Statistical Machine Learning for Spatial- and Network-Linked Data

Si Cheng

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

Ali Shojaie, Chair

Adam A. Szpiro, Chair

Lianne Sheppard

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Statistical Machine Learning for Spatial- and Network-Linked Data

Si Cheng

Co-Chairs of the Supervisory Committee:
Ali Shojaie
Department of Biostatistics

Adam A. Szpiro
Department of Biostatistics

Statistical machine learning techniques offer versatile tools for prediction, estimation and inference across a wide range of applications. However, the ability of existing methods to handle data with dependence induced by complex spatial or network structures is limited, despite the increasing potential of such data due to recent advances in data collection technologies. This dissertation develops statistical machine learning methodologies that are well suited for such settings and require weaker assumptions than many existing alternatives. We start our discussion with an intuitive variable importance measure for a broad class of black-box spatial prediction models in Chapter 2. We then introduce a flexible dimensional reduction algorithm for spatial data in Chapter 3, which leads to superior performance in downstream modeling tasks while preserving approximation accuracy. In Chapter 4, we propose a computationally efficient estimation and inference procedure for doubly-stochastic spatial point processes that does not rely on certain common but stringent model assumptions. In Chapter 5, we investigate estimation and inference for direct and indirect causal effects of treatments in imperfectly randomized trails, with the presence of cross-unit interference on random graphs.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

As I age, the flow of time has become less noticeable, and the past five years feels arguably shorter than a never-ending rainy season in Seattle. Amid these fleeting moments, I want to express my gratitude to all the remarkable individuals who have been by my side, offering support, guidance, and inspiration.

I would like to thank my PhD advisors, Ali Shojaie and Adam Szpiro, for consistently offering valuable input on my research, and embracing both my spontaneous and calculated thoughts, ideas, and decisions – whether academic or personal. They have helped me grow into not only a more independent researcher, but also a better communicator and collaborator. I will cherish all the conversations we had throughout these years, which made my journey in the world of statistics less stressful and more enjoyable.

I would like to extend my thanks to my committee members, Lianne Sheppard, Jon Wakefield, Forrest Crawford and Steve Mooney, for always asking insightful questions and providing fresh and unique perspectives (e.g., of esteemed epidemiologists or enthusiastic Bayesian thinker) on my research. Forrest, who was also my Master's advisor, has helped tremendously and been a role model since my earliest exploration of biostatistics. Whenever in doubt of myself or my decisions, I have always remembered, and will always remember, his encouragement that constantly motivates me to be faithful to my true self, and dedicatedly pursue my passion despite possible obstacles.

I want to also thank Magali Blanco and Tim Larson who provided helpful feedback on the scientific aspects of Chapters 2 and 3, and all members of the SLAB Lab, Sheppard Lab and Crawford Lab for the intriguing discussions we have had. I am grateful for the thoughtful guidance and support I have received from Shuangge (Steven) Ma, Haiqun Lin, Yanlin Tang, Bendong Lou and Jiachen Ye as I planned for my graduate school and career as a researcher or biostatistician in longer terms. I would like to thank Whitney Kiker and Katie Kerr for our enjoyable collaborations together, which have stimulated my interest in applying statistics to solve real-world, clinical

problems. I appreciate all the academic, administrative and computational resources provided by the UW Biostatistics Department, and particularly the support from Minh Vo, Sharon Browning, Gitana Garofalo, Ken Rice, Timothy Thornton