact that $\Phi_2(\Delta_1, \Delta_2, -r) = \Phi(\Delta_1) - \Phi(\Delta_1, -\Delta_2, r)$, we can derive

$$
\begin{aligned}
G_S(r) &= 12[\Phi_2(0, -\Delta, \frac{r}{\sqrt{2}}) + p\Phi_2(0, \Delta, -\frac{r}{\sqrt{2}})] - 3 \\
&= 12[(1-p)\Phi_2(0, -\Delta, \frac{r}{\sqrt{2}}) + \frac{p}{2}] - 3 \\
&= 12[(1-p)F_\Delta(\frac{r}{\sqrt{2}})] + (6p-3) \\
\implies \frac{\partial G_S(r)}{\partial r} &= \frac{12(1-p)}{\sqrt{2}} F'_\Delta(\frac{r}{\sqrt{2}}) > 0.
\end{aligned}
\tag{S12}
$$

This proves the statement.

S2 Proof of Theorem 1. (Asymptotics)

We establish asymptotical results under a sequence of finite-populations P_{N_ν} converging to a super-population P, following the framework proposed by Rubin-Bleuer, Kratina, et al., 2005. We assume that a finite sample of size n_ν is drawn from finite population of size N_ν according to a sampling design $p(s)$ and $N_\nu, n_\nu \to \infty$ with $limsup \frac{n_\nu}{N_\nu} < 1$ as $\nu \to \infty$. The superpopulation associated with N_ν is embedded with a probability space $(\Omega, \mathcal{F}, \xi)$. Below, all the distributions and convergence refer to the joint process of first choosing a finite population from the super-population and then drawing a sample from the finite population using a probability sampling. The combination of the two levels of convergence is inline with Theorem 6.1 of Rubin-Bleuer, Kratina, et al., 2005, where the authors considered the convergence of sample estimators defined by estimating equations. For next part of discussion, we denote the joint probability measure as ξp , where ξ denotes the probability

measure with respect to the model defined on finite population, and p denotes the measure with respect to the design conditioning on the finite population. We lay out the assumptions $1, 3, 4, 5$ from Wang, 2012 below under which we will prove our results.

Assumption 1. *The finite population P_N consists of a sequence of i.i.d. variables $(Y_i, X_i), i = 1, \cdots, N$.*

Assumption 2. *The following conditions hold for inclusion probabilities $w(i)$ and design variance of Horvitz–Thompson estimator of the mean - (i) $K_L \leqslant Nw(i)/n^* \leqslant K_U$ for all i , where K_L and K_U are positive constants, (ii) For any vector z_i with finite population moments, or equivalently, $\frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{z}\|^{2+\delta} < \infty$ where $\|\boldsymbol{z}\| = \sqrt{\boldsymbol{z}^T\boldsymbol{z}}$ denotes the L_2-norm of vector z, we assume $n^*\mathrm{Var}_p(\bar{\boldsymbol{z}}_w) \leqslant K_V$ for some $K_V > 0$ and $n^* = E_p(n)$.*

Assumption 3. *For any z with finite fourth population moment, $Var_p(\bar{z}_w)^{-\frac{1}{2}}(\bar{z}_w - \bar{z}_N)|\mathcal{F}_N$ converges to $N(0, \mathrm{I})$ with respect to $\mathcal{L}_p$ and $[\mathrm{Var}_p(\bar{\boldsymbol{z}}_w)]^{-1}\widehat{V}_p(\bar{\boldsymbol{z}}_w) - \mathrm{I} = O_p(n^{*^{-1/2}})$, where I is the identity matrix, the design variance–covariance matrix of $\bar{z}_w$, denoted by $[\mathrm{Var}_p(\bar{\boldsymbol{z}}_w)]$, is positive definite, and $\widehat{V}_p(\bar{\boldsymbol{z}}_w) = \frac{1}{N^2}\sum_{i\in A}\sum_{j\in A}\Omega_{ij}\boldsymbol{z}_i\boldsymbol{z}_j^T$, where Ω_{ij} means design-dependent weights associated with each pair (i, j).*

Here, convergence with respect to $\mathcal{L}_p$ means "convergence in the law of sampling design", conditioning on the realized population, $\bar{z}_w$ denotes the Horvitz-Thompson estimate of z from finite sample and $\bar{z}_N$ means the mean in the finite population.

Assumption 4. *Let $D_{t,N}$ denote the set of all distinct $(i_1, i_2, \cdots, i_t)$-tuples from P_N. We have*

$$limsup_{N\to\infty} \frac{N^4}{n^{*2}} \max_{(i_1,i_2,i_3,i_4)\in D_{4,N}} |\mathrm{E}_p(\mathrm{I}_{i_1} - w(i_1))(\mathrm{I}_{i_2} - w(i_2))(\mathrm{I}_{i_3} - w(i_3))(\mathrm{I}_{i_4} - w(i_4))| \leqslant M_1 < \infty$$

$$limsup_{N\to\infty} \frac{N^3}{n^{*2}} \max_{(i_1,i_2,i_3)\in D_{3,N}} |\mathrm{E}_p(\mathrm{I}_{i_1} - w(i_1))^2(\mathrm{I}_{i_2} - w(i_2))(\mathrm{I}_{i_3} - w(i_3))| \leqslant M_2 < \infty$$

$$limsup_{N\to\infty} \frac{N^2}{n^{*2}} \max_{(i_1,i_2)\in D_{2,N}} |\mathrm{E}_p(\mathrm{I}_{i_1} - w(i_1))^2(\mathrm{I}_{i_2} - w(i_2))^2| \leqslant M_3 < \infty \tag{S13}$$

almost surely for all populations. Here I_k *denotes the indicator that k-th unit is chosen in the finite sample.*

Now, we introduce the following notations for the Horvitz-Thompson estimators of population quantities from finite sample -

$$\begin{aligned}
\hat{F}_{0n}(t) &= \frac{1}{\sum_{i:Y_i=0} \frac{1}{w(i)}} \sum_{i:Y_i=0} \frac{1}{w(i)} I(X_i \leqslant t) \\
\hat{F}_{1n}(t) &= \frac{1}{\sum_{i:Y_i=1} \frac{1}{w(i)}} \sum_{i:Y_i=1} \frac{1}{w(i)} I(X_i \leqslant t) \\
\hat{F}_{n}(t) &= \frac{1}{\sum_{i=1}^{n} \frac{1}{w(i)}} \sum_{i=1}^{n} \frac{1}{w(i)} I(X_i \leqslant t) \\
\hat{p}_w &= \frac{\sum_{i:=1}^{n} \frac{1}{w(i)} \mathbb{I}(Y_i = 1)}{\sum_{i=1}^{n} \frac{1}{w(i)}}.
\end{aligned} \tag{S14}$$

We will also use the fact that $\hat{M}_Y = \mathbb{I}(\hat{p}_w > \frac{1}{2})$. We denote the true distribution functions of the random variables $X, X_0 = (X|Y = 0), X_1 = (X|Y = 1)$ as F_X, F_{0X} and F_{1X} respectively. Now, as the random variable $\mathbb{I}(Y_i = 1)$

has finite fourth moment, using Assumption $A3$, similar to proof of Theorem 1 in Wang, 2012, it can be shown that, $\sqrt{n}((\hat{p}_w, \hat{F}_{0n}(t), \hat{F}_{1n}(t), \hat{F}_n(t))' - (p, F_{0X}(t), F_{1X}(t), F_X(t))')$ converges weakly to $N(0, \Sigma)$ for some covariance matrix Σ that depends on second-order design probabilities. The convergence of $\hat{p}_w$, the Horvitz-Thompson estimator of the sample mean, is discussed thoroughly in Corollary 1.3.6.1 of Fuller, 2011. We can treat the distribution functions as random elements in $\mathcal{D}[0,1]$, the space of all right continuous functions defined on $[0,1]$, and $\hat{p}_w$ as a constant process. Then similar to Lumley and Scott, 2013 and Wang, 2012, we can infer that $\sqrt{n}((\hat{p}_w, \hat{F}_{0n}, \hat{F}_{1n}, \hat{F}_n)' - (p, F_{0X}, F_{1X}, F_X)')$ converges weakly to the Gaussian process $T = (T_{w_1}, T_{w_2}, T_{w_3}, T_{w_4})$ with the covariance kernel that depends on second order sampling probabilities. Following Example 20.12 in Vaart, 2000

and assuming $g(x) = I(x > \frac{1}{2})$, we can rewrite $\hat{r}_Q$ and $\hat{r}_S$ as follows

$$
\begin{aligned}
\hat{r}_Q &= \frac{1}{\sum_{i=1}^{n} \frac{1}{w(i)}} \sum_{i=1}^{n} \frac{1}{w(i)} sgn((Y_i - \hat{M}_Y)(X_i - \hat{M}_X)) \\
&= \frac{1}{\sum_{i:Y_i=1} \frac{1}{w(i)}} \hat{p}_w \mathbb{I}(\hat{M}_Y = 0) \sum_{i:Y_i=1} \frac{1}{w(i)} 2(\mathbb{I}(\hat{F}_X(X_i) > \frac{1}{2}) - 1) - \\
&\frac{1}{\sum_{i:Y_i=0} \frac{1}{w(i)}} (1 - \hat{p}_w) \mathbb{I}(\hat{M}_Y = 1) \sum_{i:Y_i=0} \frac{1}{w(i)} 2(\mathbb{I}(\hat{F}_X(X_i) > \frac{1}{2}) - 1) \\
&= \frac{1}{\sum_{i:Y_i=1} \frac{1}{w(i)}} \hat{p}_w \mathbb{I}(\hat{p}_w \leqslant \frac{1}{2}) \sum_{i:Y_i=1} \frac{1}{w(i)} 2(\mathbb{I}(\hat{F}_X(X_i) > \frac{1}{2}) - 1) - \\
&\frac{1}{\sum_{i:Y_i=0} \frac{1}{w(i)}} (1 - \hat{p}_w) \mathbb{I}(\hat{p}_w > \frac{1}{2}) \sum_{i:Y_i=0} \frac{1}{w(i)} 2(\mathbb{I}(\hat{F}_X(X_i) > \frac{1}{2}) - 1) \\
&= \hat{p}_w \mathbb{I}(\hat{p}_w \leqslant \frac{1}{2}) \int g(\hat{F}_n) d\hat{F}_{1n} - (1 - \hat{p}_w) \mathbb{I}(\hat{p}_w > \frac{1}{2}) \int g(\hat{F}_n) d\hat{F}_{0n} - \\
&- (\hat{p}_w \mathbb{I}(\hat{p}_w \leqslant \frac{1}{2})(1 - \hat{p}_w) \mathbb{I}(\hat{p}_w > \frac{1}{2}))
\end{aligned}
\tag{S15}
$$

$$
\begin{aligned}
\hat{r}_S &= 12 \frac{1}{\sum_{i=1}^{n} \frac{1}{w(i)}} \sum_{i=1}^{n} \frac{1}{w(i)} [\hat{F}_Y(Y_i) \hat{F}_X(X_i)] - 3 \\
&= 12[(1 - \hat{p}_w)^2 \int \hat{F}_n d\hat{F}_{0n} + \hat{p}_w \int \hat{F}_n d\hat{F}_{1n}] - 3
\end{aligned}
\tag{S16}
$$

where $g(x) = I(x > \frac{1}{2})$ in Equation S15. In Theorem 1 of Lumley and Scott, 2013, the convergence of $\int g(\hat{F}_n) d\hat{F}_{1n} - \int g(\hat{F}_n) d\hat{F}_{0n}$ has been proved, where the function $g()$ has to be, (i) differentiable with bounded derivative and continuous on $[0, 1]$, or, (ii) differentiable on $(0, 1)$ and $\int_{0,1} g(y)^{2+\delta}$ is finite for some $\delta > 0$, or, (iii) an indicator function of a subinterval (a, b)

of $[0, 1]$. Our function g satisfies criteria (iii), so, proceeding similarly as the proof of Theorem 1 in Lumley and Scott, 2013 and applying functional delta method (Kosorok, 2008, Theorem 12.1) to Equation S15 S16, we get that $\sqrt{n}(\hat{r}_Q - r_Q)$ converges to normal distribution. We should keep in mind that, the functional delta method requires Hadamard differentiability of the functional at (p, F_{0X}, F_{1X}, F_X) and for the quadrant correlation (Equation S15), the functional is not Hadamard differentiabile at $p = \frac{1}{2}$, so, we leave that special case out of the proof.

Proving asymptotic normality of $\sqrt{n}(\hat{r}_S - r_S)$ is more straight-forward as we can apply functional delta method to Equation (S16) and it will follow immediately.

S3 Additional Figures

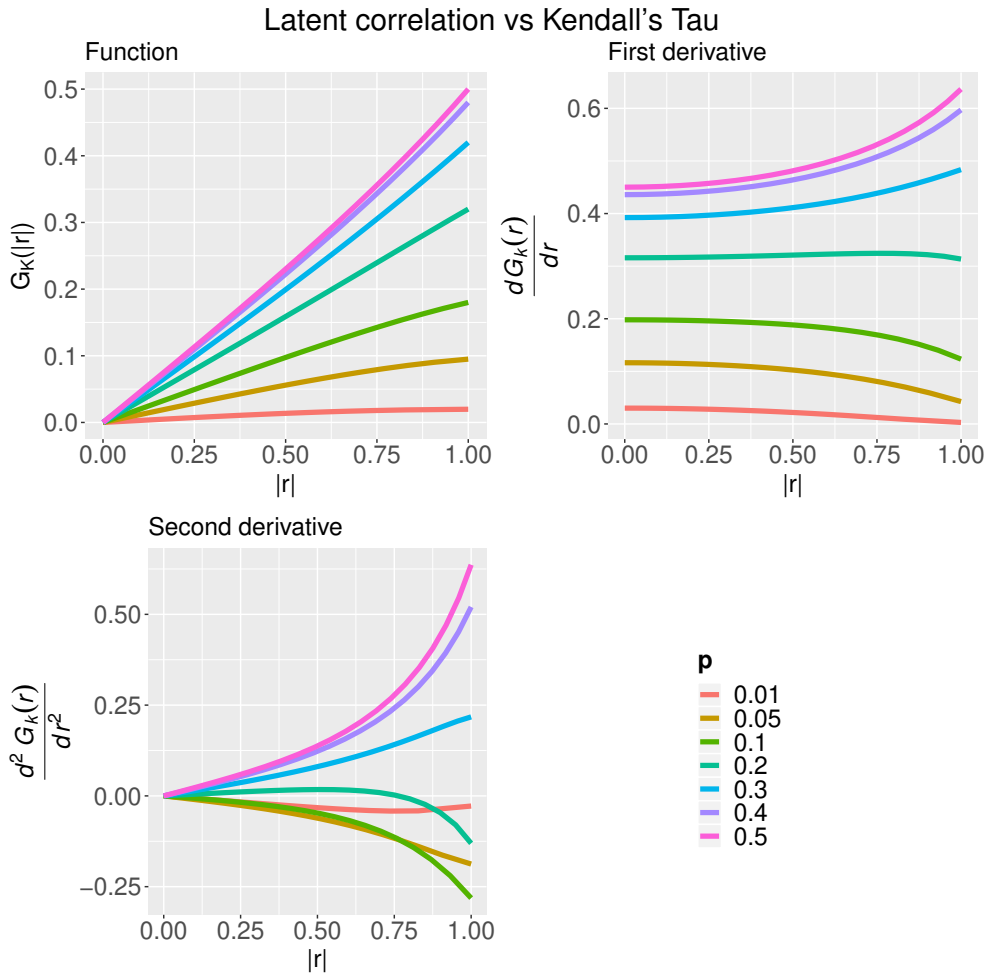


Figure S1: The relationships between the absolute value of latent correlation and Kendall's Tau, varying on p.

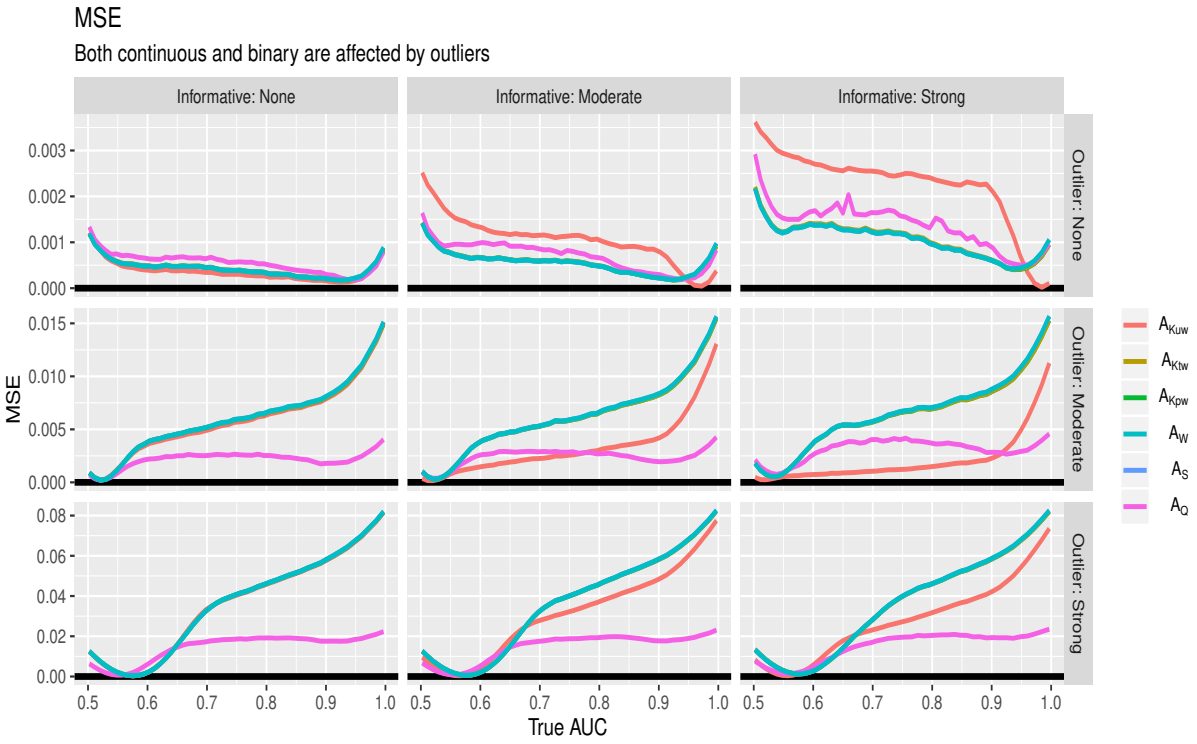


Figure S2: Simulation results: MSE of the estimators under different scenario

.

References

Kuhn, Max and Kjell Johnson (2013). *Applied predictive modeling*. Vol. 26. Springer.

Bishop, Christopher M (2006). *Pattern recognition and machine learning*. springer.

Steyerberg, Ewout W et al. (2009). *Clinical prediction models*. Vol. 381. Springer.

Steyerberg, Ewout W, Andrew J Vickers, Nancy R Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J Pencina, and Michael W Kattan (2010). "Assessing the performance of prediction models: a framework for some traditional and novel measures". In: *Epidemiology (Cambridge, Mass.)* 21.1, p. 128.

Harrell Jr, Frank E (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

Saito, Takaya and Marc Rehmsmeier (2015). "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets". In: *PloS one* 10.3, e0118432.

Kendall, Maurice George, Alan Stuart, John Keith Ord, Steven F Arnold, Anthony O'Hagan, and Jonathan Forster (1987). *Kendall's advanced theory of statistics*. Vol. 1. Griffin London.

Lobo, Jorge M, Alberto Jiménez-Valverde, and Raimundo Real (2008). "AUC: a misleading measure of the performance of predictive distribution models". In: *Global ecology and Biogeography* 17.2, pp. 145–151.

Tutz, Gerhard (2011). *Regression for categorical data*. Vol. 34. Cambridge University Press.

DeMaris, Alfred (2002). "Explained variance in logistic regression: A Monte Carlo study of proposed measures". In: *Sociological Methods & Research* 31.1, pp. 27–74.

Schemper, Michael (2003). "Predictive accuracy and explained variation". In: *Statistics in medicine* 22.14, pp. 2299–2308.

Yazici, Berna, Özlem Alpu, and Yaning Yang (2007). "Comparison of goodness-of-fit measures in probit regression model". In: *Communications in Statistics—Simulation and Computation®* 36.5, pp. 1061–1073.

Fan, Jianqing, Han Liu, Yang Ning, and Hui Zou (2017). "High dimensional semiparametric latent graphical model for mixed data". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.2, pp. 405–421.

Nelsen, Roger B (2007). *An introduction to copulas*. Springer Science & Business Media.

Joe, Harry (2014). *Dependence modeling with copulas*. Chapman and Hall/CRC.

MacCallum, Robert C, Shaobo Zhang, Kristopher J Preacher, and Derek D Rucker (2002). "On the practice of dichotomization of quantitative variables." In: *Psychological methods* 7.1, p. 19.

Sidak, Zbynek, Pranab K Sen, and Jaroslav Hajek (1999). *Theory of rank tests*. Elsevier.

Croux, Christophe and Catherine Dehon (2010). "Influence functions of the Spearman and Kendall correlation measures". In: *Statistical methods & applications* 19.4, pp. 497–515.

Lumley, Thomas and Alastair J Scott (2013). "Two-sample rank tests under complex sampling". In: *Biometrika* 100.4, pp. 831–842.

Liu, Han, Fang Han, Ming Yuan, John Lafferty, Larry Wasserman, et al. (2012). "High-dimensional semiparametric Gaussian copula graphical models". In: *The Annals of Statistics* 40.4, pp. 2293–2326.

Korn, Edward L and Barry I Graubard (2011). *Analysis of health surveys*. Vol. 323. John Wiley & Sons.

Yao, Wenliang, Zhaohai Li, and Barry I Graubard (2015). "Estimation of ROC curve with complex survey data". In: *Statistics in medicine* 34.8, pp. 1293–1303.

Leroux, Andrew, Junrui Di, Ekaterina Smirnova, Elizabeth J Mcguffey, Quy Cao, Elham Bayatmokhtari, Lucia Tabacu, Vadim Zipunnikov, Jacek K Urbanek, and Ciprian Crainiceanu (2019). "Organizing and analyzing the activity data in NHANES". In: *Statistics in Biosciences*, pp. 1–26.

Smirnova, Ekaterina, Andrew Leroux, Quy Cao, Lucia Tabacu, Vadim Zipunnikov, Ciprian Crainiceanu, and Jacek Urbanek (2019). "The predictive performance of objective measures of physical activity derived from accelerometry data for 5-year all-cause mortality in older adults: NHANES 2003-2006". In: *The Journals of Gerontology: Series A*.

Yeo, Douglas, Harold Mantel, and Tzen-Ping Liu (1999). "Bootstrap variance estimation for the national population health survey". In: *American Statistical Association, Proceedings of the Survey Research Methods Section*. Citeseer.

Varma, Vijay R, Debangan Dey, Andrew Leroux, Junrui Di, Jacek Urbanek, Luo Xiao, and Vadim Zipunnikov (2018). "Total volume of physical activity: TAC, TLAC or TAC (λ)". In: *Preventive medicine* 106, p. 233.

Yoon, Grace, Raymond J Carroll, and Irina Gaynanova (2018). "Sparse semiparametric canonical correlation analysis for data of mixed types". In: *arXiv preprint arXiv:1807.05274*.

Quan, Xiaoyun, James G Booth, and Martin T Wells (2018). "Rank-based approach for estimating correlations in mixed ordinal data". In: *arXiv preprint arXiv:1809.06255*.

Nikitin, Yakov (1995). *Asymptotic Efficiency of Nonparametric Tests*. Cambridge University Press. DOI: 10.1017/CBO9780511530081.

Rubin-Bleuer, Susana, Ioana Schiopu Kratina, et al. (2005). "On the two-phase framework for joint model and design-based inference". In: *The Annals of Statistics* 33.6, pp. 2789–2810.

Wang, Jianqiang C (2012). "Sample distribution function based goodness-of-fit test for complex surveys". In: *Computational Statistics & Data Analysis* 56.3, pp. 664–679.

Fuller, Wayne A (2011). *Sampling statistics*. Vol. 560. John Wiley & Sons.

Vaart, Aad W Van der (2000). *Asymptotic statistics*. Vol. 3. Cambridge university press.

Kosorok, Michael R (2008). *Introduction to empirical processes and semiparametric inference*. Springer.

Chapter 4

Graphical Gaussian Process Models for Highly Multivariate Spatial Data

4.1 Introduction

Multivariate spatial data abound in the natural and environmental sciences for studying features of the joint distribution of multiple spatially dependent variables (see, for example, Wackernagel, 2013; Cressie and Wikle, 2011; Banerjee, Carlin, and Gelfand, 2014). The objectives are to estimate associations over spatial locations for each variable and those among the variables. Let $y(s)$ be a $q \times 1$ vector of spatially-indexed dependent outcomes within any location $s \in \mathcal{D} \subset \mathbb{R}^d$ with $d = 2$ or 3. A multivariate spatial regression model on our spatial domain $\mathcal{D}$ specifies a univariate spatial regression model for each outcome as

$$y_i(s) = x_i(s)^{\mathrm{T}}\beta_i + w_i(s) + \epsilon_i(s)\ , \quad i = 1, 2, \ldots, q,\ s \in \mathcal{D} \tag{4.1}$$

where $y_i(s)$ is the i-th element of $y(s)$, $x_i(s)$ is a $p_i \times 1$ vector of predictors, β_i is the $p_i \times 1$ vector of slopes, each $w_i(s)$ is a spatial process and $\epsilon_i(s) \overset{ind}{\sim}$

$N(0, \tau_i^2)$ is the random noise in outcome i. We customarily assume that $w(s) = (w_1(s), w_2(s), \ldots, w_q(s))^{\mathrm{T}}$ is a multivariate Gaussian process (GP) specified by a zero mean and a cross-covariance function that introduces dependence over space and among the q variables. The cross-covariance is a matrix-valued function $C = (C_{ij}) : \mathcal{D} \times \mathcal{D} \mapsto \mathbb{R}^{q \times q}$ with $C_{ij}(s, s') = \mathrm{Cov}(w_i(s), w_j(s'))$ for any pair of locations (s, s'). Cross-covariance functions must ensure that for any finite set of locations $\mathcal{S} = \{s_1, \ldots, s_n\}$, the $nq \times nq$ matrix $C(\mathcal{S}, \mathcal{S}) = (C(s_i, s_j))$ is positive definite (p.d.).

Valid classes of cross-covariance functions have been comprehensively reviewed in Genton and Kleiber, 2015. Of particular interest are multivariate Matérn cross-covariance functions (Gneiting, Kleiber, and Schlather, 2010; Apanasovich, Genton, and Sun, 2012), where the marginal covariance functions for each $w_i(s)$ and the cross-covariance functions between $w_i(s)$ and $w_j(s')$ are Matérn functions. In its most general form, the multivariate Matérn is appealing as it ensures that each univariate process is a Matérn GP with its own range, smoothness and spatial variance although the parameters need to be constrained to ensure positive-definiteness of the cross-covariance function.

Our current focus is the increasingly commonplace *highly-multivariate* setting with a large number of dependent outcomes (e.g., $q \sim 10^2+$) at each spatial location. While substantial attention has been accorded to spatial data with massive number of locations (large n) (see, e.g., Heaton et al., 2019, for a review), the highly multivariate setting fosters separate computational issues. Likelihoods for popular cross-covariance functions, such as the multivariate Matérn, involve $O(q^2)$ parameters, and $O(q^3)$ floating point operations (flops).

Optimizing over or sampling from high-dimensional parameter spaces is inefficient even for modest values of n. Illustrations of multivariate Matérn models have typically been restricted to applications with $q \leqslant 5$.

In non-spatial settings, Gaussian graphical models are extensively used as a dimension-reduction tool to parsimoniously model conditional dependencies in highly multivariate data. any exploitable graphical structure for scalable computation, nor do they adhere to posited conditional independence relations among the outcomes as are often introduced in high-dimensional outcomes (Cox and Wermuth, 1996). Our innovation here is to develop multivariate GPs that conform to *process-level conditional independence* posited by an inter-variable graph over q dependent outcomes while attending to scalability considerations for large q.

To adapt graphical models to multivariate spatial process-based settings, we generalize notions of process-level conditional independence for discrete time-series (Dahlhaus, 2000; Dahlhaus and Eichler, 2003) to continuous spatial domains. We define multivariate *graphical Gaussian Processes (GGPs)* that satisfy process-level conditional independence as specified by an inter-variable graph. We focus on GGPs with properties deemed critical for handling multivariate spatial data. Specifically, we seek to retain the flexibility to model and interpret spatial properties of the random field for each variable separately. Except for the multivariate Matérn, most other multivariate covariance functions fail to retain this property.

We address and resolve challenges in constructing spatial processes that retain marginal properties and are also GGP. For example, while the existing

multivariate Matérn models preserve the univariate marginals as Matérn GPs, we show (Section 4.3.1) that no parametrisation of the multivariate Matérn yields a GGP. On the other hand, the literature on graphical multivariate discrete time-series models, hitherto, have not attempted to preserve marginal properties and have benefited from the regular discrete setting of equispaced time-points, in both non-parametric (Dahlhaus, 2000; Dahlhaus and Eichler, 2003; Eichler, 2008) and parametric (Eichler, 2012) analysis. We resolve both of these challenges for irregular spatial data.

Our development relies upon the seminal work of Dempster, 1972 on *covariance selection*, which ensures the existence of multivariate distributions that retain univariate marginals while satisfying conditional-independence relations specified by an inter-variable graph. While covariance selection can facilitate approximate likelihood-based inference for graphical VAR models (Eichler, 2012) by exploiting the expansion of the inverse spectral density matrix of VAR(p) models in terms of the inverse covariance matrices over finite (p) time-lags, such finite-lag representations do not typically hold for spatial covariance functions over $\mathcal{D} \subset \mathbb{R}^d$.

One of our key contributions here is to identify the construction of a marginal-retaining GGP as a *process-level covariance selection* problem. We use covariance selection on the spectral density matrix to prove existence, uniqueness and information-theoretic optimality of a marginal retaining GGP. We subsequently introduce a novel practicable method to approximate this optimal GGP by *stitching* GPs together using an inter-variable graph. Stitching relies on the orthogonal decomposition of a GP into a fixed-rank predictive

process (Banerjee et al., 2008) on a finite set of locations and a residual process. We show how to endow the predictive process with the desired conditional-independence structure via covariance selection, and use componentwise-independent residual processes to create a well defined multivariate GP that exactly preserves (i) dependencies modelled by the graph; and (ii) the marginal distributions on the entire domain. Stitching with Matérn GPs yields a *multivariate graphical Matérn GP* with a tractable likelihood for irregular spatial data such that (i) each outcome process is endowed with the original Matérn GP; (ii) we retain process-level conditional independence modelled by the graph; (iii) cross-covariances for variable pairs included in the graph are exactly or approximately Matérn.

We also demonstrate computational scalability with respect to q. We show that for decomposable graphical models, stitching facilitates drastic dimension-reduction of the parameter space and fast likelihood evaluations by obviating large matrix operations. Additionally, stitching harmonizes graphical models with parallel computing to employ a chromatic Gibbs sampler for delivering efficient fully model-based Bayesian inference. We also show how our framework can adapt to (i) deliver inference for an unknown inter-variable graph; (ii) model spatial time-series; and (iii) model multivariate spatial factor models.

4.2 Method

4.2.1 Process-level conditional independence and Graphical Gaussian Processes

We define *process-level conditional independence* for a multivariate GP $w(\cdot) = (w_1(\cdot), \ldots, w_q(\cdot))^{\mathrm{T}}$ over $\mathcal{D}$. We adapt the analogous definition for multivariate discrete time-series in Dahlhaus, 2000 to a continuous-space paradigm. Let $\mathcal{V} = \{1, \ldots, q\}$, $B \subset \mathcal{V}$ and $w_B(\mathcal{D}) = \{w_k(s) : k \in B,\ s \in \mathcal{D}\}$. Two processes $w_i(\cdot)$ and $w_j(\cdot)$ are conditionally independent given the processes $\{w_k(\cdot) \,|\, k \in V \backslash \{i,j\}\}$ if $\mathrm{Cov}(z_{iB}(s), z_{jB}(s')) = 0$ for all $s, s' \in \mathcal{D}$ and $B = \mathcal{V} \backslash \{i,j\}$, where $z_{kB}(s) = w_k(s) - \mathrm{E}[w_k(s) \,|\, \sigma(\{w_j(s') : j \in B,\ s' \in \mathcal{D}\})]$, where $\sigma(\cdot)$ is the usual σ-algebra generated by its argument. Let $\mathcal{G}_\mathcal{V} = (\mathcal{V}, E_\mathcal{V})$ be a graph, where $E_\mathcal{V}$ is a pre-specified set of edges among pairs of variables. We now define a *Graphical Gaussian Process* (GGP) with respect to (or conforming to) $\mathcal{G}_\mathcal{V}$ as follows.

Definition 4.2.1. [Graphical Gaussian Process] A $q \times 1$ GP $w(\cdot)$ is a Graphical Gaussian Process (GGP) with respect to a graph $\mathcal{G}_\mathcal{V} = (\mathcal{V}, E_\mathcal{V})$ when the univariate GPs $w_i(\cdot)$ and $w_j(\cdot)$ are conditionally independent for every $(i,j) \notin E_\mathcal{V}$. We denote such a process as GGP($\mathcal{G}_\mathcal{V}$).

Any collection of q independent GPs will trivially constitute a GGP with respect to any graph $\mathcal{G}_\mathcal{V}$. More pertinent is the ability of a GGP to approximate a full (non-graphical) GP. This is particularly relevant for inference because the full GP is computationally impracticable for large q. Theorem 4.2.1 shows that given a graph $\mathcal{G}_\mathcal{V}$ and a multivariate GP with cross-covariance function

C, there exists a *unique* and information-theoretically optimal GGP among the class of all GGP($\mathcal{G}_\mathcal{V}$). Proofs of all subsequent results are provided in the supplement.

Theorem 4.2.1. *Let $\mathcal{G}_\mathcal{V} = (\mathcal{V}, E_\mathcal{V})$ be any given graph, $C = (C_{ij})$ be a $q \times q$ stationary cross-covariance function. Let $F(\omega) = (f_{ij}(\omega))$ be the spectral density matrix corresponding to C at frequency ω. Let $f_{ii}(\cdot)$ be square-integrable for all i. Then*

(a) There exists a unique $q \times 1$ GGP($\mathcal{G}_\mathcal{V}$) $w(\cdot)$ with cross-covariance function $M = (M_{ij})$ such that $M_{ij} = C_{ij}$ for $i = j$ and for all $(i, j) \in E_\mathcal{V}$;

(b) If $\tilde{F}(\omega)$ denotes the spectral density matrix of $w(\cdot)$ and $\mathcal{F}$ is the set of spectral density matrices of all possible GGP($\mathcal{G}_\mathcal{V}$), then

$$\tilde{F}(\cdot) = \arg\min_{K(\cdot)\in\mathcal{F}} \int_\omega d_{KL}(F(\omega)\|K(\omega))d\omega \ ,$$

where $d_{KL}(F\|K) = \text{tr}(K^{-1}F) + \log\det(K)$ denotes the Kullback-Leibler divergence between two positive definite matrices F and K.

Theorem 4.2.1 shows that the optimal GGP approximating a GP, given a graph, needs to exactly preserve the marginal distributions of the univariate processes, which is also critical to retain interpretation of the spatial properties of each univariate surface. This optimal GGP also preserves cross-covariances for variable pairs included in $\mathcal{G}_\mathcal{V}$. Theorem 4.2.1, however, is of limited practical value because it does not present a convenient way to construct cross-covariances. We develop a practicable method of *stitching* q univariate random fields (Section 4.2.2) to construct marginal-preserving GGPs for modelling

irregular spatial data.

4.2.2 Stitching of Gaussian Processes

Given any $\mathcal{G}_{\mathcal{V}}$ and a cross-covariance function C, we seek a multivariate GP $w(\cdot)$ that

(i) exactly preserves the marginal distributions specified by C, i.e., $w_i(\cdot) \sim GP(0, C_{ii})\ \forall i$;

(ii) is a GGP($\mathcal{G}_{\mathcal{V}}$), i.e., satisfies process-level conditional independence according to $\mathcal{G}_{\mathcal{V}}$; and

(iii) exactly or approximately retains the cross-covariances specified by C for pairs of variables included in $\mathcal{G}_{\mathcal{V}}$, i.e., for $(i, j) \in E_{\mathcal{V}}$, $\text{Cov}(w_i(s), w_j(s')) \approx C_{ij}(s, s')$.

We visually illustrate stitching of univariate GPs to build a GGP $w(\cdot)$, satisfying (i)-(iii) above. Figure 4.1 (left) shows realizations of 4 univariate Matérn GPs $w_i(\cdot)$, $i = 1, \ldots, 4$, each with a different smoothness and spatial range. Figure 4.1 (right) shows a multivariate GGP constructed by stitching together the 4 processes using a path-graph as $\mathcal{G}_{\mathcal{V}}$ with $E_{\mathcal{V}} = \{(i, i+1) : i = 1, 2, 3\}$. We begin our construction on $\mathcal{L}$, a finite but otherwise arbitrary set of locations in $\mathcal{D}$ (the 3 locations in Figure 4.1 (right)). We first ensure that $w(\mathcal{L}) = (w_1(\mathcal{L}), \ldots, w_q(\mathcal{L}))^{\text{T}}$satisfies conditions (i)-(iii) when the domain is restricted to $\mathcal{L}$. This is achieved by stitching together the variables at the 3 locations in $\mathcal{L}$ such that there is a *thread* (edge) between two variable-location pairs if and only if there is an edge between the two corresponding variables in $\mathcal{V}$. We then stitch each of the remaining surfaces independently so that they

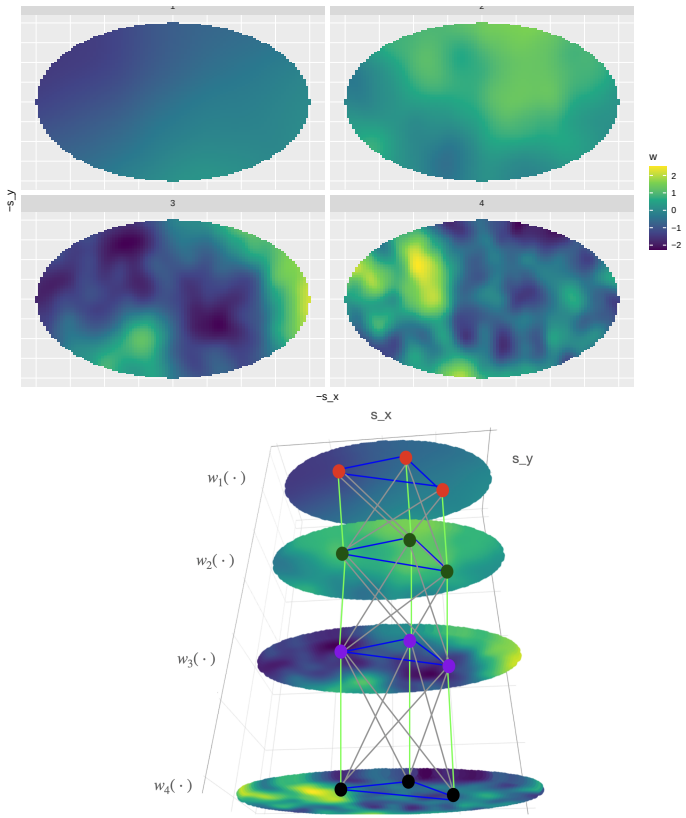


Figure 4.1: Stitching Gaussian Processes. Left: Realizations of 4 univariate GPs. Right: Realization of a multivariate (4-dimensional) GGP created by stitching together the 4 univariate GPs from the left figure using the strong product graph over the 4 variables and 3 locations.

have the same distribution as the univariate surfaces from the left panel and conforms to the graph at the process-level. This resembles stitching the four surfaces together at the locations $\mathcal{L}$, while exactly preserving each univariate surface. The graph edges serve as the threads holding the surfaces together.

Turning to the formal development, we first create $w(\mathcal{L})$—the realisation of our target process $w(\cdot)$ on $\mathcal{L}$ that satisfies properties (i)-(iii) on $\mathcal{L}$. Combining the three requirements, we model $w(\mathcal{L}) \sim N(0, M(\mathcal{L}, \mathcal{L}))$, seeking a p.d. matrix $M(\mathcal{L}, \mathcal{L})$ such that

(a) $M_{ii}(\mathcal{L}, \mathcal{L}) = C_{ii}(\mathcal{L}, \mathcal{L})$ for all $i = 1, \ldots, q$, to satisfy (i);

(b) $(M(\mathcal{L},\mathcal{L})^{-1})_{ij} = 0$ for all $(i,j) \notin E_{\mathcal{V}}$ to satisfy (ii).

(c) $M_{ij}(\mathcal{L},\mathcal{L}) = C_{ij}(\mathcal{L},\mathcal{L})$ for all $(i,j) \in E_{\mathcal{V}}$, to satisfy (iii).

Existence of such a matrix $M(\mathcal{L},\mathcal{L})$ is a covariance selection problem (Dempster, 1972).

Lemma 4.2.2 (Covariance selection (Dempster, 1972))**.** *Given a graph $\mathcal{G} = (\mathcal{S}, E)$ and any p.d. matrix $F = (F_{rs})$ indexed by $\mathcal{S} \times \mathcal{S}$, there exists a unique p.d. matrix $\widetilde{F} = (\widetilde{F}_{rs})$ such that $\widetilde{F}_{rs} = F_{rs}$ for $r = s$ or for $(r,s) \in E_{\mathcal{S}}$, and $(\widetilde{F}^{-1})_{rs} = 0$ for $(r,s) \notin E_{\mathcal{S}}$.*

To ensure that the covariances and cross-covariances are preserved over $\mathcal{L}$ for all i and all $(i,j) \in E_{\mathcal{V}}$ and the conditional independence among elements of $w(\mathcal{L})$ are inherited from $\mathcal{G}_{\mathcal{V}}$, $w(\mathcal{L})$ needs to conform to a graph with edges between variable-location pairs as in Figure 4.1. Formally, let $\mathcal{G}_{\mathcal{L}} = (\mathcal{L}, E_{\mathcal{L}})$ be the complete graph on the set of locations $\mathcal{L}$. The variable-location graph from Figure 4.1 (right) is the *strong product graph* $\mathcal{G}_{\mathcal{V}} \boxtimes \mathcal{G}_{\mathcal{L}}$. Here, $\mathcal{G}_V \boxtimes \mathcal{G}_L = (\mathcal{V} \times \mathcal{L}, E_{\mathcal{V}\times\mathcal{L}})$ with $\mathcal{V} \times \mathcal{L} = \{(i,l) : i \in \mathcal{V},\ l \in \mathcal{L}\}$ and $E_{\mathcal{V}\times\mathcal{L}}$ comprises edges between vertex-pairs (i,l) and (i',l') based upon the following strong-product adjacency rules: (i) $i = i'$ and $(l,l') \in E_{\mathcal{L}}$; or (ii) $l = l'$ and $(i,i') \in E_{\mathcal{V}}$; or (iii) $(i,i') \in E_{\mathcal{V}}$ and $(l,l') \in E_{\mathcal{L}}$.

Applying Lemma 4.2.2 with the vertex set $\mathcal{S} = \mathcal{V} \times \mathcal{L}$, positive definite matrix $F = C(\mathcal{L},\mathcal{L})$ and the graph $\mathcal{G}_V \boxtimes \mathcal{G}_L$, ensures the existence and uniqueness of a positive definite matrix $\widetilde{F} = M(\mathcal{L},\mathcal{L})$ satisfying conditions (a), (b) and (c) above. In practice, $M(\mathcal{L},\mathcal{L})$ can be obtained using an iterative proportional scaling (IPS) algorithm (Speed, Kiiveri, et al., 1986; Xu, Guo, and He, 2011).

Note that Condition (b) only ensures conditional independence of the process restricted to $\mathcal{L}$. Process-level conditional independence over the entire domain $\mathcal{D}$ follows from the subsequent extension in (4.2) as proved in Theorem 4.2.4. Having built the finite-dimensional distribution of $w(\mathcal{L})$ from $\mathcal{G}_{\mathcal{V}} \boxtimes \mathcal{G}_{\mathcal{L}}$, we now suitably extend it to a well-defined multivariate GP $w(\cdot)$ over the domain $\mathcal{D}$, which conforms to the conditional dependencies implied by $\mathcal{G}_{\mathcal{V}}$. We leverage the following well-known decomposition of a GP $w_i(\cdot)$ as sum of a finite rank *predictive process* $w_i^*(\cdot) = E(w_i(\cdot) \,|\, w_i(\mathcal{L}))$ and an independent *residual process* $z_i(\cdot)$ (Banerjee et al., 2008; Finley et al., 2009):

$$w_i(s) = w_i^*(s) + z_i(s) = C_{ii}(s, \mathcal{L})C_{ii}(\mathcal{L}, \mathcal{L})^{-1}w_i(\mathcal{L}) + z_i(s) \quad \text{for all} \quad s \in \mathcal{D}\backslash\mathcal{L}\ , \tag{4.2}$$

where each $z_i(\cdot)$ is a zero-centred Gaussian Process, independent of $w(\mathcal{L})$, with the valid covariance function $C_{ii\,|\,\mathcal{L}}(s, s') = C_{ii}(s, s') - C_{ii}(s, \mathcal{L})C_{ii}^{-1}(\mathcal{L}, \mathcal{L})C_{ii}(\mathcal{L}, s')$.

The first part of stitching ensures that $w(\mathcal{L})$ conforms to $\mathcal{G}_{\mathcal{V}}$ when restricted to $\mathcal{L}$. The next result establishes process-level conditional independence for the stitched predictive process.

Lemma 4.2.3. *The predictive process $w^*(\cdot) = (w_1^*(\cdot), \ldots, w_q^*(\cdot))^{\mathrm{T}}$ is a GGP($\mathcal{G}_{\mathcal{V}}$) on $\mathcal{D}$.*

We now extend the finite-rank GGP $w^*(\cdot)$ to a full-rank GGP $w(\cdot)$ over the entire domain $\mathcal{D}$ through (4.2). We construct $z_i(\cdot) \sim GP(0, C_{ii|\mathcal{L}})$ such that $z_i(\cdot) \perp z_j(\cdot)$ for all $i \neq j$, and $z_i(\cdot) \perp w(\mathcal{L})$ for all i. Independence among $z_i(\mathcal{L})$ and $w(\mathcal{L})$ and the marginal covariance of $z_i(\mathcal{L})$ in (4.2) ensures that each $w_i(\cdot)$

on $\mathcal{D}$ is exactly $GP(0, C_{ii})$. However, independence among the $z_i(\cdot)$'s is a neat choice ensuring that the conditional independence relations in $\mathcal{G}_\mathcal{V}$ is extended from the finite set $\mathcal{L}$ to the spatial process over $\mathcal{D}$. We prove this formally in Theorem 4.2.4.

Theorem 4.2.4. *Given a cross-covariance function C and an inter-variable graph $\mathcal{G}_\mathcal{V}$, stitching creates a valid multivariate GGP $w(\cdot)$ with a valid (p.d.) cross-covariance function M such that:*

(a) $w_i(\cdot) \sim GP(0, C_{ii})$ *, i.e.,* $M_{ii}(s, s') = C_{ii}(s, s')$ *for all* $s, s' \in \mathcal{D}$ *and for each* $i = 1, \ldots, q$*,*

(b) $w(\cdot)$ *is a* $GGP(\mathcal{G}_\mathcal{V})$ *on* $\mathcal{D}$*,*

(c) if $(i, j) \in E_\mathcal{V}$*, then* $M_{ij}(s, s') = C_{ij}(s, s')$ *for all* $s, s' \in \mathcal{L}$*.*

Stitching produces a multivariate GP $w(\cdot)$ that exactly satisfies the first two conditions sought in Section 4.2.1. Regarding Condition (iii), we point out some differences between the GGP ensured by Theorem 4.2.1 and the one produced by stitching. For pairs of variables $(i, j) \in E_\mathcal{V}$, the cross-covariance for the former is exactly the same as the given cross-covariance C_{ij} on the entire domain $\mathcal{D}$, whereas for the latter $M_{ij}(s, s') = C_{ij}(s, s')$ for locations in $\mathcal{L}$. For a pair $s, s' \notin \mathcal{L}$ and $i \neq j$ it is straightforward to verify that

$$M_{ij}(s, s') = C_{ii}(s, \mathcal{L})C_{ii}(\mathcal{L}, \mathcal{L})^{-1}M(\mathcal{L}, \mathcal{L})_{ij}C_{jj}(\mathcal{L}, \mathcal{L})^{-1}C_{jj}(\mathcal{L}, s') \,. \tag{4.3}$$

Stitching, thus, produces a computationally feasible GGP with desired full-rank marginal covariance and process-level conditional independence at the

expense of allowing a fixed rank cross-covariance. Choosing $\mathcal{L}$ to be reasonably dense (well-spaced) in $\mathcal{D}$, we have $M_{ij}(s, s') \approx C_{ij}(s, s')$ for $(i, j) \in E_{\mathcal{V}}$, $s, s' \in \mathcal{D} \backslash \mathcal{L}$. Hence, condition (iii) is satisfied exactly on $\mathcal{L}$ and approximately on $\mathcal{D} \backslash \mathcal{L}$ for the stitched GP.

4.3 Highly multivariate Graphical Matérn Gaussian processes

4.3.1 Incompatibility of multivariate Matérn with graphical models

Theorems 4.2.1 and 4.2.4 establish, respectively, the existence of and the construction of a marginal-preserving GGP given any valid cross-covariance C and any inter-variable graph $\mathcal{G}_{\mathcal{V}}$. We are particularly interested in developing a novel class of *multivariate graphical Matérn GPs* that are GGP($\mathcal{G}_{\mathcal{V}}$) such that each univariate process is a Matérn GP. This is appealing for inference as we retain the ability to interpret the parameters for each univariate spatial process. We achieve this using stitching, which is necessary as we argue below that no non-trivial parametrisation of the existing multivariate Matérn GP yields a GGP.

The isotropic multivariate Matérn cross-covariance function on a d-dimensional domain is $C_{ij}(s, s') = \sigma_{ij} H_{ij}(\|s - s'\|)$, where $H_{ij}(\cdot) = H(\cdot \mid \nu_{ij}, \phi_{ij})$, H being the Matérn correlation function (Apanasovich, Genton, and Sun, 2012). If $\theta_{ij} = \{\sigma_{ij}, \nu_{ij}, \phi_{ij}\}$, then for a multivariate Matérn GP the ith individual variable is a Matérn GP with parameters θ_{ii}. This is attractive because it endows each univariate process with its own variance σ_{ii}, smoothness

ν_{ii}, and spatial decay ϕ_{ii}. Another nice property is that under this model, $\Sigma = (\sigma_{ij}) = \text{Cov}(w(s))$ is the covariance matrix for $w(s)$ within each location s. The cross-correlation parameters ν_{ij} and ϕ_{ij} for $i \neq j$, are generally hard to interpret, especially since ν_{ij} does not correspond to the smoothness of any surface. Recent work by Kleiber, 2017 on the concept of *coherence* has facilitated some interpretation of these parameters. The *parsimonious multivariate Matérn* model of Gneiting, Kleiber, and Schlather, 2010 emerges from this general specification as a special case with $\nu_{ij} = (\nu_{ii} + \nu_{jj})/2$ and $\phi_{ij} = \phi$.

To ensure a valid multivariate Matérn cross-covariance function, it is sufficient to constrain the intra-site covariance matrix $\Sigma = (\sigma_{ij})$ to be of the form (Theorem 1, Apanasovich, Genton, and Sun, 2012)

$$\sigma_{ij} = \; b_{ij} \frac{\Gamma(\frac{1}{2}(\nu_{ii}+\nu_{jj}+d))\Gamma(\nu_{ij})}{\phi_{ij}^{2\Delta_A+\nu_{ii}+\nu_{jj}}\Gamma(\nu_{ij}+\frac{d}{2})} \text{ where } \Delta_A \geqslant 0, \text{ and } B = (b_{ij}) > 0, \text{ i.e., is p.d.} \tag{4.4}$$

This is equivalent to Σ being constrained as $\Sigma = (B \odot (\gamma_{ij}))$, where γ_{ij} are constants collecting the terms in (4.4) involving only ν_{ij}'s and ϕ_{ij}'s, and $\odot$ denotes the Hadamard (element-wise) product. Similarly, the spectral density matrix takes the form $F(\omega) = (B \odot (g_{ij}(\omega)))$, where $g_{ij}(\omega)$ are functions involving the parameters ϕ_{ij} and ν_{ij}. The matrix $B = (b_{ij})$'s are the $O(q^2)$ parameters (free of ϕ_{ij}'s or ν_{ij}'s) that are constrained to ensure B is positive-definite. Process-level conditional independences introduce zeros in the inverse of the spectral density matrix for stationary processes (see, e.g., Theorem 2.4 in Dahlhaus, 2000). This implies that, for any parametrisation of the multivariate Matérn GP to be a GGP, we need $(F(\omega)^{-1})_{ij} = 0$ for every $(i, j) \notin E_{\mathcal{V}}$ and almost all ω. From the Hadamard product $F(\omega) = (B \odot (g_{ij}(\omega)))$, it is clear that zeros

in B^{-1} or Σ^{-1} do not generally imply zeros in $F^{-1}(\omega)$ for the multivariate Matérn. An exception occurs when each component is posited to have the same smoothness ν and the same spatial decay parameter ϕ, whence both Σ and $F(\omega)$ become proportional to B. In this case, zeros in B^{-1} (specified according to $\mathcal{G}_{\mathcal{V}}$) will correspond to zeros in Σ^{-1} and $F^{-1}(\omega)$ yielding a GGP with respect to $\mathcal{G}_{\mathcal{V}}$. However, assuming $\nu_{ij} = \nu$ and $\phi_{ij} = \phi$ for all (i, j) implies that the univariate GPs have the same smoothness and rate of spatial decay, which is restrictive. Beyond this separable model, there is, to the best of our knowledge, no known parameter choice for the multivariate Matérn GPs that will allow it to be a GGP($\mathcal{G}_{\mathcal{V}}$).

4.3.2 Computational considerations for stitching

Stitching univariate processes corresponding to a valid multivariate Matérn cross-covariance C and a graph $\mathcal{G}_{\mathcal{V}}$ yields a multivariate graphical Matérn GP such that (i) the univariate processes are exactly Matérn; (ii) the multivariate process conforms to process-level conditional independence relations as specified by $\mathcal{G}_{\mathcal{V}}$; and (iii) the cross-covariances for pairs of variables in $\mathcal{G}_{\mathcal{V}}$ are exactly or approximately Matérn (see Eq. 4.3). For each $i = 1, 2, \ldots, q$ let D_i be the set of n_i locations where the i-th variable has been observed. The joint probability density of $w_i(D_i)$ and $w(\mathcal{L})$ is specified by $w(\mathcal{L}) \sim N(0, M(\mathcal{L}, \mathcal{L}))$ and

$$w_i(D_i) \,|\, w(\mathcal{L}) \overset{ind}{\sim} N(C_{ii}(D_i, \mathcal{L})C_{ii}(\mathcal{L}, \mathcal{L})^{-1}w_i(\mathcal{L}), C_{ii\,|\,\mathcal{L}}(D_i, D_i)) \;\text{ for } i = 1, \ldots, q\,. \tag{4.5}$$

The covariance matrix for $\{w_i(D_i) : i = 1, \ldots, q\} \,|\, w(\mathcal{L})$ is block-diagonal with

variable-specific blocks and is cheap to compute if all of the n_i's are small. If some n_i's are large, we can use one of the several variants of scalable GPs for very large number of locations (Heaton et al., 2019). For example, a nearest neighbour GP (NNGP, Datta et al., 2016) yields a sparse approximation of $C_{ii\,|\,\mathcal{L}}(D_i, D_i)$ with linear complexity, but the joint distribution still preserves the conditional independence implied by $\mathcal{G}_\mathcal{V}$.

When q is large, note that $\{w_i(D_i) : i = 1, \ldots, q\}\,|\,w(\mathcal{L})$ in (4.5) has q conditionally independent factors and is easy to compute in parallel. However, the likelihood for $w(\mathcal{L}) \sim N(0, M(\mathcal{L}, \mathcal{L}))$ presents the bottleneck for this highly multivariate case. In particular, there are two challenges for large q. As discussed earlier, the multivariate Matérn C required for stitching needs to constrain $B = (b_{ij})$ to be p.d. on an $O(q^2)$-dimensional parameter space. Searching in such a high-dimensional space is difficult for large q and verifying positive definiteness of B incurs an additional cost of $O(q^3)$ flops. Second, evaluating $w(\mathcal{L}) \sim N(0, M(\mathcal{L}, \mathcal{L}))$ involves matrix operations for the $nq \times nq$ matrix $M(\mathcal{L}, \mathcal{L})$. While the precision matrix, $M(\mathcal{L}, \mathcal{L})^{-1}$, is sparse because of $\mathcal{G}_V$, its determinant is usually not available in closed form and the calculation can become prohibitive even for small n.

4.3.3 Decomposable variable graphs

To facilitate scalability in highly multivariate settings, we consider decomposable inter-variable graphs. For $\mathcal{G}_\mathcal{V} = (\mathcal{V}, E)$, and a triplet (A, B, O) of disjoint subsets $\mathcal{V}$, O is said to *separate* A from B if every path from A to B passes through O. If $\mathcal{V} = A \cup B \cup O$, and O induces a complete subgraph of $\mathcal{V}$, then

Table 4.1: Properties of any q-dimensional multivariate Matérn GP of Gneiting, Kleiber, and Schlather, 2010 or Apanasovich, Genton, and Sun, 2012 and a multivariate graphical Matérn GP stitched using a decomposable graph $\mathcal{G}_V$ with largest clique size q^* (typically $\ll q$), length of perfect ordering p, and maximal number of cliques p^* sharing a common vertex.

Model attributes	Multivariate Matérn	Multivariate Graphical Matérn
Number of parameters	$O(q^2)$	$O(\lvert E_{\mathcal{V}}\rvert + q)$
Parameter constraints	$O(q^3)$	$O(p^*(q^{*3}))$ (worst case)
Storage	$O(n^2q^2)$	$O(pn^2q^{*2})$ (worst case)
Time complexity	$O(n^3q^3)$	$O(pn^3q^{*3})$ (worst case)
Conditionally independent processes	No	Yes
Univariate components are Matérn GPs	Yes	Yes

(A, B, O) is said to decompose $\mathcal{G}_{\mathcal{V}}$. The graph $\mathcal{G}_{\mathcal{V}}$ is said to be decomposable if it is complete or if there exists a proper decomposition (A, B, O) into decomposable subgraphs $\mathcal{G}_{A\cup O}$ and $\mathcal{G}_{B\cup O}$. Several naturally occurring dependence structures like low-rank dependence or autoregressive dependence correspond to decomposable graphs (see Section 4.4). More generally, if a graph is non-decomposable, it can be embedded in a larger decomposable graph. Hence, assuming decomposability is conspicuous in graphical models (see, e.g., Dobra et al., 2003; Wang and West, 2009) since fitting Bayesian graphical models is cumbersome for non-decomposable graphs (Roverato, 2002; Atay-Kayis and Massam, 2005).

For stitching of Matérn GPs using decomposable graphs we can significantly reduce the dimension of the parameter space, storage and computational burden. Let $K_1, \cdots, K_p$ be a sequence of subsets of the vertex set $\mathcal{V}$ for an undirected graph $\mathcal{G}_{\mathcal{V}}$. Let, $F_m = K_1 \cup \cdots \cup K_m$ and $S_m = F_{m-1} \cap K_m$. The sequence $\{K_m\}$ is said to be *perfect* if (i) for every $l > 1$, there is an $m < l$ such that $S_l \subset K_m$; and (ii) the *separator* sets S_m are complete for all m. If $\mathcal{G}_V$ is

decomposable, then it has a perfect clique sequence (Lauritzen, 1996) and the joint density of $w(\mathcal{L})$ can be factorized as follows.

Corollary 4.3.0.1. *If $\mathcal{G}_{\mathcal{V}}$ has a perfect clique sequence $\{K_1, K_2, \cdots, K_p\}$ with separators $\{S_2, \ldots, S_m\}$, then the GGP likelihood on $\mathcal{L}$ can be decomposed as*

$$f_M(w(\mathcal{L})) = \frac{\Pi_{m=1}^{p} f_C(w_{K_m}(\mathcal{L}))}{\Pi_{m=2}^{p} f_C(w_{S_m}(\mathcal{L}))}\,, \tag{4.6}$$

where f_A denotes the density of a GP over $\mathcal{L}$ with covariance function A for $A \in \{M, C\}$.

Corollary 4.3.0.1 helps us manage the dimension and constraints of the parameter space and the computational complexity of stitching. For an arbitrary $\mathcal{G}_V$, the parameter space for the stitching covariance function M is the same as the parameter space $\{\theta_{ij} | 1 < i, j \leqslant q\}$ for the original covariance function C. For a decomposable $\mathcal{G}_V$, the likelihood (4.6) and, in turn, the stitched GGP is only specified by the parameters $\{\theta_{ij} \,|\, (i = j) \text{ or } (i, j) \in E_{\mathcal{V}}\}$. Therefore, the dimension of the parameter space reduces from $O(q^2)$ to $O(|E_{\mathcal{V}}| + q)$, where $|E_{\mathcal{V}}|$ is the number of edges on $\mathcal{G}_{\mathcal{V}}$, which is small for sparse graphs. When using a multivariate Matérn cross-covariance C for stitching, the parameter space for B in the stitched graphical Matérn is the intersection of the parameter spaces of the low-dimensional clique-specific multivariate Matérn covariance functions $C_{K_1}, \ldots, C_{K_p}$. Hence, the parameter space becomes $\{b_{ij} | (i = j) \text{ or } (i, j) \in E_{\mathcal{V}}\}$ and needs to satisfy the constraint that $B_{K_l} = (b_{ij})_{i,j \in K_l}$ is p.d. for all $l = 1, \ldots, p$. This reduces the computational complexity of parameter constraints from $O(q^3)$ to at most $O(p^* q^{*3})$, where q^* is the largest clique size and p^* is the maximum number of cliques sharing a common vertex. The precision matrix

of $w(\mathcal{L})$ satisfies (Lemma 5.5, Lauritzen, 1996)

$$M(\mathcal{L},\mathcal{L})^{-1} = \sum_{m=1}^{p} [C^{-1}_{[K_m \boxtimes \mathcal{G}_\mathcal{L}]}]^{\mathcal{V}\times\mathcal{L}} - \sum_{m=2}^{p} [C^{-1}_{[S_m \boxtimes \mathcal{G}_\mathcal{L}]}]^{\mathcal{V}\times\mathcal{L}}, \tag{4.7}$$

where, for any symmetric matrix $A = (a_{ij})$ with rows and columns indexed by $\mathcal{U} \subset \mathcal{V} \times \mathcal{L}$, $A^{\mathcal{V}\times\mathcal{L}}$ denotes a $|\mathcal{V}\times\mathcal{L}| \times |\mathcal{V}\times\mathcal{L}|$ matrix such that $(A^{\mathcal{V}\times\mathcal{L}})_{ij} = a_{ij}$ if $(i,j) \in \mathcal{U}$, and $(A^{\mathcal{V}\times\mathcal{L}})_{ij} = 0$ elsewhere. From (4.6) and (4.7) we see that the stitching likelihood evaluation avoids the large matrix $M(\mathcal{L},\mathcal{L})$ and all matrix operations are limited to the sub-matrices of $M(\mathcal{L},\mathcal{L})$ corresponding to the cliques $K_m \boxtimes \mathcal{G}_\mathcal{L}$ and separators $S_m \boxtimes \mathcal{G}_\mathcal{L}$. The entire process requires at most $O(pn^3q^{*3})$ flops and $O(pn^2q^{*2})$ storage, where p is the length of the perfect ordering. Table 4.1 summarizes these gains from stitching with decomposable graphs.

The computational efficiency of stitching is clear from the above. In addition, the following result shows that the GGP likelihood from stitching yields unbiased estimating equations for all parameters included in the GGP (all marginal and cross-covariance parameters for any pairs of variables included in $\mathcal{G}_\mathcal{V}$) under model misspecification when the data is generated from a multivariate Matérn GP, but is modelled as a graphical Matérn GP with a decomposable $\mathcal{G}_\mathcal{V}$.

Proposition 4.3.1. Let $w(\cdot) \sim GP(0, C(\cdot,\cdot))$, where C is a valid $q \times q$ multivariate Matérn cross-covariance function with parameters $\{\theta_{ij} : 1 \leqslant i,j \leqslant q\}$, and $f_M(w(\mathcal{L}))$ denotes the multivariate graphical Matérn GP likelihood (4.6) from stitching using a decomposable graph $\mathcal{G}_\mathcal{V}$. Then $E(\partial \log f_M(w(\mathcal{L}))/\partial\theta_{ij}) = 0$ for any $i = j$ or $(i,j) \in E_\mathcal{V}$.

4.3.4 Chromatic Gibbs sampler

With a valid process specification for $w(\cdot)$, we cast (4.1) into a hierarchical model over the n observed locations in $\mathcal{S}$ and sample from the posterior distribution derived from

$$p(\beta, \tau, \theta) \times N(w(\mathcal{S}) \,|\, 0, C_\theta(\mathcal{S}, \mathcal{S})) \times \prod_{j=1}^{n} N(y(s_j) \,|\, X(s_j)\beta + w(s_j), D_\tau) \,, \quad (4.8)$$

where $X(s_j) = \text{diag}(x_1(s_j)^\text{T}, x_2(s_j)^\text{T}, \ldots, x_q(s_j)^\text{T})$ is $q \times (\sum_{i=1}^{q} p_i)$, $\beta = (\beta_1^\text{T}, \beta_2^\text{T} \ldots, \beta_q^\text{T})^\text{T}$, $w(\mathcal{S}) = (w(s_1)^\text{T}, w(s_2)^\text{T} \ldots, w(s_n)^\text{T})^\text{T}$, $D_\tau = \text{diag}(\tau_1^2, \tau_2^2, \ldots, \tau_q^2)$, θ is the set of parameters in the cross-covariance function and $p(\beta, \tau, \theta)$ is a prior distribution on model parameters. Besides the computational benefits described in Table 4.1, stitched GGP models are also amenable to parallel computing. In a Bayesian implementation of a stitched GGP model (described in Section S2.1 of the Supplement), we can exploit the graph $\mathcal{G}_V$ and deploy a chromatic Gibbs sampler (Gonzalez et al., 2011) to simultaneously update batches of random variables in parallel. Let η_i be the vector grouping variable-specific parameters (regression coefficients, spatial parameters, noise variance and latent spatial random effects). Under a graph colouring of $\mathcal{G}_V$, η_i and $\eta_{i'}$ can be updated simultaneously if i and i' share the same colour, as illustrated in Figure 4.2 (left).

This brings down the number of sequential steps in sampling of the η_i's from q to the chromatic number $\chi(\mathcal{G}_V)$. We can also employ a chromatic sampling scheme for the b_{ij}'s, but using a different graph. We exploit the fact that the parameters b_{ij} and $b_{i'j'}$ belongs to the same factor in (4.6) for a pair of edges (i, j) and (i', j') in $E_\mathcal{V}$ if and only if the variables i, j, i', j' belongs to

Figure 4.2: Chromatic sampling for GGP with a gem graph between 5 variables: Left: Gem graph and colouring used for chromatic sampling of the variable-specific parameters. Right: Colouring of the corresponding edge graph $\mathcal{G}_E(\mathcal{G}_V)$ used for chromatic sampling of the cross-covariance parameters b_{ij}'s.

the same clique. Thus, if $\mathcal{G}_E(\mathcal{G}_V) = (E_\mathcal{V}, E^*)$ denotes this graph on the set of edges $E_\mathcal{V}$, i.e., there is an edge $((i,j),(i',j'))$ in this new graph $\mathcal{G}_E(\mathcal{G}_V)$ if $\{i,i',j,j'\}$ are in some clique K of $\mathcal{G}_V$, then we can batch the updates of b_{ij}'s based on the colouring of the graph $\mathcal{G}_E(\mathcal{G}_V)$ (Figure 4.2 (right)). The number of such sequential batch updates will be the chromatic number $\chi(\mathcal{G}_E(\mathcal{G}_V))$, a potentially drastic reduction from $|E_\mathcal{V}|$ sequential updates for b_{ij}.

4.4 Extensions

4.4.1 Factor models

The construction of GGP and its implementation described in Sections 4.2 and 4.3 assumes a known graphical model. Here, we describe different avenues for choosing or estimating the graph and offer extensions of GGP to model different spatial and spatiotemporal structures.

In many multivariate spatial models, the inter-variable graphical model arises naturally and is decomposable. A large subset of multivariate spatial models are process-level factor models (emerge from more general linear models of coregionalization (LMC)), where each of the q observed univariate processes are a weighted sum of $r \leqslant q$ latent univariate factor processes with

the weights being component-specific (Schmidt and Gelfand, 2003; Gelfand et al., 2004; Wackernagel, 2013). In general, a linear model of coregionalization can be expressed as

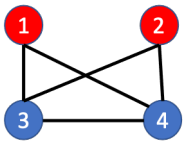
$$w_i(s) = \sum_{j=1}^{r} a_{ij}(s) f_j(s) + \xi_i(s) \,, \tag{4.9}$$

where each $f_j(\cdot)$ is a latent factor process such that $f(\cdot) = (f_1(\cdot), \ldots, f_r(\cdot))^{\mathrm{T}}$ is a multivariate GP, $a_{ij}(\cdot)$'s are component-specific weight functions and $\xi(\cdot)$ are independent processes representing the idiosyncratic spatial variation in $w_i(\cdot)$ not explained by the latent factors. If q is large, choosing $r \ll q$ in (4.9) also facilitates dimension reduction (Lopes, Salazar, and Gamerman, 2008; Ren and Banerjee, 2013; Taylor-Rodriguez et al., 2019; Zhang and Banerjee, 2021). We next show that any linear model of coregionalization can be formulated as a GGP with a decomposable graph on the elements of $w(\cdot)$ and $f(\cdot)$.

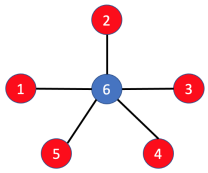
Proposition 4.4.1. Consider the linear model of coregionalization (4.9) where $f(\cdot)$ is an $r \times 1$ multivariate GP with a complete graph between component processes, and $\xi_i(\cdot)$'s are independent univariate GPs. Then $(w_1, w_2, \ldots, w_q, f_1, \ldots, f_r)^{\mathrm{T}}$ is a GGP on vertices $\{1, \ldots, q+r\}$ and a decomposable graph $\{(i,j) | i \in 1, \ldots, (q+r), j \in (q+1), \ldots, (q+r), j \neq i\}$.

Proposition 4.4.1 dictates that the assumption of multivariate dependence induced through factor processes can be translated into a decomposable graph between the observed and factor processes. Hence, GGPs can be used as a richer alternative to the linear model of coregionalization. While the linear model of coregionalization enforces all processes $w_i(\cdot)$ to have the smoothness of the roughest $f_j(\cdot)$ (Genton and Kleiber, 2015), the GGP enables us to model

and interpret the spatial smoothness of each component process (e.g., with the graphical Matérn GP). The complete graph between the r component processes of $f(\cdot)$ can be assumed without loss of generality as even for a sparse graph between latent factors (e.g., when the factors are independent processes), they will generally be conditionally dependent given the observed processes $w(\cdot)$, thereby yielding the same joint graph. Due to $r \ll q$, this joint graph of observed and latent processes will still be sparse even after considering all possible edges between latent processes. Figure 4.3 illustrates two examples of the decomposable graphs arising from linear model of coregionalization.



(a) 2 observed (red) and 2 latent (blue) processes



(b) 5 observed (red) and 1 latent (blue) processes

Figure 4.3: Decomposable graphs for (a) a full rank and (b) a low-rank linear model of coregionalization.

An alternative approach to linear models of coregionalization builds multivariate spatial processes by sequentially modelling a set of univariate GPs conditional from some ordering of the q variables (Cressie and Zammit-Mangion, 2016). A sparse partial ordering can facilitate dimension reduction for large q. This approach does not attempt to preserve marginals or introduce process-level conditional independence. However, a partial ordering yields a directed acyclic graph (DAG), which, when moralised, produces a decomposable undirected graph that can be used in our stitched GGPs.

4.4.2 Non-separable spatial time-series modelling

GGPs are natural candidates for non-separable (in space-time), non-stationary (in time) modelling of univariate or multivariate spatial time-series. Consider a univariate spatial time-series modelled as a GP $\{w(s,t)\}$ for $s \in \mathcal{D}$ evolving over a discrete set of time points $t \in \mathcal{T} = \{1, 2, \ldots, T\}$. We envision this as a $T \times 1$ GP $w(s) = (w_1(s), \ldots, w_T(s))^{\mathrm{T}}$, where $w_t(s) = w(s,t)$. Temporal evolution of processes is often encapsulated using a directed acyclic graph (DAG), which, when moralized, produces an undirected graph $\mathcal{G}_{\mathcal{T}}$ over $\mathcal{T}$. We can then recast the spatial time-series model as a $T \times 1$ GGP with respect to $\mathcal{G}_{\mathcal{T}}$. A multivariate Matérn used for stitching will produce a GGP with each $w_t(\cdot)$ being a Matérn GP with parameters θ_{tt}. Time-specific process variances and spatial parameters enrich the model without imposing stationarity of the spatial process over time and space-time separability (Gneiting, 2002).

Any autoregressive (AR) structure over time corresponds to a decomposable moralized graph $\mathcal{G}_{\mathcal{T}}$. For example, the $AR(1)$ model corresponds to a path graph with edges $\{(t, t+1) \,|\, t = 1, \ldots, T-1\}$, $q^* = 2$ and $p^* = 2$. An $AR(2)$ is specified by the DAG $t-2 \to t$ and $t-1 \to t$ for all $t \in \{3, \ldots, T\}$ (Figure S8a in the Supplement), which, when moralized, yields the sparse decomposable graph $\mathcal{G}_{\mathcal{T}}$ (with $q^* = 3$) in Figure S8b of the Supplement. Hence, Corollary 4.3.0.1 accrues computational gains for GGP models for autoregressive spatial time-series. An added benefit of using the GGP is that the auto-regression parameters need not be universal, but can be time-specific, thus relaxing another restrictive stationarity condition.

GGP allows the marginal variances and autocorrelations of the processes

to vary over time and be estimated in an unstructured manner. However, more structured temporal models for stochastic volatility can be easily accommodated by a GGP if forecasting the process at a future time-point is of interest. This can be achieved by adding a model for the time-specific variances like the log-AR(1) model as considered in Jacquier, Polson, and Rossi, 1993. Bayesian estimation of these model parameters has been discussed in Jacquier, Polson, and Rossi, 2002 and can be seamlessly incorporated into our Bayesian framework for estimation of GGP parameters.

Multivariate spatial time-series can also be modelled using GGP. We envision q variables recorded at T time-points resulting in qT variables. We now specify $\mathcal{G}_{\mathcal{V}\times\mathcal{T}}$ on the variable-time set. Common specifications for multivariate time-series like graphical vector autoregressive (VAR) structures (Dahlhaus and Eichler, 2003) will yield decomposable $\mathcal{G}_{\mathcal{V}\times\mathcal{T}}$. For example, consider the non-separable graphical-VAR of order 1 with $q = 2$ and specified by the DAG $(1, t-1) \to (1, t)$, $(1, t-1) \to (2, t)$, and $(2, t-1) \to (2, t)$ (Figure S8c of the Supplement). This yields the decomposable $\mathcal{G}_{\mathcal{V}\times\mathcal{T}}$ in Figure S8d of the Supplement, also with $q^* = 3$.

4.4.3 Graph estimation

Sections 4.4.1 and 4.4.2 present settings where the decomposable graph for a GGP arises naturally. For gridded spatial data, one can use a spatial graphical lasso to estimate the graph from the sparse inverse spectral density matrix (Jung, Hannak, and Goertz, 2015), and plug-in the estimated graph in subsequent estimation of GGP likelihood parameters. For irregularly located spatial

data, we now extend our framework in (4.8) to infer about the graphical model itself along with the GGP parameters by adapting an MCMC sampler for decomposable graphs (Green and Thomas, 2013).

The *junction graph* G of a decomposable $\mathcal{G}_{\mathcal{V}}$ is a complete graph with the cliques of $\mathcal{G}_{\mathcal{V}}$ as its nodes. Every edge in the junction graph is represented as a link, which is the intersection of the two cliques, and can be empty. A *spanning tree* of a graph is a subgraph comprising all the vertices of the original graph and is a tree (acyclic graph). Suppose a spanning tree J of the junction graph of G satisfies the following property: for any two cliques C and D of the graph, every node in the unique path between C and D in the tree contains $C \cap D$. Then J is called the *junction tree* for the graph $\mathcal{G}_{\mathcal{V}}$ (see Figure 2 of Thomas and Green, 2009, for an illustration). A junction tree exists for $\mathcal{G}_{\mathcal{V}}$ if and only if $\mathcal{G}_{\mathcal{V}}$ is decomposable. Also, a decomposable graph can have many junction trees but each junction tree represents a unique decomposable graph. This allows us to transform a prior on decomposable graphs to a prior on the junction trees. If $\mu(\mathcal{G}_{\mathcal{V}}(J))$ is the number of junction trees for the decomposable graph $\mathcal{G}_{\mathcal{V}}$ corresponding to J, then a prior π on decomposable graphs gives rise to a prior $\tilde{\pi}$ on the junction trees as $\tilde{\pi}(J) = \pi(\mathcal{G}_{\mathcal{V}}(J))/\mu(\mathcal{G}_{\mathcal{V}}(J))$. In our application, we assume π to be uniform over all decomposable graphs with a pre-specified maximum clique size, i.e., $\tilde{\pi}(J) \propto 1/\mu(\mathcal{G}_{\mathcal{V}}(J))$.

With junction trees as a representative state variable for the graph, the jumps are governed by constrained addition or deletion of single/multiple edges so that the resulting tree is also a junction tree for some decomposable

graph. Each graph corresponds to a different GGP model using a specific subset of the cross-covariance parameters. To embed sampling this graph within the Gibbs sampler in Section S2.1, jumps between graphs need to be coupled with introduction or deletion of cross-covariance parameters depending on addition or deletion of edges. We use the reversible jump MCMC (rjMCMC) algorithm of Barker and Link, 2013 to carry out the sampling of the graph and cross-covariance parameters and lay out the details in Section S2.3.

4.4.4 Asymmetric covariance functions

Our examples of stitching have primarily involved the isotropic (symmetric) multivariate Matérn cross-covariances. Symmetry implies $C_{ij}(s, s') = C_{ij}(s', s)$ for all i, j, s, s' and is not a necessary condition for validity of a cross-covariance function. An asymmetric cross-covariance function (Apanasovich and Genton, 2010; Li and Zhang, 2011) C^a can be specified in-terms of a symmetric cross-covariance C as $C^a_{ij}(s, s') = C^a_{ij}(s - s') = C_{ij}(s - s' + (a_i - a_j))$, where a_i, $i = 1, \ldots, q$ are distinct variable specific parameters. Stitching works with any valid cross-covariance function, and if C^a is used for stitching, then the resulting graphical cross-covariance M^a will also be asymmetric, satisfying $M^a_{ij}(s, s') = C^a_{ij}(s, s')$ for all $(i, j) \in E_{\mathcal{V}}$, and $s, s' \in \mathcal{L}$.

4.4.5 Response model

We outline a Gibbs sampler in Section S2.1 of the Supplement for the multivariate spatial linear model in (4.1), where the latent $q \times 1$ process $w(s)$ is modelled as a GGP. If $|\mathcal{L}| = n$, then the algorithm needs to sample $\sim O(nq)$

latent spatial random effects $w(\mathcal{L})$ at each iteration.

A popular method for estimating spatial process parameters in (4.1) is to integrate out the spatial random effects $w(\mathcal{L})$ and directly use the marginalized (or collapsed) likelihood for the response process $y(\cdot) = (y_1(\cdot), \ldots, y_q(\cdot))^{\mathrm{T}}$, which is also a multivariate GP. However, $w(\cdot)$ modelled as a GGP does not ensure that the marginalized $y(\cdot)$ will be a GGP. We demonstrate this in Figure 4.4(a) with a path graph $\mathcal{G}_{\mathcal{V}}$ between 3 latent processes $w_1(\cdot)$, $w_2(\cdot)$ and $w_3(\cdot)$. The response processes $y_i(\cdot) = w_i(\cdot) + \epsilon_i(\cdot)$ have complete graphs. This is because $\mathrm{Cov}(y) = \mathrm{Cov}(w) + \mathrm{Cov}(\epsilon)$, and the zeros in $\mathrm{Cov}(w)^{-1}$ do not correspond to zeros in $\mathrm{Cov}(y)^{-1}$. Hence, modelling the latent spatial process as a GGP and subsequent marginalization is inconvenient because the marginalized likelihood for y will not factorize like (4.6).

(a) GGP on the latent process.

(b) GGP on the response process.

Figure 4.4: Comparison of induced graphs for 3 processes (obeying a path graph) from marginalized model and latent model. Blue edges indicate the dependencies modelled and red edges denote the marginal dependencies induced from the model construction.

Instead, we can directly create a GGP for the response process by stitching the marginal cross-covariance function $\mathrm{Cov}(y(s), y(s+h)) = C(h) + D(h)$ using $\mathcal{G}_{\mathcal{V}}$, where $D(h) = \mathrm{diag}(\tau_1^2, \ldots, \tau_q^2) I(h = 0)$ is the diagonal white-noise covariance function. With a Matérn cross-covariance C, the resulting GGP

model for $y(\cdot)$ endows each univariate GP $y_i(\cdot)$ with mean $x_i(\cdot)^{\mathrm{T}}\beta_i$ and retaining the marginal covariance function $C_{ii}(h) + \tau_i^2 I(h = 0)$ (i.e., Matérn plus a nugget). The cross-covariance between $y_i(\cdot)$ and $y_j(\cdot)$ is also Matérn for $(i, j) \in E_{\mathcal{V}}$ and locations in $\mathcal{L}$. For $(i, j) \notin \mathcal{G}_V$, the response processes $y_i(\cdot)$ and $y_j(\cdot)$ will be conditionally independent. We outline the Gibbs sampler for this *response GGP* in Section S2.2 of the Supplement.

The response model drastically reduces the dimensionality of the sampler from $O(nq + |E_{\mathcal{V}}|)$ for the latent model to $O(q + |E_{\mathcal{V}}|)$. What we gain in terms of convergence of the chain is traded off in interpretation of the latent process. As we see in Figure 4.4(b), using a graphical model on the response process leads to a complete graph among the latent process. If, however, conditional independence on the latent processes is not absolutely necessary, then the marginalized GGP model is a pragmatic alternative for modelling highly multivariate spatial data.

4.5 Simulations

4.5.1 Known graph

We conducted multiple simulation experiments to compare three models: (a) PM: Parsimonious Multivariate Matérn of Gneiting, Kleiber, and Schlather, 2010; (b) MM: Multivariate Matérn of Apanasovich, Genton, and Sun, 2012 with $\nu_{ij} = \nu_{ii} = \nu_{jj} = \frac{1}{2}$, and $\Delta_A = 0$ and $\phi_{ij}^2 = (\phi_{ii}^2 + \phi_{jj}^2)/2$; and (c) GM: Graphical Matérn (GGP on the latent process, stitched using multivariate Matérn model (b)).

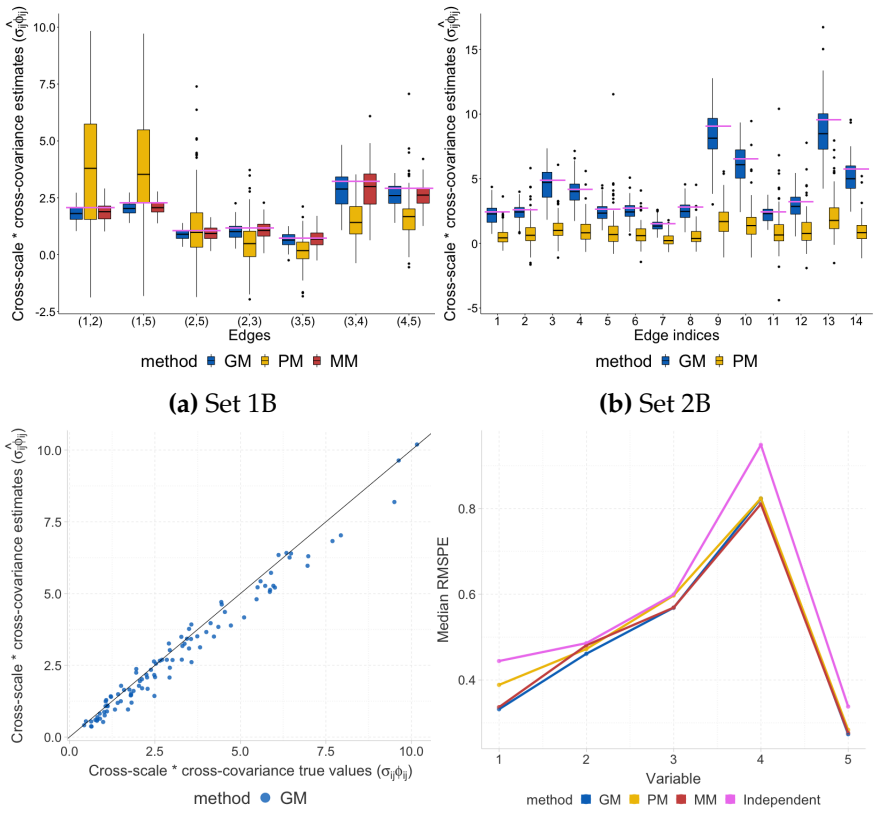
We consider the 6 settings in Table 4.2. In Sets 1A, 2A, and 3A, we generate

Table 4.2: Different simulation scenarios considered for the comparison between methods.

Set	q	Graph $\mathcal{G}_\mathcal{V}$	B	Nugget	Locations	Data model	Fitted models
1A	5	Gem (Figure 4.2(a))	Random	No	Same location for all variables	GM	GM, MM, PM
1B	5	Gem (Figure 4.2(a))	Random	No	Same location for all variables	MM	GM, MM, PM
2A	15	Path	$b_{i-1,i} = \rho_i$	Yes	Partial overlap in locations for variables	GM	GM, PM
2B	15	Path	$b_{i-1,i} = \rho_i$	Yes	Partial overlap in locations for variables	MM	GM, PM
3A	100	Path	$b_{i-1,i} = \rho_i$	Yes	Partial overlap in locations for variables	GM	GM
3B	100	Path	$b_{i-1,i} = \rho_i$	Yes	Partial overlap in locations for variables	MM	GM

data from GM. Set 1A has $q = 5$ and uses a gem graph (Figure 4.2 (a)). For Set 2A, we considered $q = 15$ outcomes and used a path graph, while Set 3A considers the highly multivariate case with $q = 100$ outcomes and a path graph. Sets 1B–3B are same as Sets 1A–3A, respectively, except that we generate data from MM. Thus the scenarios 1A–3A correspond to correctly specified settings for the GGP, while scenarios 1B–3B serve as misspecified examples where data is generated from MM. For all scenarios, we generated data on $n = 250$ locations uniformly chosen over a grid. We simulated 1 covariate $x_j(s_i)$ for each variable j, generated independently from a $N(0, 4)$ distribution and the true regression coefficients β_j from *Unif(-2,2)* for $j = 1, 2, \ldots, q$. The ϕ_{ii} and σ_{ii} were equispaced numbers in $(1, 5)$, while the b_{ij}'s where chosen as in Table 4.2. For all of the candidate models, each component of the q-variate process is a Matérn GP. Following the recommendation outlined in Apanasovich, Genton, and Sun, 2012, the marginal parameters θ_{ii} for the univariate Matérn processes were estimated apriori using only the data for the i-th variable. The `BRISC` R-package (Saha and Datta, 2018) was used for estimation.

To compare estimation performance, we primarily focus on the cross-covariance parameters b_{ij}, $(i, j) \in E_\mathcal{V}$, as they specify the cross-covariances in stitching. Specifically, we compare the estimates of $\sigma_{ij}\phi_{ij} = \Gamma(1/2)b_{ij}$, which are the b_{ij}'s rescaled to be at the same scale as the marginal microergodic



(a) Set 1B **(b)** Set 2B

(c) Set 3B **(d)** Set 1B

Figure 4.5: Performance of graphical Matérn under misspecification: (a), (b) and (c): Estimates of the cross-covariance parameters $\sigma_{ij}\phi_{ij} = \Gamma(1/2)b_{ij}$, $(i,j) \in E_{\mathcal{V}}$ for the sets 1B, 2B and 3B respectively. The pink lines in Figures (a) and (b) indicate true parameter values. (d): Median RMSPE for GM, MM, PM and Independent GP model for Set 1B.

parameters $\sigma_{ii}\phi_{ii}$. Model evaluations under the correctly specified settings of 1A–3A are provided in Supplementary Figure S9, which reveals that the GGP accurately estimates cross-covariance parameters for all the edges in the graph for all 3 scenarios. Figures 4.5 (a), (b), and (c), evaluate the estimates of GM for the misspecified settings 1B, 2B and 3B, respectively. For Set 1B we see that MM, and GM produce reasonable estimates of the true cross-covariance parameters included, whereas the estimates from PM are biased and more

variable. For Set 2B the estimates of PM are once again biased, while GM is more accurate.

For the highly multivariate settings in Sets 3A and 3B, neither PM nor MM can be implemented because B involves $4,950$ parameters and likelihood evaluation requires inverting a $25,000 \times 25,000$ matrix in each iteration. Hence, we only compare the estimates from GGP to the truth. Figure S9c shows that the GGP performs well in the highly multivariate setting with misspecification (3B) with GM once again accurately estimating all the b_{ij}'s for $(i,j) \in E_{\mathcal{V}}$. These simulations under misspecification confirm the accuracy of GGP in estimating b_{ij} for the MM for pairs (i,j) included in the graph and aligns with the conclusion from Proposition 4.3.1.

We also evaluate the impact of misspecification on the predictive performance. Figure 4.5d plots the root mean square predictive error (RMSPE) based on hold-out data for Set 1B. In addition to the models listed in Table 4.2 we also consider a model where each component GP is an independent Matérn GP serving as a reference for the impact of not modelling dependence. We find GM performs competitively with MM (the correctly specified model) yielding nearly identical RMSPEs for all the 5 variables. PM yields higher RMSPE for variables 1 and 3, while the independent model is, unsurprisingly, the least accurate. Additional analyses and discussions are in the Supplementary materials (Section S3). These include comparison of marginal parameter estimates (Section S3.1), impact of excluding edges on estimation of cross-correlation functions (Section S3.2), comparison of GGP with dynamic linear models for spatial time series (Section S3.4), comparison of GGP with linear model of

coregionalization (Section S3.3), and comparison among different variants of the GM model (Section S3.5).

4.5.2 Unknown graph

We also evaluated our model when the graph is unknown and is sampled using the reversible jump MCMC sampler described in Section 4.4.3. We consider simulation scenarios in Sets 1A and 2A from Table 4.2, where the true multivariate process is a graphical Matérn. We assess the accuracy of inferring about the graphical model and the estimates of the cross-covariance parameters. We visualise the estimated edge probabilities for Set 2A in Figure 4.6(a). The blue edges correspond to the true edges, while red ones correspond to false edges. The width of the edges are proportional to the posterior probability of selecting that edge. We see that most of the false edges have narrow width indicating their low selection probability. We report the top 20 probable edges estimated by our model in Table S1 of the Supplement and observe that our approach ranks all the 14 true edges higher than any of the false edges in terms of marginal probability. Figure 4.6(b) shows that the cross-covariance parameters corresponding to true edges are also estimated correctly. The results for Set 1A are similar and presented in Figure S10.

4.6 Spatial modelling of $PM_{2.5}$ time-series

We demonstrate an application of GGP for non-stationary (in time) and non-separable (in space-time) modelling of spatial time-series (Section 4.4.2). We model daily levels of $PM_{2.5}$ measured at monitoring stations across 11 states of the north-eastern US and Washington DC for a three month period from

February, 01, 2020, until April, 30th, 2020. The data is publicly available from the website of the United States Environmental Protection Agency (EPA).

We selected $n = 99$ stations with at least two months of measured data for both 2020 and 2019. Meteorological variables such as temperature, barometric pressure, wind-speed and relative humidity are known to affect $PM_{2.5}$ levels. Since all of the pollutant monitoring stations do not measure all these covariates, we collected the data from NCEP North American Regional Reanalysis (NARR) database, and merged it with the available weather data from EPA to impute daily values of these covariates at pollutant monitoring locations using multilevel B-spline smoothing. Also to adjust for baseline $PM_{2.5}$ levels, for each station and day in 2020, we included a 7-day moving average of the $PM_{2.5}$ data for that station centered around the same day of 2019 as a baseline covariate We adjust for weekly periodicity of $PM_{2.5}$ levels by subtracting day-of-the-week specific means from raw $PM_{2.5}$ values. Following Section 4.4.2, we view the spatial time-series at $n = 99$ locations and $T = 89$ days as a highly multivariate (89-dimensional) spatial data set. Neither the parsimonious Matérn nor the multivariate Matérn were implementable as they involve $89^2/2 \approx 4000$ cross-covariance parameters and 9000×9000 matrix computations ($99 \times 89 \approx 9000$) in each iteration.

We used a graphical Matérn GP with an $AR(1)$ graph based upon exploratory analysis that revealed autocorrelation among pollutant processes on consecutive days after adjusting for covariates. The marginal parameters for day t were σ_{tt}, ϕ_{tt} and τ_t^2. The autoregressive cross-covariance between

days $t-1$ and t is $b_{t-1,t}$. Hence, GGP offers the flexibility to model non-separability across space and time, time-varying marginal spatial parameters and autoregressive coefficients.

We first present a subgroup analysis breaking 89 days worth of data into 6 fortnights. Data for each fortnight is only 14 or 15 dimensional and, hence, we are able to analyse each chunk separately using the parsimonious Matérn (PM). Figure 4.7a presents hold-out RMSPE and reveals that GM and PM produce very similar predictive performance when analysing each fortnight of data separately. We analyse the full dataset using the GGP model (GM) as other multivariate Matérn GPs like PM are precluded by the highly multivariate setting. The GGP model involves only 88 cross-covariance parameters. Since the largest clique size in an AR(1) graph is 2, the largest matrix we deal with for the data at 99 stations is only 198×198. We also consider spatiotemporal models that can model non-stationary and non-separable relationships in the data. Gneiting, 2002 developed general classes of non-separable spatiotemporal models. However, these models assume a stationary temporal process. More importantly, likelihood for this model will involve a dense 9000×9000 matrix over the set of all space-time pairs and is generally impracticable for modelling long spatial time-series.

For the full analysis, we compare GGP with a spatial dynamic linear model (Stroud, Müller, and Sansó, 2001; Gelfand, Banerjee, and Gamerman, 2005) that, like GGP, can parsimoniously model the temporal evolution using an $AR(1)$ structure and allows both non-separability and time-specific parameters. We use the `SpDynLm` function currently offered in the `spBayes` package

(see Section S3.4 of the Supplement for details). Predictive performance is similar for both models with respect to both point predictions (Figure 4.7b) and interval predictions (Figure S11). Figure 4.7c plots variance estimates (in the log-scale) over time of the latent processes. The implementation in `spBayes` uses the customary random-walk prior to model for the AR(1) evolution. This enforces these marginal variances to be monotonically increasing resulting in unrealistically large variance estimates for later time-points. The estimates from GGP show substantial variation across time with generally a decreasing trend going from February to April. The estimates and credible intervals for the auto-correlation parameters $r_{t,t-1}$ (normalized $b_{t,t-1}$) from GGP are presented in Figure 4.7d. There is large variation in these estimates across time with many spikes indicating high positive autocorrelation. Quantitatively, 95% Bayesian credible intervals for 40 out of the 88 (45%) $r_{t,t-1}$ estimates from GM exclude 0 providing strong evidence in favour of non-stationary auto-correlation across time. `SpDynLm` does not have an analogous auto-correlation parameter, and, hence, cannot be compared in this regard.

The estimated average residual spatial surface, $y_{\mathcal{P}}(s) = (1/|\mathcal{P}|)\sum_{t\in\mathcal{P}}(y_t(s) - x_t(s)^{\mathrm{T}}\hat{\beta}_t)$, is depicted in Figure 4.7e for two choices of the time-period $\mathcal{P}$–the first two weeks of February, 2020 (left), and the last two weeks of April, 2020 (right). These two periods represent the beginning and end of the time period for our study and also correspond to before and during lock-downs imposed in the north-eastern US due to COVID-19. We observe a slight decrease in the magnitude of the residual process from February (median across locations: 0.181) to April (median across locations:

0.164) (Figure S12) suggesting a decrease in the $PM_{2.5}$ levels during this period even after accounting for the meteorological covariates and the previous year's level as a baseline. The residuals for April also showed much lesser variability compared to that in February, suggesting a decrease in the latent process variance over time. This agrees with the estimates of σ_{tt} from GGP (Figure 4.7c) and contradicts the strongly increasing variance estimates from SpDynLm (see Section S3.4 for a broader discussion).

4.7 Discussion

This high-dimensional problem we address here accounts for large number of variables and is distinctly different from the burgeoning literature on high-dimensional problems referring to the massive number of spatial locations. A future direction will be to simultaneously address the problem of big n and big q by extending stitching to nearest neighbor location graphs with sparse variable graphs. Relaxing the assumption of linear covariate effects $x_i^{\mathrm{T}}\beta_i$ in (4.1) can also be pursued as discussed recently by Saha, Basu, and Datta, 2021. A multivariate analogue of this would benefit from the sparse precision matrices available from stitching (4.7). Finally, the idea of stitching can be transported to the discrete spatial (areal) setting to create multivariate analogs of the interpretable Directed Acyclic Graph Auto-regressive (DAGAR) models (Datta et al., 2019), where stitching would preserve the univariate marginals being exactly DAGAR distributions.

4.7.1 Acknowledgement

A. Datta gratefully acknowledges financial support from the National Science Foundation Division of Mathematical Sciences grant DMS-1915803. S. Banerjee gratefully acknowledges support from NSF grants, DMS-1916349 and DMS-2113778, and from NIH grants NIEHS-R01ES030210 and NIEHS-R01ES027027. This work was completed through a fellowship supported by a Joint Graduate Training Program between the Department of Biostatistics at the Johns Hopkins Bloomberg School of Public Health and the Intramural Research Program of the National Institute of Mental Health. The authors are grateful to the Editor, Associate Editor and anonymous reviewers for their feedback which helped to improve the manuscript.

S1 Proofs

of Theorem 4.2.1. Part (a). For the original GP, $F(\omega) = \{f_{ij}(\omega)\}$ is a valid spectral density matrix (SDM). Therefore, following Cramer's Theorem (Cramér, 1940; Parra and Tobar, 2017), $F(\omega)$ is positive definite (p.d.) or (almost) every frequency ω. Using Lemma 4.2.2 we derive a unique $\widetilde{F}(\omega) = (\widetilde{f}_{ij}(\omega))$, which is also positive definite and satisfies $\widetilde{F}(\omega)_{ij} = F(\omega)_{ij} = f_{ij}(\omega)$ for $i = j$ or $(i, j) \in E_{\mathcal{V}}$, and $\widetilde{F}(\omega)^{-1}_{ij} = 0$ for $(i, j) \notin E_{\mathcal{V}}$. The square-integrability assumption of $f_{ii}(\omega)$ is sufficient to ensure that $\int |\widetilde{f}_{ij}(\omega)| d\omega < \infty$ using the Cauchy-Schwarz inequality. Thus, we have a spectral density matrix $\widetilde{F}(\omega)$, which is positive definite for (almost) all ω, $\widetilde{f}_{ii}(\omega) = f_{ii}(\omega) > 0$ for all i, ω, and $\int |\widetilde{f}_{ij}(\omega)| d\omega < \infty$ for all i, j. By Cramer's theorem, there exists a GP $w(\cdot)$ with spectral density matrix $\widetilde{F}(\omega)$ and some cross-covariance function M. As

by construction $\widetilde{f}_{ij}(\omega) = f_{ij}(\omega)$ for $i = j$ or $(i,j) \in E_{\mathcal{V}}$, we have $M_{ij} = C_{ij}$ for $i = j$ or $(i,j) \in E_{\mathcal{V}}$. Since $\widetilde{F}^{-1}(\omega)_{ij} = 0$ for $(i,j) \notin E_{\mathcal{V}}$ and almost all ω, using the result of Dahlhaus, 2000, $w(s)$ has process-level conditional independence on $\mathcal{D}$ as specified by $\mathcal{G}_{\mathcal{V}}$, completing the proof.

Part (b). Let $K(\omega) \in \mathcal{F}$. Then by definition $K(\omega)$ corresponds to the SDM of a GGP with respect to $\mathcal{G}_{\mathcal{V}}$. By Theorem 2.4 in Dahlhaus, 2000), $(K(\omega)^{-1})_{ij} = 0$ for all $(i,j) \notin E_{\mathcal{V}}$ and almost all ω. Let $S(K)$ denote the collection of ω for which this happens. From the construction of $\tilde{F}$ in part (a), for each $\omega \in S(K)$ we thus have $(K(\omega)^{-1})_{ij} = (\tilde{F}(\omega)^{-1})_{ij} = 0$ for all $(i,j) \notin E_{\mathcal{V}}$ and $\tilde{F}(\omega)_{ij} = F(\omega)_{ij}$ for all $(i,j) \in E_{\mathcal{V}}$. Using property (c) of Dempster, 1972, we have

$$\operatorname{tr}(K(w)^{-1}F(\omega)) + \log\det(K(\omega)) \geqslant \operatorname{tr}(\tilde{F}(w)^{-1}F(\omega)) + \log\det(\tilde{F}(\omega)) \;\forall\; \omega \in S(K).$$

As $S(K)^c$ has measure zero, integrating this over $S(K)$ produces the inequality in part (b).

□

of Lemma 4.2.3. Consider any $(i,j) \notin E_{\mathcal{V}}$, and $s, s' \in \mathcal{D}$. Let $B = \mathcal{V}\backslash\{i,j\}$ and $\mathcal{A}$ denote the σ-algebra $\sigma(\{w_k^*(s) \,|\, s \in \mathcal{D}, k \in B\})$. As $w_k^*(s) = C_{kk}(s,\mathcal{L})C_{kk}(\mathcal{L},\mathcal{L})^{-1}w_k(\mathcal{L})$ is a deterministic function of $w_k(\mathcal{L})$ for all k, s, we have $\mathcal{A} \subseteq \sigma(w_B(\mathcal{L}))$. On the other hand, as $w_k(s) = w_k^*(s)$ for all $s \in \mathcal{L}$ (predictive process agrees with the original process at the knot locations) we have $\sigma(w_B(\mathcal{L})) \subseteq \sigma(\mathcal{A})$. Hence, $\sigma(w_B(\mathcal{L})) = \sigma(\mathcal{A})$. Now we have

$$
\begin{aligned}
Cov(w_i^*(s), w_j^*(s') \mid \mathcal{A}) =& Cov(w_i^*(s), w_j^*(s') \mid \sigma(w_B(\mathcal{L})) \\
=& C_{ii}(s, \mathcal{L}) C_{ii}(\mathcal{L}, \mathcal{L})^{-1} Cov(w_i(\mathcal{L}), w_j(\mathcal{L}) \mid \sigma(w_B(\mathcal{L})) \\
& C_{jj}(\mathcal{L}, \mathcal{L})^{-1} C_{jj}(\mathcal{L}, s') \\
=& 0.
\end{aligned}
$$

The last equality follows directly from the construction of stitching for $w(\mathcal{L})$.

□

of Theorem 4.2.4. For two arbitrary locations $s_1, s_2 \in \mathcal{D}$, we can calculate the covariance function from our construction as follows:

$$
\begin{aligned}
M_{ij}(s_1, s_2) &= Cov\left(C_{ii}(s_1, \mathcal{L}) C_{ii}(\mathcal{L}, \mathcal{L})^{-1} w_i(\mathcal{L}) + z_i(s_1),\right. \\
&\qquad \left. C_{jj}(s_2, \mathcal{L}) C_{jj}(\mathcal{L}, \mathcal{L})^{-1} w_j(\mathcal{L}) + z_j(s_2)\right) \\
&= C_{ii}(s_1, \mathcal{L}) C_{ii}(\mathcal{L}, \mathcal{L})^{-1} Cov(w_i(\mathcal{L}), w_j(\mathcal{L})) C_{jj}(\mathcal{L}, \mathcal{L})^{-1} C_{jj}(\mathcal{L}, s_2)) + \\
&\qquad \mathbb{I}(i = j) C_{ii \mid \mathcal{L}}(s_1, s_2) \\
&= \mathbb{I}(i = j)[C_{ii}(s_1, \mathcal{L}) C_{ii}(\mathcal{L}, \mathcal{L})^{-1} C_{ii}(\mathcal{L}, s_2)) + C_{ii \mid \mathcal{L}}(s_1, s_2)] + \\
&\qquad \mathbb{I}(i \neq j) C_{ii}(\boldsymbol{s}_1, \mathcal{L}) C_{ii}(\mathcal{L}, \mathcal{L})^{-1} M_{ij}(\mathcal{L}, \mathcal{L}) C_{jj}(\mathcal{L}, \mathcal{L})^{-1} C_{jj}(\mathcal{L}, s_2) \\
&= \mathbb{I}(i = j) C_{ii}(s_1, s_2) + \\
&\qquad \mathbb{I}(i \neq j) C_{ii}(s_1, \mathcal{L}) C_{ii}(\mathcal{L}, \mathcal{L})^{-1} M_{ij}(\mathcal{L}, \mathcal{L}) C_{jj}(\mathcal{L}, \mathcal{L})^{-1} C_{jj}(\mathcal{L}, s_2))
\end{aligned}
\tag{S1}
$$

The second equality follows from the independence of z_i and z_j for $i \neq j$, the third equality uses $M_{ii}(\mathcal{L}\mathcal{L}) = C_{ii}(\mathcal{L}, \mathcal{L})$ and the fourth uses the form of the conditional covariance function $C_{ii \mid \mathcal{L}}$ from (4.2). It is now immediate, that w_i has the covariance function C_{ii} on the entire domain $\mathcal{D}$, proving Part (a).

To prove part (b), without loss of generality we only consider $q = 3$ processes $w_1(s), w_2(s), w_3(s)$ which is constructed via stitching, with the assumption that $(1, 3) \notin E_{\mathcal{V}}$. First, we will show that, for any two locations

$s_1, s_2 \in \mathcal{D}$, $w_1(s_1)$ is conditionally independent of $w_3(s_2)$ given $w_2(\mathcal{L})$, which we denote as $w_1(s_1) \perp\!\!\!\perp w_3(s_2) \,|\, w_2(\mathcal{L})$.

As $(1,3) \notin E_{\mathcal{V}}$, the sets $\{1 \times \mathcal{L}\} = \{(1,s) \,|\, s \in \mathcal{L}\}$ and $\{3 \times \mathcal{L}\}$ are separated by $\{2 \times \mathcal{L}\}$ in the graph $\mathcal{G}_V \boxtimes \mathcal{G}_L$. Hence, using the global Markov property of Gaussian graphical models, we have $w_1(\mathcal{L}) \perp\!\!\!\perp w_3(\mathcal{L}) \,|\, w_2(\mathcal{L})$.

For any $s_1, s_2 \in \mathcal{D}$ we have, similar to (S1),

$$\begin{aligned} &Cov(w_1(s_1) w_3(s_2) \,|\, w_2(\mathcal{L})) \\ &= C_{11}(s_1, \mathcal{L}) C_{11}(\mathcal{L}, \mathcal{L})^{-1} Cov\left(w_1(\mathcal{L}), w_3(\mathcal{L}) \,|\, w_2(\mathcal{L})\right) C_{33}(\mathcal{L}, \mathcal{L})^{-1} C_{33}(\mathcal{L}, s_2) = 0. \end{aligned}$$

Hence, $w_1(s_1) \perp\!\!\!\perp w_3(s_2) \,|\, w_2(\mathcal{L})$ for any $s_1, s_2 \in \mathcal{D}$. Now

$$\begin{aligned} &Cov(w_1(s_1),\ w_3(s_2) \,|\, \sigma\left(\{w_2(s) \,|\, s \in \mathcal{D}\}\right)) \\ &= Cov(w_1(s_1),\ w_3(s_2) \,|\, \sigma\left(w_2(\mathcal{L}), \{z_2(s) \,|\, s \in \mathcal{D}\}\right)) \\ &= Cov(w_1(s_1),\ w_3(s_2) \,|\, \sigma\left(w_2(\mathcal{L})\right)) = 0. \end{aligned} \tag{S2}$$

The second inequality follows due to the agreement of the two conditioning σ-algebras (similar to the argument in the proof of Lemma 4.2.3). The third inequality follows from the fact that for any three random variables X, Y and Z such that X and Y are independent of Z, $E(X|Y,Z) = E(X|Y)$. Equation (S2) establishes process level conditional independence for $w_1(\cdot)$ and $w_3(\cdot)$ given $w_2(\cdot)$, thereby proving part (b).

Finally, if $(i,j) \in E_{\mathcal{V}}$, and $(s_1, s_2) \in \mathcal{L}$, in (S1), we will have $M_{ij}(s_1, s_2) = C_{ij}(s_1, s_2)$ directly from the construction of $M(\mathcal{L}, \mathcal{L})$. This proves part (c).

□

of Corollary 4.3.0.1. Recall from the construction of $M(\mathcal{L}, \mathcal{L})$ that the Gaussian random vector $w(\mathcal{L})$ satisfies the graphical model $\mathcal{G} = \mathcal{G}_{\mathcal{V}} \boxtimes \mathcal{G}_{\mathcal{L}}$, where $\mathcal{G}_{\mathcal{L}}$ is the complete graph between n locations. The strong product graph $\mathcal{G}$ is

decomposable and $K_m \boxtimes \mathcal{G}_{\mathcal{L}}; m = 1, \cdots, p$ form a perfect sequence for $\mathcal{G}$ with $S_m \boxtimes \mathcal{G}_{\mathcal{L}}; m = 2, \cdots, p$ being the separators. Thus, using results (3.17) and (5.44) from Lauritzen, 1996, we are able to factorize $w(\mathcal{L})$ as (4.6). □

of Proposition 4.3.1. Suppose, we observe a Multivariate Matérn process. Under the assumption of Graphical Gaussian Processes, the resulting maximum likelihood estimating equations for parameters θ_{ij} belonging to cliques or separators ($i = j$ or $(i, j) \in E_{\mathcal{V}}$) are given by -

$$\frac{\partial \log(f_M(w(\mathcal{L}))}{\partial \theta_{ij}} = 0$$

$$\implies \frac{\partial}{\partial \theta_{ij}} \left(\sum_K \log(f_C(w_K(\mathcal{L})) - \sum_S \log(f_C(w_S(\mathcal{L})) \right) = 0 \tag{S3}$$

$$\implies \frac{\partial}{\partial \theta_{ij}} \left(\sum_{K \ni (i,j)} \log(f_C(w_K(\mathcal{L})) - \sum_{S \ni (i,j)} \log(f_C(w_S(\mathcal{L})) \right) = 0 \,,$$

where for a subset a, $\log(f_C(w_a(\mathcal{L})) = -\frac{1}{2} w_a(\mathcal{L})^T C_a(\theta)^{-1} w_a(\mathcal{L}) - \log|\det(C_a(\theta)|$.

Below, we will show that for every subset (clique or separator), the maximum likelihood estimating equation is unbiased.

$$
\begin{aligned}
& E_w \left[\frac{\partial}{\partial \theta_{ij}} (-\frac{1}{2} w_a(\mathcal{L})^{\mathrm{T}} C_a(\theta)^{-1} w_a(\mathcal{L}) - \log |\det(C_a(\theta))|) \right] \\
& = E_w \left[-\frac{1}{2} \operatorname{tr} \left(C_a(\theta)^{-1} w_a(\mathcal{L}) w_a(\mathcal{L})^{\mathrm{T}} C_a(\theta)^{-1} \frac{\partial C_a(\theta)}{\partial \theta_{ij}} \right) + \frac{1}{2} \operatorname{tr} \left(C_a(\theta)^{-1} \frac{\partial C_a(\theta)}{\partial \theta_{ij}} \right) \right] \\
& = -\frac{1}{2} \operatorname{tr} \left(C_a(\theta)^{-1} E_w \left[w_a(\mathcal{L}) w_a(\mathcal{L})^{\mathrm{T}} \right] C_a(\theta)^{-1} \frac{\partial C_a(\theta)}{\partial \theta_{ij}} \right) + \frac{1}{2} \operatorname{tr} \left(C_a(\theta)^{-1} \frac{\partial C_a(\theta)}{\partial \theta_{ij}} \right) \\
& = -\frac{1}{2} \operatorname{tr} \left(C_a(\theta)^{-1} \frac{\partial C_a(\theta)}{\partial \theta_{ij}} \right) + \frac{1}{2} \operatorname{tr} \left(C_a(\theta)^{-1} \frac{\partial C_a(\theta)}{\partial \theta_{ij}} \right) = 0 \, .
\end{aligned}
\tag{S4}
$$

Since the Graphical Gaussian process likelihood is made up of the sums and differences of individual clique and separator likelihoods, the above result ensures that under true parameter values θ_{ij} of the Multivariate Matérn, we obtain the following which concludes our proof -

$$
E_w \left[\frac{\partial}{\partial \theta_{ij}} \left(\sum_{K \ni (i,j)} \log(f_C(w_K(\mathcal{L})) - \sum_{S \ni (i,j)} \log(f_C(w_S(\mathcal{L})) \right) \right] = 0 \tag{S5}
$$

□

of Proposition 4.4.1. We only need to prove $\operatorname{Cov}(w_i(s), w_j(s') \,|\, \sigma(\{f_j(s) \,|\, j \in 1, \ldots, r, s \in \mathcal{D}\}) = 0$ for all $i \neq j$ and $s, s' \in \mathcal{D}$. From Equation (4.9), we

have $w_i(s) = a_i(s)^{\mathrm{T}} f(s) + \xi_i(s)$ where $a_i(s) = (a_{i1}(s), \ldots, a_{ir}(s))^{\mathrm{T}}$.

$$\mathrm{Cov}(w_i(s),\, w_j(s') \,|\, \sigma\left(\{f_j(s) \,|\, j = 1, \ldots, r, s \in \mathcal{D}\}\right))$$

$$= \mathrm{Cov}(a_i(s)^{\mathrm{T}} f(s) + \xi_i(s),\; a_j(s)^{\mathrm{T}} f(s') + \xi_j(s') \,|\, \sigma\left(\{f_j(s) \,|\, j = 1, \ldots, r, s \in \mathcal{D}\}\right))$$

$$= \mathrm{Cov}(a_i(s)^{\mathrm{T}} f(s),\; a_j(s)^{\mathrm{T}} f(s') \,|\, \sigma\left(\{f_j(s) \,|\, j = 1, \ldots, r, s \in \mathcal{D}\}\right)) + \mathrm{Cov}(\xi_i(s), \xi_j(s'))$$

$$= 0 + 0 = 0$$

because $a_i(s)^{\mathrm{T}} f(s)$'s are deterministic functions of the conditioning σ-algebra, and ξ_i's are independent of each other and of the factor processes.

Thus we have proved that any pair of observed processes are conditionally independent given the latent processes. When translated into a Graphical Gaussian processes framework, we will observe no edges between the observed nodes and each observed node will be connected to all the factor (latent) nodes. Additionally, we assume all possible connections (a complete graph) between the factor nodes in their marginal distribution. This gives us a complete graph between the vertices $\{q + j | j \in 1, \ldots, r\}$. Therefore, the graphs on the joint set of observed and factor processes will be decomposable with the perfect ordering of cliques $K_1, \ldots, K_q$ where $K_i = \{i\} \cup \{q + i | i \in 1, \ldots, r\}$. □

S2 Implementation

S2.1 Gibbs sampler for GGP model for the latent processes

Let $y_i = (y_i(s_{i1}), y_i(s_{i2}), \ldots, y_i(s_{in_i}))^{\mathrm{T}}$ be the $n_i \times 1$ vector of measurements for the i-th response or outcome over the set of n_i locations in $\mathcal{D}$. Let $X_i = (x_i(s_{i1}), x_i(s_{i2}), \cdots, x_i(s_{in_i}))^{\mathrm{T}}$ be the known $n_i \times p_i$ matrix of predictors on the

set $\mathcal{S}_i = \{s_{i1}, \cdots, s_{in_i}\}$. We specify the spatial linear model as $y_i = X_i\beta_i + w_i + \epsilon_i$, where β_i is the $p_i \times 1$ vector of regression coefficients, ϵ_i is the $n_i \times 1$ vector of normally distributed random independent errors with marginal common variance τ_i^2, and w_i is defined analogously to y_i for the latent spatial process corresponding to the i-th outcome. The distribution of each w_i is derived from the specification of $w(s)$ as the $q \times 1$ multivariate graphical Matérn GP with respect to a decomposable $\mathcal{G}_\mathcal{V}$. Let $\{\phi_{ii}, \sigma_{ii}, \tau_i^2 | i = 1\ldots, q\}$ denote the marginal parameters for each component Matérn process $w_i(\cdot)$.

We elucidate the sampler using a GGP constructed by stitching the simple multivariate Matérn (Apanasovich, Genton, and Sun, 2012), where $\nu_{ij} = (\nu_{ii} + \nu_{jj})/2$, $\Delta_A = 0$ in (4.4) and $\phi_{ij}^2 = (\phi_{ii}^2 + \phi_{jj}^2)/2$. Hence, the only additional cross-correlation parameters are $\{b_{ij} | (i, j) \in E_\mathcal{V}\}$. Any of the other multivariate Matérn specifications in Apanasovich, Genton, and Sun, 2012 that involve more parameters to specify ν_{ij}'s and ϕ_{ij}'s can be implemented in a similar manner. We consider partial overlap between the variable-specific location sets and take $\mathcal{L} = \cup_i \mathcal{S}_i$ as the reference set for stitching. If there is total lack of overlap between the data locations for each variable, we can simply take $\mathcal{L}$ to be a set of locations sufficiently well distributed in the domain and the Gibbs sampler can be designed analogously.

Conjugate priors are available for $\beta_i \overset{ind}{\sim} N(\mu_i, V_i)$ and $\tau_i^2 \overset{ind}{\sim} IG(a_i, b_i)$, where IG is the Inverse-Gamma distribution. There are no conjugate priors for the process parameters. For ease of notation, the collection $M_{a,b}$ the submatrix of M indexed by sets a and b, $M_a = M_{a,a}$, and $M_{a|b} = M_a - M_{a,b}M_b^{-1}M_{b,a}$. Similarly, we denote $w(a)$ to be the vector stacking $w_i(s)$ for all $(i, s) \in a$. We

denote cliques by K and separators by S in the perfect ordering of the graph $\mathcal{G}_V$.

The full-conditional distributions for the Gibbs updates of the parameters are as follows.

$$p(\beta_i \mid \cdot) \sim N((X_i^{\mathrm{T}} X_i + V_i^{-1})^{-1}(\mu_i + X_i^{\mathrm{T}}(y_i - w_i)), \tau_i^2 (X_i^{\mathrm{T}} X_i + V_i^{-1})^{-1}) \; ;$$

$$p(\tau_i^2 \mid \cdot) \sim IG(a + \frac{n_i}{2}, b + \frac{(y_i - X_i^{\mathrm{T}}\beta_i - w_i)^{\mathrm{T}}(y_i - X_i^{\mathrm{T}}\beta_i - w_i)}{2}) \; ;$$

$$p(\sigma_{ii}, \phi_{ii}, \nu_{ii} \mid \cdot) \propto \frac{\prod_{K \ni i} \frac{1}{|M_{K\times\mathcal{L}}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} w(K\times\mathcal{L})^{\mathrm{T}} M_{K\times\mathcal{L}}^{-1} w(K\times\mathcal{L})\right)}{\prod_{S \ni i} \frac{1}{|M_{S\times\mathcal{L}}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} w(S\times\mathcal{L})^{\mathrm{T}} M_{S\times\mathcal{L}}^{-1} w(S\times\mathcal{L})\right)} \times p(\sigma_{ii}) p(\phi_{ii}) p(\nu_{ii}) \; ;$$

$$p(b_{ij} \mid \cdot) \propto \frac{\prod_{K \ni (i,j)} \frac{I(B_K>0)}{|M_{K\times\mathcal{L}}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} w(K\times\mathcal{L})^{\mathrm{T}} M_{K\times\mathcal{L}}^{-1} w(K\times\mathcal{L})\right)}{\prod_{S \ni (i,j)} \frac{1}{|M_{S\times\mathcal{L}}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} w(S\times\mathcal{L})^{\mathrm{T}} M_{S\times\mathcal{L}}^{-1} w(S\times\mathcal{L})\right)} \times p(b_{ij})$$

$$\text{for } (i.j) \in E_{\mathcal{V}} \; .$$

To update the latent random effects w, let $\mathcal{L} = \{s_1, \ldots, s_n\}$ and $o_i = \mathrm{diag}(I(s_1 \in \mathcal{S}_i), \ldots, I(s_n \in \mathcal{S}_i))$ denote the vector of missing observations for the i-th outcome. With $X_i(\mathcal{L}) = (x_i(s_1), \ldots, x_i(s_n))^{\mathrm{T}}$, $y_i(\mathcal{L})$ and $w_i(\mathcal{L})$ defined similarly,

we obtain

$$p(w_i(\mathcal{L})\,|\,\cdot) \sim N\left(\mathcal{M}_i^{-1}\mu_i, \mathcal{M}_i^{-1}\right) \text{ , where}$$

$$\mathcal{M}_i = \frac{1}{\tau_i^2}\text{diag}(o_i) + \sum_{K\ni i} M^{-1}_{\{i\}\times\mathcal{L}|(K\setminus\{i\})\times\mathcal{L}} - \sum_{S\ni i} M^{-1}_{\{i\}\times\mathcal{L}|(S\setminus\{i\})\times\mathcal{L}} \text{ ,}$$

$$\mu_i = \frac{(y_i(\mathcal{L}) - x_i(\mathcal{L})^{\mathrm{T}}\beta_i)\odot o_i}{\tau_i^2} +$$

$$\sum_{K\ni i} T_i(K) w((K\setminus\{i\})\times\mathcal{L}) - \sum_{S\ni i} T_i(S) w((S\setminus\{i\})\times\mathcal{L}) \text{ ,}$$

$$T_i(A) = M^{-1}_{\{i\}\times\mathcal{L}|(A\setminus\{i\})\times\mathcal{L}} M_{\{i\}\times\mathcal{L},(A\setminus\{i\})\times\mathcal{L}} M^{-1}_{(A\setminus\{i\})\times\mathcal{L}}\text{, for } A\in\{K,S\}.$$

The Gibbs sampler evinces the multifaceted computational gains. The constraints on the parameters b_{ij} no longer require checking the positive-definiteness of B, which would require $O(q^3)$ flops for each check. Instead, due to decomposability it is enough to check for positive definiteness of the (at most q^* dimensional) sub-matrices B_K of B corresponding to the cliques of $\mathcal{G}_\mathcal{V}$. The largest matrix inversion across all these updates is of the order $nq^* \times nq^*$, corresponding to the largest clique. The largest matrix that needs storing is also of dimension $nq^* \times nq^*$. These result in appreciable reduction of computations from any multivariate Matérn model that involves $nq \times nq$ matrices and positive-definiteness checks for $q \times q$ matrices at every iteration.

Finally, for generating predictive distributions, note that, as a part of the Gibbs sampler, we are simultaneously imputing w_i at the locations $\mathcal{L}\setminus\mathcal{S}_i$. Subsequently, we only need to sample $y_i(\mathcal{L}\setminus\mathcal{S}_i)\,|\,\cdot \sim N(X_i(\mathcal{L}\setminus\mathcal{S}_i)'\beta_i + w_i(\mathcal{L}\setminus\mathcal{S}_i), \tau_i^2 I)$.

S2.2 Gibbs sampler for GGP model for the response processes

Let $y(\mathcal{L}) = (y_1(\mathcal{L}), \cdots, y_q(\mathcal{L}))^{\mathrm{T}}$, $X(\mathcal{L}) = \mathrm{bdiag}(X_1(\mathcal{L}), \ldots, X_q(\mathcal{L}))$, and $\beta = (\beta_1^{\mathrm{T}}, \cdots, \beta_q^{\mathrm{T}})^{\mathrm{T}}$. We will consider the joint likelihood

$$\mathbf{y}(\mathcal{L}) \,|\, X(\mathcal{L}), \beta, \{\phi_{ii}, \sigma_{ii}, \tau_i^2\}_{\{i=1,\ldots,q\}}, \{b_{ij}\}_{\{(i,j)\in E_{\mathcal{V}}\}} \sim N\left(X(\mathcal{L})\beta, M^*_{V\times\mathcal{L}}\right) \quad \text{(S6)}$$

and impute the missing data $y_i(\mathcal{L}\backslash\mathcal{S}_i)$ in the sampler. Let $\mathcal{T}_i = \{i\} \times (\mathcal{L}\backslash\mathcal{S}_i)$, $U_i(A) = (A \times \mathcal{L})\backslash\mathcal{T}_i$ for $A \in \{K, S\}$ and $\beta(A)$ be the vector stacking up β_j for $j \in A$. Also, for any $U \subseteq \mathcal{V} \times \mathcal{L}$, let $\tilde{X}(U) = bdiag(\{X_j(U \cap (\{j\} \times \mathcal{L}))|j \ni U \cap (\{j\} \times \mathcal{L}) \neq \{\}\}$. We have the following updates:

$$y_i(\mathcal{T}_i) \,|\, \cdot \sim N(X_i(\mathcal{T}_i)\beta_i +$$

$$H_i^{-1}\left(\sum_{K\ni i} M^{*-1}_{\mathcal{T}_i|U_i(K)} M^{*-1}_{\mathcal{T}_i,\, U_i(K)} M^{*-1}_{U_i(K)} (y(U_i(K)) - \tilde{X}(U_i(K))\beta(K)) - \right.$$

$$\left. \sum_{S\ni i} M^{*-1}_{\mathcal{T}_i|U_i(S)} M^{*-1}_{\mathcal{T}_i,\, U_i(S)} M^{*-1}_{U_i(S)} (y(U_i(S)) - \tilde{X}(U_i(S))\beta(S))\right), H_i^{-1})$$

where $H_i = \sum_{K\ni i} M^{*-1}_{\mathcal{T}_i|U_i(K)} - \sum_{S\ni i} M^{*-1}_{\mathcal{T}_i|U_i(S)}$

Once again the updates require inversion or storage of matrices of size at most $nq^* \times nq^*$. The updates for the other parameters are similar to that in the sampler of Section S2.1 of the Supplement with the cross-covariance M^* replacing M. The only exception is τ_i^2, which no longer has conjugate full conditionals and are also now updated using Metropolis random walk steps within the Gibbs sampler akin to the other spatial parameters.

S2.3 Reversible jump MCMC algorithm

We use the reversible jump MCMC (rjMCMC) algorithm of Barker and Link, 2013 to carry out the multimodel inference by sampling of the graph and estimating the cross-covariance parameters specific to the graph. We embed the graph sampling described in Section 4.4.3 within the Gibbs sampler in Section S2.1. Jumps between graphs need to be coupled with introduction or deletion of cross-covariance parameters depending on addition or deletion of edges. In order to facilitate this, we need a bijection between the parameter sets of the GGP models corresponding to two different graphs. This is achieved by creating a universal parameter (palette) ψ from which all model-specific (graph-specific) parameters can be computed. For example, if we assume the k-th graph $\mathcal{G}_{\mathcal{V}_k} = (\mathcal{V}, E_k)$ has θ_k as the cross-covariance parameter vector, then we need to define an invertible mapping g_k such that $g_k(\psi) = c(\theta_k, u_k)$, where u_k's are irrelevant to graph k. In our case, we define ψ to be the concatenated vector of length $\frac{q(q-1)}{2}$ containing all pairwise cross-covariance parameters, i.e. $\psi = (\psi_{12}, \psi_{13}, \cdots, \psi_{23}, \cdots, \psi_{(q-1),q})$. We define $g_k(\psi) = \psi^{(k)} = [\{\psi_{ij}^{(k)} : (i,j) \in E_k\}, \{\psi_{ij}^{(k)} : (i,j) \notin E_k\}]$ to be the permuted vector of ψ.

Using the above setup, we now devise our two-step sampling strategy for the graphs. From the current junction tree J, we propose a move to a new junction tree J' by adding or deleting edges. Following Green and Thomas, 2013 we calculate the proposal probabilities as $\kappa(J, J')$. The acceptance probability of the new junction tree J' is $\alpha(J, J') = \min\left(1, \frac{p(y|\psi, J', \cdot)\tilde{\pi}(J')\kappa(J, J')}{p(y|\psi, J, \cdot)\tilde{\pi}(J)\kappa(J', J)}\right)$.

Exploiting the factorisation (4.6) of stitched GGP likelihoods for decomposable graphs, we can simplify computations in $\alpha(J, J')$. Let K_J, S_J be the set

of cliques and separators for J. Let $K^+(J,J'), K^-(J,J')$ and $S^+(J,J'), S^-(J,J')$ denote, respectively, the cliques and separators added and deleted by the proposed move to J'. The ratio for a proposed move from J to J' is

$$p(J \longrightarrow J' \mid \psi, \cdot) = \frac{\prod_{K \in K^+(J,J')} \frac{I(B_K>0)}{|M_{K\times\mathcal{L}}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} w_K(\mathcal{L})^{\mathrm{T}} M^{-1}_{K\times\mathcal{L}} w_K(\mathcal{L})\right)}{\prod_{K \in K^-(J,J')} \frac{I(B_K>0)}{|M_{K\times\mathcal{L}}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} w_K(\mathcal{L})^{\mathrm{T}} M^{-1}_{K\times\mathcal{L}} w_K(\mathcal{L})\right)} \times$$

$$\frac{\prod_{S \in S^-(J,J')} \frac{1}{|M_{S\times\mathcal{L}}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} w_S(\mathcal{L})^{\mathrm{T}} M^{-1}_{S\times\mathcal{L}} w_S(\mathcal{L})\right) \kappa(J,J') \mu(\mathcal{G}_{\mathcal{V}}(J))}{\prod_{S \in S^+(J,J')} \frac{1}{|M_{S\times\mathcal{L}}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} w_S(\mathcal{L})^{\mathrm{T}} M^{-1}_{S\times\mathcal{L}} w_S(\mathcal{L})\right) \kappa(J',J) \mu(\mathcal{G}_{\mathcal{V}}(J'))}.$$

The terms corresponding to the cliques $K^-(J,J')$ and separators $S^-(J,J')$ have already been computed from the existing tree J. We only need to evaluate the likelihood factors corresponding to new cliques $K^+(J,J')$ and separators $S^+(J,J')$ added in the proposed tree J'. This makes the jumps between junction trees computationally efficient for the GGP likelihood. Subsequent to moving to a new tree, we modify the Gibbs' sampler (Section S2.1) to sample the cross-correlation parameters as below.

$$p(\psi_{ij}^{(k)}; (i,j) \in E_k \mid J, \cdot) \propto \frac{\prod_{K_J \ni (i,j)} \frac{I(B_K>0)}{|M_{K\times\mathcal{L}}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} w_K(\mathcal{L})^{\mathrm{T}} M^{-1}_{K\times\mathcal{L}} w_K(\mathcal{L})\right)}{\prod_{S_J \ni (i,j)} \frac{1}{|M_{S\times\mathcal{L}}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} w_S(\mathcal{L})^{\mathrm{T}} M^{-1}_{S\times\mathcal{L}} w_S(\mathcal{L})\right)} \times p(\psi_{ij}^{(k)})$$

$$p(\psi_{ij}^{(k)}; (i,j) \notin E_k \mid J, \cdot) \propto p(\psi_{ij}^{(k)}).$$

S2.4 Co-ordinate descent

To conduct estimation and prediction using GGP in a frequentist setting, we outline a co-ordinate descent algorithm for maximum likelihood estimation (assuming a known graph). We illustrate the implementation for the case where each of the q variables are measured at $\mathcal{L}$. The case of spatial misalignment can be handled by an EM algorithm to impute the missing responses for each variable. For the frequentist setup, we use the GGP model for the response. From Corollary 4.3.0.1, the joint likelihood can be factored into sub-likelihoods corresponding to specific cliques and separators. Let $\theta^{(t)}$ denote the values of the spatial parameters θ at the t-th iteration, and $M^*_{\mathcal{L}} = M^*_{\mathcal{L}}(\theta)$ denote the GGP covariance matrix of $y(\mathcal{V} \times \mathcal{L})$ from stitching. Let $\theta_{ii} = \{\sigma^2_{ii}, \phi_{ii}, \nu_{ii}\}$, $\theta_{-i} = \theta \backslash \theta_{ii}$, $\theta_{-ij} = \theta \backslash \{b_{ij}\}$. Letting $\tilde{X}(\mathcal{L}) := \tilde{X}(\mathcal{V} \times \mathcal{L})$ we immediately have the following updates of the parameters:

$$\beta^{(t+1)} = \left(\tilde{X}(\mathcal{L})^{\mathrm{T}} M^{*-1}_{\mathcal{L}}(\theta^{(t)}) \tilde{X}(\mathcal{L})\right) \tilde{X}(\mathcal{L})^{\mathrm{T}} M^{*-1}_{\mathcal{L}}(\theta^{(t)}) y(\mathcal{L}),$$

$$\theta^{(t+1)}_{ii} = \arg\min_{\theta_{ii}} \Big[\sum_{K \ni i} l_K(\theta_{ii}) - \sum_{S \ni i} l_S(\theta_{ii}) \Big], \text{ where for any } A \subset \mathcal{V},$$

$$l_A(\theta_{ii}) = \log(|M^*_{A \times \mathcal{L}}(\theta_{ii}, \theta^{(t)}_{-i})|) +$$

$$(y(A \times \mathcal{L}) - \tilde{X}(A \times \mathcal{L})\beta(A))^{\mathrm{T}} M^{-*1}_{A \times \mathcal{L}}(\theta_{ii}, \theta^{(t)}_{-i})(y(A \times \mathcal{L}) - \tilde{X}(A \times \mathcal{L})\beta(A)),$$

$$b^{(t+1)}_{ij} = \arg\min_{b_{ij}} \Big[\sum_{K \ni (i,j)} \big(\tilde{\ell}_K(b_{ij}) - \log(I(B_K > 0))\big) - \sum_{S \ni (i,j)} \tilde{\ell}_S(b_{ij}) \Big], \text{ for } (i,j) \in E_{\mathcal{V}},$$

where $\tilde{\ell}_A(b_{ij}) = \log(|M^*_{A \times \mathcal{L}}(b_{ij}, \theta^{(t)}_{-ij})|) +$

$$(y(A \times \mathcal{L}) - \tilde{X}(A \times \mathcal{L})\beta(A))^{\mathrm{T}} M^{-*1}_{A \times \mathcal{L}}(b_{ij}, \theta^{(t)}_{-ij})(y(A \times \mathcal{L}) - \tilde{X}(A \times \mathcal{L})\beta(A)).$$

The update of β involves the large $nq \times nq$ matrix $M_{\mathcal{L}}^{*-1}$. However, from (4.7), $M_{\mathcal{L}}^{*-1}$ can be expressed as sum of sparse matrices, each requiring at-most $O(n^3 q^{*3})$ storage and computation arising from inverting matrices of the form $C_{K\boxtimes\mathcal{L}} + D_{K\boxtimes\mathcal{L}}$. For updates of the spatial parameters θ_{ii} and b_{ij}, coordinate descent moves along the respective parameter and optimizing the negative log-likelihood which is expressed in terms of the corresponding negative log-likelihoods of the cliques and separators containing that parameter. This process is iterated until convergence. Each iteration of the co-ordinate descent has the same complexity of parameter dimension, same computation and storage costs and parameter constraint check as each iteration of the Gibbs sampler, and hence is comparably scalable.

S3 Additional data analyses results

S3.1 Estimation of marginal parameters

Comparison of the estimates of the marginal (variable-specific) parameters $\theta_{ii} = (\sigma_{ii}, \phi_{ii})'$ is of lesser importance because stitching ensures that each univariate process is Matérn GP, similar to the competing multivariate Matérn model. The estimates of the marginal microergodic parameters $\sigma_{ii}\phi_{ii}$ are plotted in Figure S1 of the Supplement and reveal similar trends to Figure 4.5, with MM and GM accurately estimating the parameters while PM producing poor estimates due to parameter constraints imposed by its simplifying assumptions. Also, the estimates of the regression coefficients β_j were accurate for all models, and are not presented.

S3.2 Estimates of cross-correlation function under misspecification

We also assess the impact of GGP not excluding parameters b_{ij} for all $(i, j) \notin E_{\mathcal{V}}$ on the estimates of the cross-correlation functions for these variable pairs. Since these parameters are not in the GGP, we can only compare the true cross-covariance function between these variables pairs against the one indirectly estimated by GGP.

For the misspecified case of Set 1B, Figure S2 shows that GM estimates the cross-correlation function pretty well for the assumed edges (blue background) and show bias for some of the variable pairs not included in $\mathcal{G}_{\mathcal{V}}$ (white background). The accurate estimation for the covariance functions (diagonal plots) is attributable to the GGP exactly preserving the marginal distributions of the multivariate Matérn. Similarly, the estimates of the cross-covariance parameters for $(i, j) \in E_{\mathcal{V}}$ is expectedly accurate (as is concluded in Proposition 4.3.1).

The bias observed in estimates of the cross-correlation for some pairs of $(i, j) \notin E_{\mathcal{V}}$ is also unsurprising. The multivariate Matérn used to generate the data does not follow any graphical model or any other low-rank structure, and any form of dimension-reduction (like modelling dependencies with a sparse graph) will lead to some tradeoff in terms of accuracy and scalability. For analyzing highly multivariate spatial data, even if the variables truly doesn't conform to any graphical model, the MM is not a feasible option due to its high-dimensional parameter space and computing requirements (Table 4.1) and hence dimension-reduction is necessary. Hence, using GGP

with a reasonably chosen graphical model that does not exclude important variable pairs is a necessary dimension-reduction step. While it is challenging to establish a bound for the bias for excluded edges in GGP, we have proved that the marginal-preserving GGP is the information-theoretically optimal approximation of a full GP among the class of all GGP (Theorem 4.2.1) .

We see from Figure that S2 that the bias from GM is worse than that of PM for some $(i, j) \notin E_{\mathcal{V}}$ (e.g., $(1, 4)$ or $(2, 4)$). On the other hand, estimates for PM are worse for some $(i, j) \notin E_{\mathcal{V}}$ (e.g., $(1, 3)$) and for a majority of the pairs (i, i) and $(i, j) \in E_{\mathcal{V}}$. PM achieves dimension-reduction by imposing simplifying parameter constraints which degrades its estimation accuracy substantially for most parameters. Moreover, PM cannot even be implemented in the truly highly multivariate settings (like sets 3A and 3B) due to requiring $O(q^3)$ for likelihood evaluation (see Table 4.1). The GGP offers drastic improvement in scalability over these alternatives, and for highly multivariate settings, maybe the only viable option guaranteeing accurate estimation of a large subset of the full model parameters. Additionally, we see that exclusion of edges does not severely impact the prediction quality of GM.

S3.3 Comparison with linear model of coregionalization

To compare relative performance of linear model of coregionalization and GGP in modelling low-rank processes, we consider the following simulation scenarios: (i) data is generated from an linear model of coregionalization; (ii) data is generated from a Graphical Matérn (GM) respecting the graphical model that would arise from the linear model of coregionalization in scenario

(i). For each simulation setting we fit GM and linear model of coregionalization with two factor processes (using *spMisalignLM* function from the R package *spBayes* for our setting of variable-specific locations). Since *spMisalignLM* can be implemented only when the number of observed and latent processes are equal, for generating the data we considered two observed process based on two independent factor processes. This linear model of coregionalization leads to the graphical model from Figure 4.3a. Hence, for scenario (ii) we generate data from a graphical Matérn using this graph to generate correlated factor processes.

Since the two models correspond to different sets of parameters, we cannot compare them directly. Instead, in Figure S3 we compare the estimates of the entire correlation and cross-correlation functions. We observe that the impact of misspecification is more pronounced for the linear model of coregionalization; when the true model is graphical Matérn the estimate of the correlation function for the second variable by linear model of coregionalization is quite poor. In comparison, the graphical Matérn estimates the correlation and cross-correlation functions reasonably well both in the correctly specified and in the misspecified case.

We also compare the predictive performance of the models. For all the simulations performed, we leave out 20% of the data to create test sets in order to evaluate prediction accuracy of the models. Figure S4 presents the comparison of the prediction and the true values for both models and both data generation scenarios. GM reports marginally improved root mean square prediction error than linear model of coregionalization in all of the situations.

S3.4 Comparison with spatial dynamic linear models

The simulation set 3A corresponds to a highly multivariate setting ($q = 100$), where the multivariate process truly follows a graphical model (path graph among the 100 variables). None of the competing multivariate approaches besides the GGP can model such a graphical structure. These alternatives (PM and MM) also do not scale to our highly multivariate settings. In Section 4.4.2 we have illustrated how common univariate and multivariate spatial time-series correspond to decomposable graphical models among the variables and can be modelled using GGP. The path graph in setting 3A corresponds to the decomposable graph resulting from an AR(1) temporal structure. Hence, for this set we compare the performance of the GGP with a dynamic linear model (DLM) (Stroud, Müller, and Sansó, 2001; Gelfand, Banerjee, and Gamerman, 2005; Finley, Banerjee, and Gelfand, 2012) commonly used for modelling spatial time-series with AR(1) temporal evolution.

In particular, we compare GGP with the spatial dynamic linear model (*SpDynLm*) of Finley, Banerjee, and Gelfand, 2012 which is set in the GP-based mixed-effect modeling setup similar to (4.1), as opposed to spatial basis function based approach of (Stroud, Müller, and Sansó, 2001). SpDynLM models the spatial process $w_t(\cdot) = w(\cdot, t)$ at time t as

$$w_t(s) = w_{t-1}(s) + \delta_t(s); \quad \delta_t(\cdot) \sim GP(0, C_{tt}), \tag{S7}$$

i.e., at each time-point the spatial process is a sum of the process at the previous time point and an independent time-specific GP. The rest of the model is the same as in our setup (Eq. 4.1) with SpDynLM enforcing an auto-regressive

evolution model for the regression coefficients β_t as well.

Both SpDynLm and any GGP with a path graph between the time-specific variables model an AR(1) evolution over time. However, any DLM using an additive model of the type (S7) for the temporal evolution of the latent processes $w_t(\dot{c})$, unfortunately, enforces the processes $w_t(\cdot)$ to have the same smoothness at all time-points t. Thus, even if the $\delta_t(\cdot)$'s are modelled using Matérn GPs with time-specific smoothness, range and variance parameters, none of the processes $w_t(\cdot)$ will be Matérn GPs and each will have the smoothness of the roughest of the independent processes $\delta_t(\dot{c})$.

Another major restriction of the SpDynLM model is that (S7) is the customary *random walk prior* (Stroud, Müller, and Sansó, 2001) for $w_t(\cdot)$ which imposes the assumption that $\text{Var}(w_t(s)) > \text{Var}(w_{t-1}(s))$ for all t, s, i.e., that the process variance is monotonically increasing over time. For most spatiotemporal processes this assumption is unlikely to hold. The GGP, on the other hand, ensures that the processes $w_t(\cdot)$ for each time t can be modelled using a Matérn GP with time-specific variance parameters.

The dynamic model in (S7) also implicitly assumes a constant (over time) auto-regression coefficient of 1. While this can be easily relaxed by replacing $w_{t-1}(s)$ with $\rho_t w_{t-1}(s)$ in (S7), the current implementation of SpDynLm does not allow modelling such a non-stationary auto-correlation coefficient ρ_t. In a GGP with a path graph, the cross-correlation parameter $b_{t,t-1}$ between the processes at two consecutive times is time-specific, thereby allowing it to capture non-stationary auto-regressive structures.

For Set 3A, the estimation accuracy of GM has already been demonstrated

in Figure S1(e) (for the marginal parameters) and in Figure S9(c) (for the cross-covariance parameters representing the auto-regression). In particular, Figure S9(c) demonstrates the capability of GGP to successfully estimate non-stationary (time-specific) cross-covariance parameters. The competing model SpDynLm does not possess an autocovariance parameter that can been compared with these cross-covariances. However, we compare estimates of the marginal process variances from GGP and the SpDynLm function with the truth in Figure S5a. We observe that while GGP accurately captures the marginal variances of the processes $w_t(\cdot)$ for each time t, the estimates from SpDynLm are monotonically increasing with time and far exceeding the true values. This demonstrates the detrimental implications of the model in S7 leading to variances exploding with time and, hence, prohibiting any meaningful insight regarding the underlying processes from these parameter estimates.

We also compare the models based on their predictive performance on hold-out data. We use the implementation of SpDynLm in the SpBayes R-package (Finley, Banerjee, and Gelfand, 2013). Figure S5b plots the median RMSPE for each variable (time-point). We see that SpDynLm produces higher predictive error (RMSPE=0.868) than GM (RSMPE=0.8) for most time-points. The numbers reported in parentheses are averaged across variables.

Overall, while predictive performance between GGP and DLM is competitive, the GGP is flexible and interpretable allowing estimation of spatial properties of the latent process $w_t(\cdot)$ for each time. The current implementation of SpDynLM imposes unnecessary constraints of monotonically increasing

latent process variance with time leading to meaningless estimates of these parameters. More importantly, DLM assumes common smoothness over time, thereby offering no avenue to quantitatively study smoothness of the process at each time which can reveal important scientific phenomenon, e.g., pollutant surfaces can be smooth on days where the pollutant is driven by regional sources, but will be much less smooth with high local variations on days where there are significant local sources of emission.

S3.5 Comparison between different implementations of GGP

We have implemented 3 different variants of the GGP model. Besides the main focus on Bayesian model with GGP (GM) on the latent spatial processes (implementations details in Section S2.1), we have also discussed GGP on the response process ($\text{GM}_{response}$) in Section 4.4.5 (implementation details in Section S2.2), and have presented a frequentist estimation scheme for the parameters with maximum likelihood estimation (GM_{MLE}) using co-ordinate descent (see Section S2.4). In this Section we compare the performances of the 3 variants of GGP.

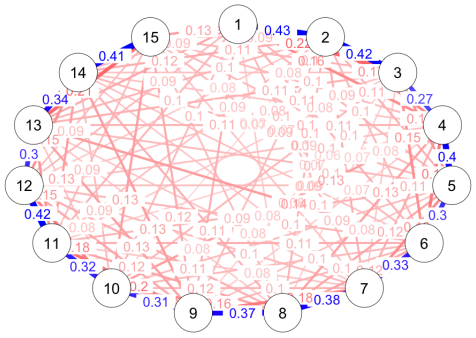
The MLE-based methods preclude misalignment among data locations for the different variables and excludes nugget processes $\epsilon_i(\cdot)$ in (4.1). Set 1B conforms to such assumptions. Hence, we compare GM with GM_{MLE} for this set. Since there is no nugget, GM and $\text{GM}_{response}$ are the same for this set. Figure S6 plots the true covariance and cross-covariance parameters and their estimates from GM and GM_{MLE} showing that the Bayesian and frequentist implementations yield similar estimates.

We then compare the two Bayesian implementations of GGP: GGP on the latent process (GM); and GGP on the response process ($GM_{response}$). Figure S7 plots the estimates of the covariance and cross-covariance parameters, and prediction RMSPE based on hold-out data for the two variants for Set 2B. We see that they produce similar estimates and predictive performance.

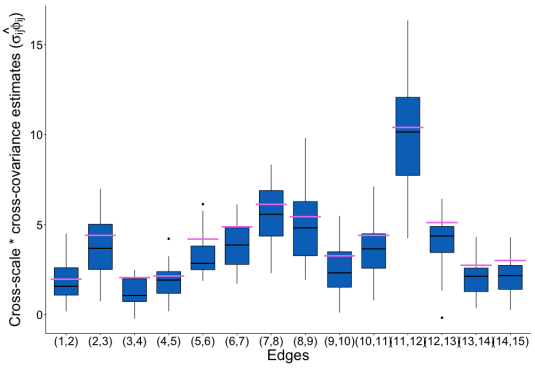
S4 Additional figures and tables

Set 1A (True = Gem graph)		Set 2A (True = Path graph)	
Edges	Probability	Edges	Probability
(3, 4)	**0.64**	**(1, 2)**	**0.43**
(1, 2)	**0.57**	**(11, 12)**	**0.42**
(1, 5)	**0.57**	**(2, 3)**	**0.42**
(2, 3)	**0.55**	**(14, 15)**	**0.41**
(2, 5)	**0.50**	**(4, 5)**	**0.40**
(3, 5)	**0.48**	**(7, 8)**	**0.38**
(4, 5)	**0.46**	**(8, 9)**	**0.37**
(1, 3)	0.45	**(13, 14)**	**0.34**
(2, 4)	0.43	**(6, 7)**	**0.33**
(1, 4)	0.42	**(10, 11)**	**0.32**
		(9, 10)	**0.31**
		(5, 6)	**0.30**
		(12, 13)	**0.30**
		(3, 4)	**0.27**
		(1, 3)	0.22
		(13, 15)	0.21
		(9, 11)	0.20
		(7, 9)	0.18
		(4, 6)	0.18
		(10, 12)	0.18

Table S1: Posterior probabilities of including an edge when estimating the graph in a GGP. The rows of the table are ordered from highest to lowest. (a) Set 1A (all edges), (b) Set 2A (edges with the top 20 highest selection probabilities). Bold numbers indicate true edges.

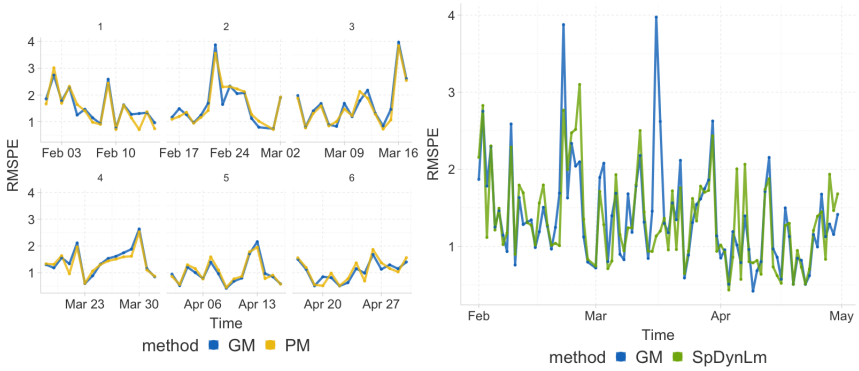


(a) Posterior edge selection probabilities for Set 2A.



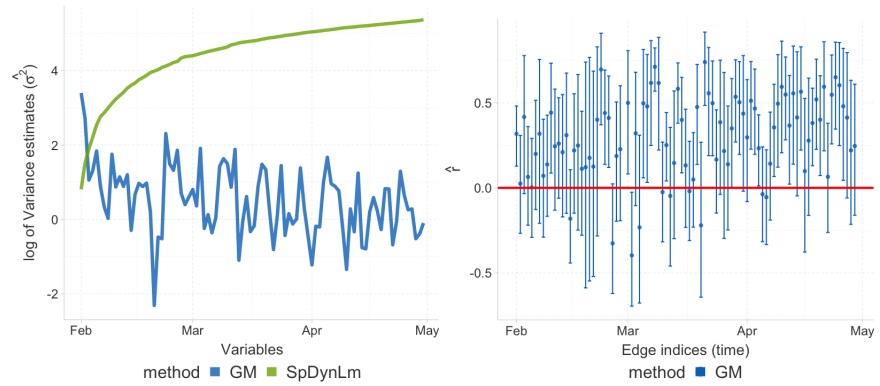
(b) Cross-covariance parameter estimates for Set 2A while estimating the unknown graph

Figure 4.6: Performance of GGP with unknown graph for Set 2A: (a): Marginal edge probabilities estimated from the reversible jump MCMC sampler. Blue edges denote the true edges and red denotes the non-existent edges. Edges are weighted proportional to the estimated posterior selection probabilities. (b) GM estimates of cross-correlation parameters (b_{ij}) corresponding to true edges when the graph is unknown, with horizontal pink lines indicating the true values.



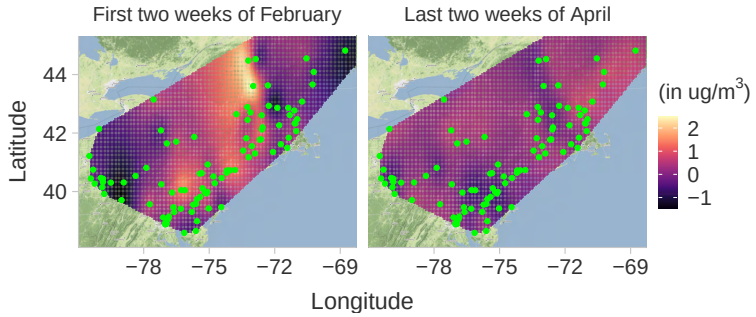
(a) Prediction performance for fortnightly analysis

(b) Prediction performance for full analysis



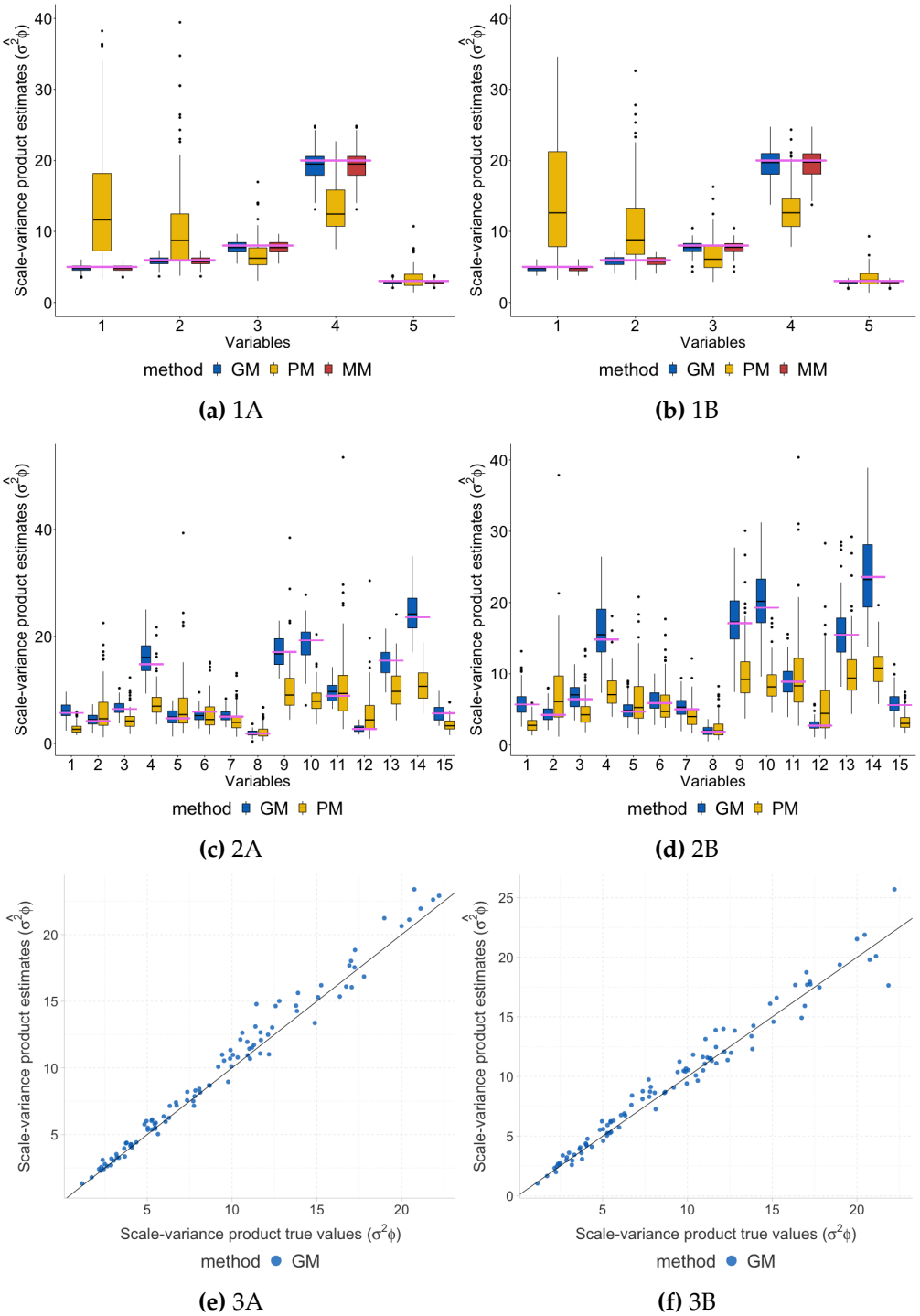
(c) Log-variance estimates for the full analysis

(d) Estimates of time-specific cross-correlations



(e) Average residual surfaces

Figure 4.7: $PM_{2.5}$ analysis: (a) Daily RMSPE for the 6 fortnightly analyses, (b) Daily RMSPE for the full analyses, (c) Estimates of the time-specific process variances, (d) Estimates and credible intervals of the cross-correlation parameters $r_{t,t-1}$ (corresponding to the cross-covariances $b_{t,t-1}$), (e) Estimates of the residual spatial processes from GM (after adjusting for covariates and baseline) for first two weeks of February and last two weeks of April.

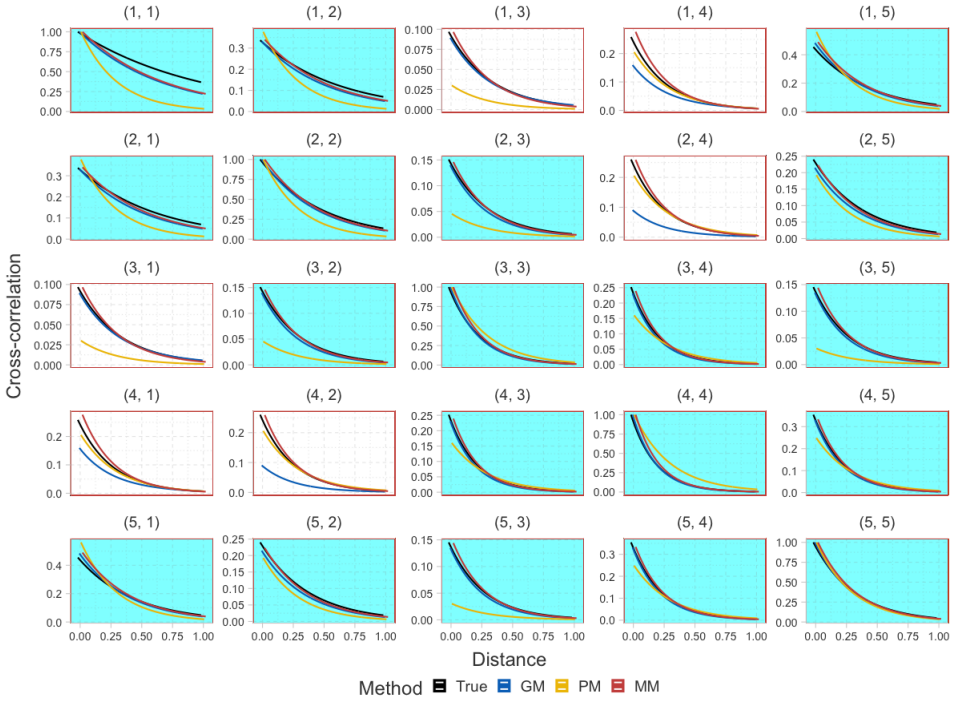


(a) 1A **(b)** 1B

(c) 2A **(d)** 2B

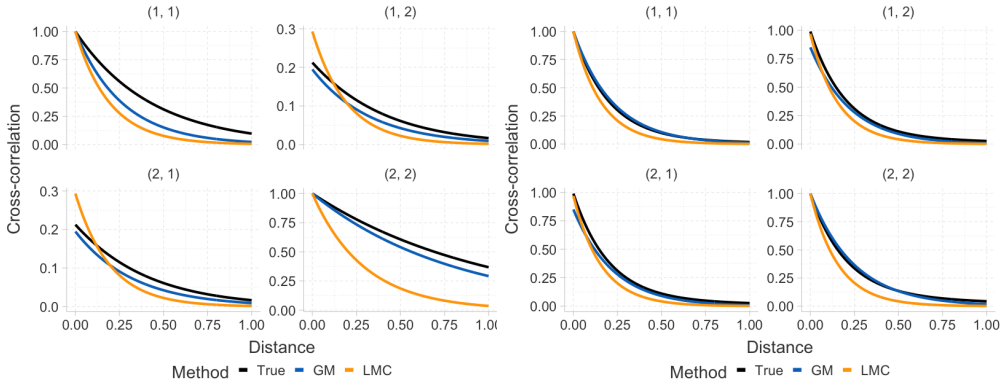
(e) 3A **(f)** 3B

Figure S1: Estimates of the marginal parameters $\sigma_{ii}\phi_{ii}$, $i \in \mathcal{V}$, for the 6 simulation settings. The horizontal pink lines in Figures (a) and (b) indicate the true parameter values.



(a) Set 1B

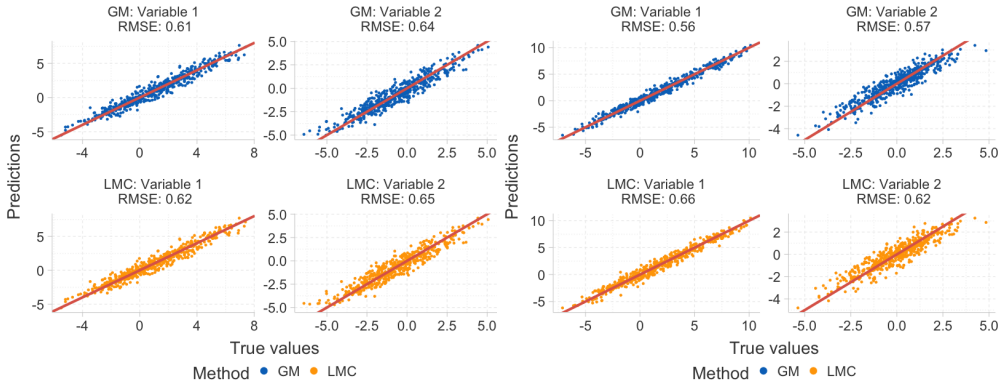
Figure S2: Estimates of cross-correlation functions (GM, PM, MM) compared to the truth in Set 1B. The grids correspond to specific pair of the cross-correlations. The sky blue shaded grids correspond to edges in the gem graph assumed for GM.



(a) True process: GM

(b) True process: linear model of coregionalization

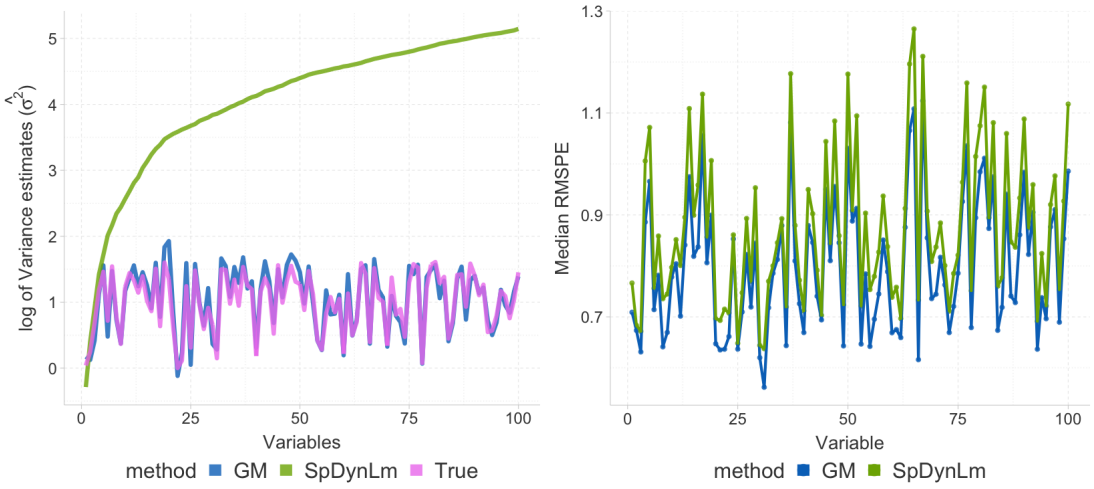
Figure S3: Estimates of cross-correlation functions for the two observed processes. The grids correspond to specific pair of the cross-correlations.



(a) True process: GM

(b) True process: linear model of coregionalization

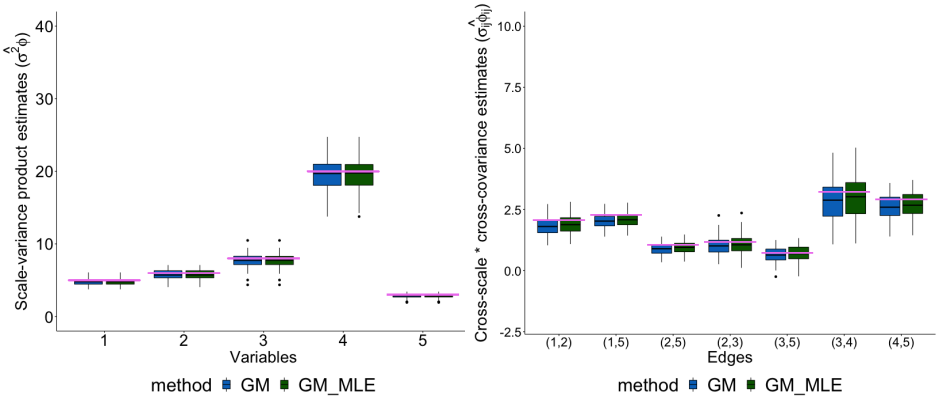
Figure S4: Truth vs prediction for test sets in different simulation scenarios with prediction RMSE reported.



(a) Variance estimates (in log-scale).

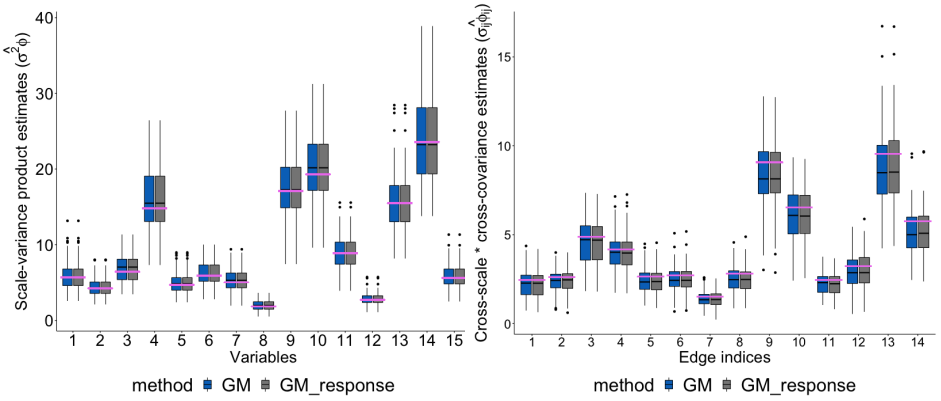
(b) Median RMSPE over seeds for each variable (time-point).

Figure S5: Comparison between GGP and SpDynLm for modelling AR(1) spatial time series: (a) Variance estimates (in log-scale) for GM and SpDynLm compared to the true values for Set 3A. (b) Median RMSPE over seeds for each variable (time) for GM and SpDynLm for Set 3A.

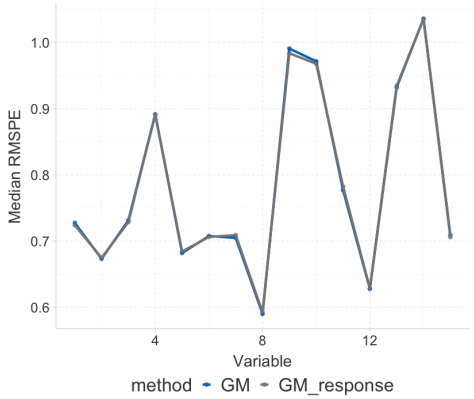


(a) Set 1B (Marginal parameters) **(b)** Set 1B (Cross-covariance parameters)

Figure S6: Comparison of performance of Graphical Matérn (GM) and Graphical Matérn frequentist (GM_{MLE}) : (a) Estimates of the scale-covariance product parameters $\sigma_{ii}\phi_{ii} = i \in \mathcal{V}$, (b) Estimates of the cross-covariance parameters $\sigma_{ij}\phi_{ij} = \Gamma(1/2)b_{ij}$, $(i, j) \in E_{\mathcal{V}}$ for Set 1B. The horizontal pink lines in Figures (a) and (b) indicate true parameter values.

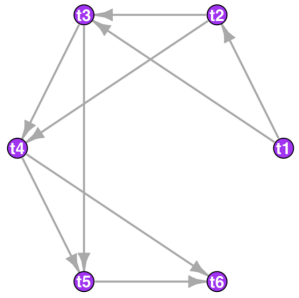


(a) Set 2B (Marginal parameters) **(b)** Set 2B (Cross-covariance parameters)

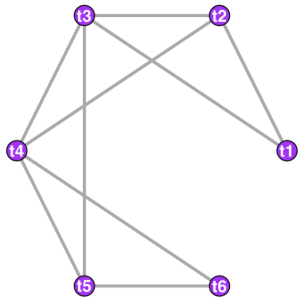


(c) Set 2B (Predictions)

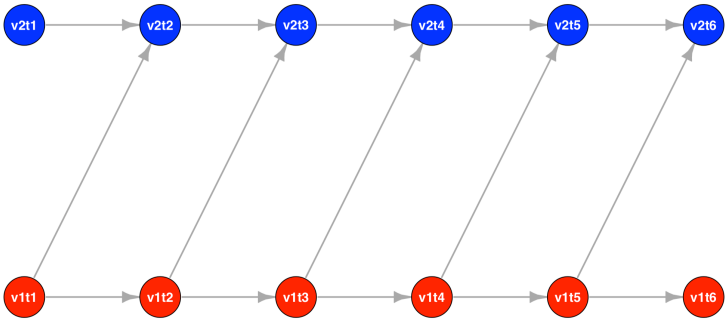
Figure S7: Comparison of performance of Graphical Matérn (GM) and Graphical Matérn response ($\text{GM}_{\text{response}}$) : (a) Estimates of the scale-covariance product parameters $\sigma_{ii}\phi_{ii}, i \in \mathcal{V}$, (b) Estimates of the cross-covariance parameters $\sigma_{ij}\phi_{ij} = \Gamma(1/2)b_{ij}$, $(i,j) \in E_{\mathcal{V}}$ and (c) median RMSPE for Set 2B. The horizontal pink lines in Figures (a) and (b) indicate true parameter values.



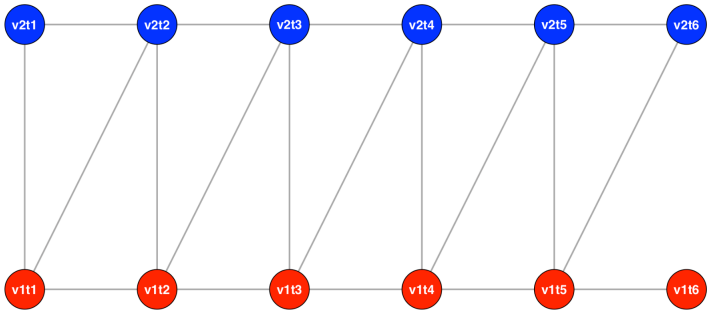
(a) DAG for a univariate AR(2) model



(b) Moralized $\mathcal{G}_{\mathcal{T}}$ for a univariate AR(2) model

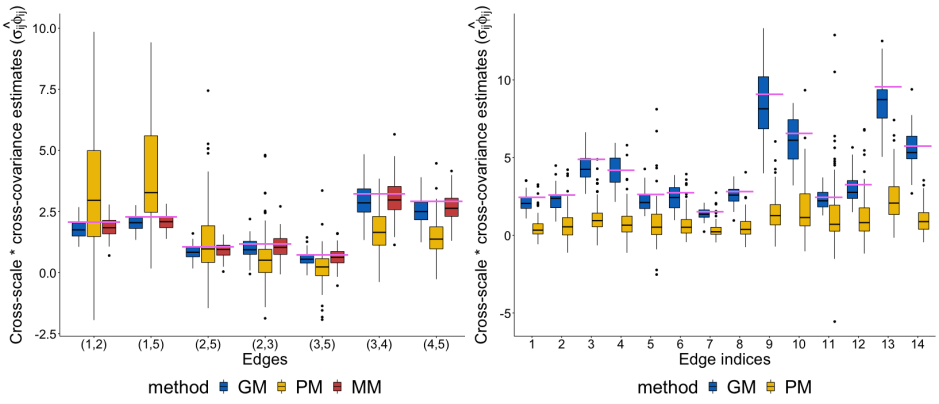


(c) DAG for the graphical VAR model example of Section 4.4.2

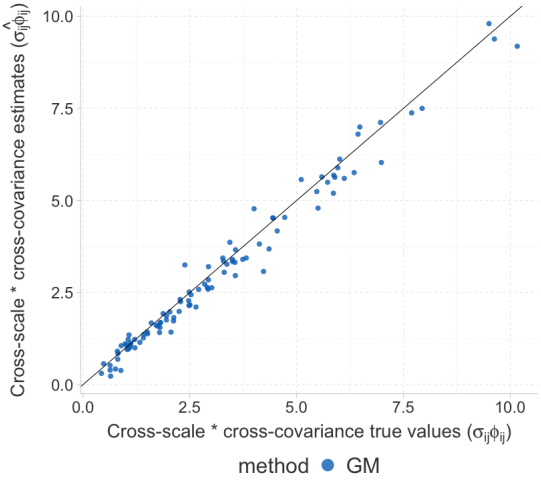


(d) Moralized $\mathcal{G}_{\mathcal{V}\times\mathcal{T}}$ for the graphical VAR model of Figure (c)

Figure S8: Graphical models for autoregressive spatial time-series.

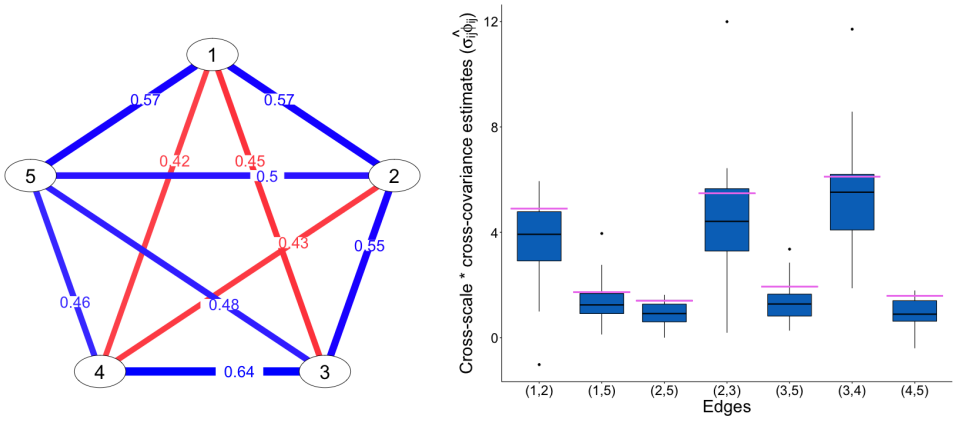


(a) Set 1A **(b)** Set 2A



(c) Set 3A

Figure S9: Estimation performance of graphical Matérn in the correctly specified case: (a), (b) and (c): Estimates of the cross-covariance parameters $\sigma_{ij}\phi_{ij} = \Gamma(1/2)b_{ij}$, $(i, j) \in E_{\mathcal{V}}$ for the 3 simulation sets (1A, 2A and 3A) where the graphical Matérn is correctly specified. The horizontal pink lines in Figures (a) and (b) indicate true parameter values.



(a) Posterior edge selection probabilities for Set 1A.

(b) Cross-covariance parameter estimates for Set 1A while estimating the unknown graph

Figure S10: Performance of GGP with unknown graph for Set 1A: (a): Marginal edge probabilities estimated from the reversible jump MCMC sampler. Blue edges denote the true edges and red denotes the non-existent edges. Edges are weighted proportional to the estimated posterior selection probabilities. (b) GM estimates of cross-correlation parameters (b_{ij}) corresponding to true edges when the graph is unknown.

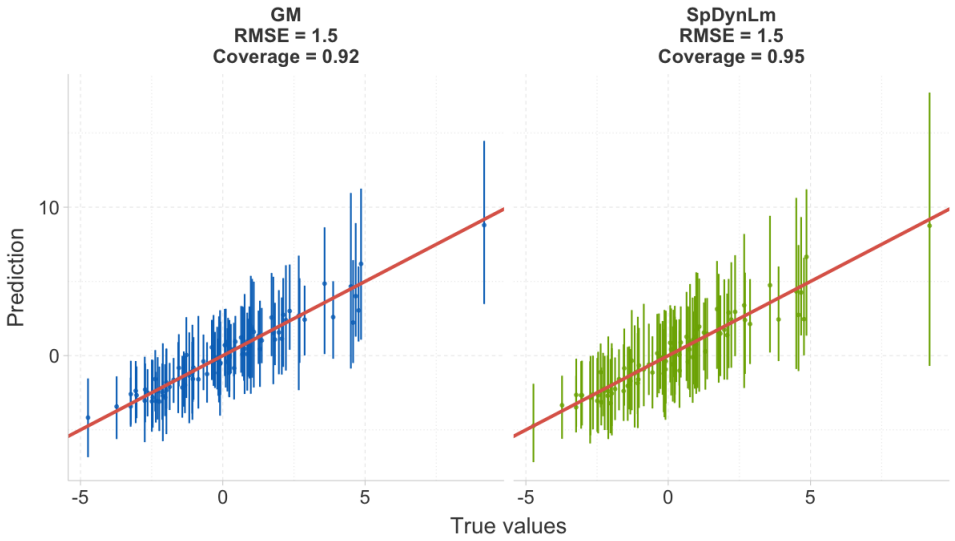


Figure S11: Truth vs prediction for test set data compared among GM and SpDynLM

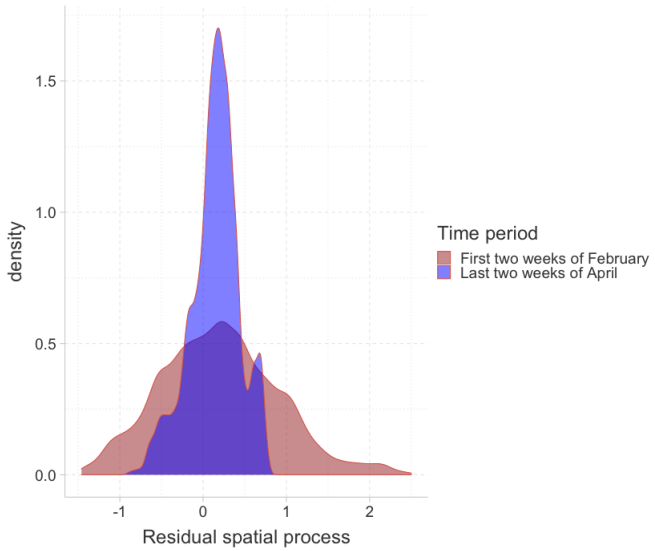


Figure S12: Density of residual spatial process values (across locations) for two different time periods - first two weeks of February and last two weeks of April

References

Wackernagel, Hans (2013). *Multivariate geostatistics: an introduction with applications*. Springer Science & Business Media.

Cressie, Noel A. C. and Christopher K. Wikle (2011). "Statistics for spatio-temporal data". In: Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley. ISBN: 978-0-471-69274-4. URL: http://opac.inria.fr/record=b1133266.

Banerjee, S., B. P. Carlin, and A. E. Gelfand (2014). "Hierarchical Modeling and Analysis for Spatial Data". In: Second. Boca Raton, FL: Chapman & Hall/CRC.

Genton, Marc G and William Kleiber (2015). "Cross-covariance functions for multivariate geostatistics". In: *Statistical Science*, pp. 147–163.

Gneiting, Tilmann, William Kleiber, and Martin Schlather (2010). "Matérn cross-covariance functions for multivariate random fields". In: *Journal of the American Statistical Association* 105.491, pp. 1167–1177.

Apanasovich, Tatiyana V, Marc G Genton, and Ying Sun (2012). "A valid Matérn class of cross-covariance functions for multivariate random fields with any number of components". In: *Journal of the American Statistical Association* 107.497, pp. 180–193.

Heaton, Matthew J, Abhirup Datta, Andrew O Finley, Reinhard Furrer, Joseph Guinness, Rajarshi Guhaniyogi, Florian Gerber, Robert B Gramacy, Dorit Hammerling, Matthias Katzfuss, et al. (2019). "A case study competition among methods for analyzing large spatial data". In: *Journal of Agricultural, Biological and Environmental Statistics* 24.3, pp. 398–425. DOI: 10.1007/s13253-018-00348-w.

Cox, David R and Nanny Wermuth (1996). *Multivariate Dependencies: Models, Analysis and Interpretation*. Chapman and Hall/CRC.

Dahlhaus, Rainer (2000). "Graphical interaction models for multivariate time series". In: *Metrika* 51.2, pp. 157–172.

Dahlhaus, Rainer and Michael Eichler (2003). "Causality and graphical models in time series analysis". In: *Oxford Statistical Science Series*, pp. 115–137.

Eichler, Michael (2008). "Testing nonparametric and semiparametric hypotheses in vector stationary processes". In: *Journal of Multivariate Analysis* 99.5, pp. 968–1009.

Eichler, Michael (2012). "Fitting graphical interaction models to multivariate time series". In: *arXiv:1206.6839*.

Dempster, Arthur P (1972). "Covariance selection". In: *Biometrics*, pp. 157–175.

Banerjee, Sudipto, Alan E Gelfand, Andrew O Finley, and Huiyan Sang (2008). "Gaussian predictive process models for large spatial data sets". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.4, pp. 825–848.

Speed, Terence P, Harri T Kiiveri, et al. (1986). "Gaussian Markov distributions over finite graphs". In: *The Annals of Statistics* 14.1, pp. 138–150.

Xu, Ping-Feng, Jianhua Guo, and Xuming He (2011). "An improved iterative proportional scaling procedure for Gaussian graphical models". In: *Journal of Computational and Graphical Statistics* 20.2, pp. 417–431.

Finley, Andrew O, Huiyan Sang, Sudipto Banerjee, and Alan E Gelfand (2009). "Improving the performance of predictive process modeling for large datasets". In: *Computational statistics & data analysis* 53.8, pp. 2873–2884.

Kleiber, William (2017). "Coherence for multivariate random fields". In: *Statistica Sinica*, pp. 1675–1697.

Datta, Abhirup, Sudipto Banerjee, Andrew O. Finley, and Alan E. Gelfand (2016). "Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets". In: *Journal of the American Statistical Association* 111.514, pp. 800–812. DOI: 10.1080/01621459.2015.1044091. eprint: http://dx.doi.org/10.1080/01621459.2015.1044091. URL: http://dx.doi.org/10.1080/01621459.2015.1044091.

Dobra, Adrian et al. (2003). "Markov bases for decomposable graphical models". In: *Bernoulli* 9.6, pp. 1093–1108.

Wang, Hao and Mike West (2009). "Bayesian analysis of matrix normal graphical models". In: *Biometrika* 96.4, pp. 821–834.

Roverato, Alberto (2002). "Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models". In: *Scandinavian Journal of Statistics* 29.3, pp. 391–411.

Atay-Kayis, Aliye and Hélène Massam (2005). "A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models". In: *Biometrika* 92.2, pp. 317–335.

Lauritzen, Steffen L (1996). *Graphical models*. Vol. 17. Clarendon Press.

Gonzalez, Joseph, Yucheng Low, Arthur Gretton, and Carlos Guestrin (2011). "Parallel Gibbs sampling: From colored fields to thin junction trees". In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 324–332.

Schmidt, Alexandra M and Alan E Gelfand (2003). "A Bayesian coregionalization approach for multivariate pollutant data". In: *Journal of Geophysical Research: Atmospheres* 108.D24.

Gelfand, Alan E, Alexandra M Schmidt, Sudipto Banerjee, and CF Sirmans (2004). "Nonstationary multivariate process modeling through spatially varying coregionalization". In: *Test* 13.2, pp. 263–312.

Lopes, Hedibert Freitas, Esther Salazar, and Dani Gamerman (2008). "Spatial Dynamic Factor Analysis". In: *Bayesian Analysis* 3(4), pp. 759 –792.

Ren, Qian and Sudipto Banerjee (2013). "Hierarchical factor models for large spatially misaligned data: A low-rank predictive process approach". In: *Biometrics* 69.1, pp. 19–30.

Taylor-Rodriguez, Daniel, Andrew O Finley, Abhirup Datta, Chad Babcock, Hans-Erik Andersen, Bruce D Cook, Douglas C Morton, and Sudipto Banerjee (2019). "Spatial factor models for high-dimensional and large spatial data: An application in forest variable mapping". In: *Statistica Sinica* 29, p. 1155.

Zhang, Lu and Sudipto Banerjee (2021). "Spatial factor modeling: A Bayesian matrix-normal approach for misaligned data". In: *Biometrics*. DOI: https://doi.org/10.1111/biom.13452. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.13452. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13452.

Cressie, Noel and Andrew Zammit-Mangion (2016). "Multivariate spatial covariance models: a conditional approach". In: *Biometrika* 103.4, pp. 915–935.

Gneiting, Tilmann (2002). "Nonseparable, stationary covariance functions for space–time data". In: *Journal of the American Statistical Association* 97.458, pp. 590–600.

Jacquier, E, NG Polson, and PE Rossi (1993). "Priors and models for multivariate stochastic volatility". In: *Unpublished manuscript, Graduate School of Business, University of Chicago*.

Jacquier, Eric, Nicholas G Polson, and Peter E Rossi (2002). "Bayesian analysis of stochastic volatility models". In: *Journal of Business & Economic Statistics* 20.1, pp. 69–87.

Jung, Alexander, Gabor Hannak, and Norbert Goertz (2015). "Graphical lasso based model selection for time series". In: *IEEE Signal Processing Letters* 22.10, pp. 1781–1785.

Green, Peter J and Alun Thomas (2013). "Sampling decomposable graphs using a Markov chain on junction trees". In: *Biometrika* 100.1, pp. 91–110.

Thomas, Alun and Peter J Green (2009). "Enumerating the junction trees of a decomposable graph". In: *Journal of Computational and Graphical Statistics* 18.4, pp. 930–940.

Barker, Richard J and William A Link (2013). "Bayesian multimodel inference by RJMCMC: A Gibbs sampling approach". In: *The American Statistician* 67.3, pp. 150–156.

Apanasovich, Tatiyana V and Marc G Genton (2010). "Cross-covariance functions for multivariate random fields based on latent dimensions". In: *Biometrika* 97.1, pp. 15–30.

Li, Bo and Hao Zhang (2011). "An approach to modeling asymmetric multivariate spatial covariance structures". In: *Journal of Multivariate Analysis* 102.10, pp. 1445–1453.

Saha, Arkajyoti and Abhirup Datta (2018). "BRISC: bootstrap for rapid inference on spatial covariances". In: *Stat* 7.1, e184.

Stroud, Jonathan R., Peter Müller, and Bruno Sansó (2001). "Dynamic models for spatiotemporal data". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.4, pp. 673–689. DOI: https://doi.org/10.1111/1467-9868.00305. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9868.00305. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00305.

Gelfand, A. E., S. Banerjee, and D. Gamerman (2005). "Spatial Process Modelling for Univariate and Multivariate Dynamic Spatial Data". In: *Environmetrics* 16, pp. 465–479.

Saha, Arkajyoti, Sumanta Basu, and Abhirup Datta (2021). "Random forests for spatially dependent data". In: *Journal of the American Statistical Association* just-accepted, pp. 1–46.

Datta, Abhirup, Sudipto Banerjee, James S Hodges, and Leiwen Gao (2019). "Spatial disease mapping using directed acyclic graph auto-regressive (DAGAR) models". In: *Bayesian analysis* 14.4, p. 1221.

Cramér, Harald (1940). "On the theory of stationary random processes". In: *Annals of Mathematics*, pp. 215–230.

Parra, Gabriel and Felipe Tobar (2017). "Spectral mixture kernels for multi-output Gaussian processes". In: *Advances in Neural Information Processing Systems*, pp. 6681–6690.

Finley, Andrew O, Sudipto Banerjee, and Alan E Gelfand (2012). "Bayesian dynamic modeling for large space-time datasets using Gaussian predictive processes". In: *Journal of geographical systems* 14.1, pp. 29–47.

Finley, Andrew O, Sudipto Banerjee, and Alan E Gelfand (2013). "spBayes for large univariate and multivariate point-referenced spatio-temporal data models". In: *arXiv:1310.8192*.

Chapter 5

Discussion and Conclusion

This dissertation provided three major methodological contributions to modeling mixed data types and spatial data from public health surveillance. In this discussion, we will focus on the limitations and possible future extensions of the proposed methods to solve broader public health problems.

Chapter 2 focused on a joint modeling of multivariate mixed data types in cross-sectional studies. We want to extend the scope of the methods to model longitudinal outcomes coming from mobile health studies. With modern technologies such as wearables and smartphones, intensive longitudinal studies emerged over the last decade. These studies collect repeated measurements of mixed type (e.g. continuous, truncated, binary, ordinal, and others) within days and over weeks. For example, we could have joint measurements on physical activity (continuous) and mood/ pain scores (ordinal) on the same subjects collected repeatedly throughout the day. Thus, we can extend SGCRM to deal with mixed type intensive longitudinal data.

Methods developed in Chapter 3 are limited to applying in a univariate classification problem with a binary outcome and a continuous predictor. We

can extend the work in several interesting ways to cover broader scenarios. Ideas from Chapter 2 can define latent R_l^2 for any mixed type outcome and mixed type predictor. Similarly, the approach can be extended to scenarios with multiple predictors. Moreover, we can compare the asymptotic efficiency of the various rank-based estimators of AUC and R_l^2.

Chapter 4 described Graphical Gaussian Process (GGP) models, which are scalable with large number of variables at every location. However, these models still have cubic complexity regarding the number of locations. Therefore, we want to extend our method to make it computationally tractable for a large number of locations. Nearest Neighbor Gaussian Processes provide an accurate and scalable solution for the problem of large number of locations in a univariate setting. Therefore, we can extend this approach to devise an algorithm, that embeds the nearest neighbor Gaussian process in our GGP framework. These Nearest Neighbor Graphical Gaussian Process (NNGGP) models can analyze big spatial data with many variables and locations. Apart from the obvious applications in environmental sciences, this new approach can significantly impact spatial transcriptomics, where we collect millions of gene expressions from cells distributed over a tissue.

As an overarching broad goal, we can combine methods proposed in this dissertation to model multivariate stochastic processes of mixed data type. One specific example can be multivariate mixed spatial data coming from ecological applications. Vegetation cover or species distribution is often reported in ordinal scales in ecology. Here, the researchers want to learn the association between various environmental factors and ecological variables

to promote conservation efforts. Similarly, the prevalence of a disease can be coded in high/low (binary) throughout different regions. The goal here would be to learn the association between disease prevalence and multiple pollutants. Here, we can propose a solution to build joint models for mixed-type multivariate spatial data.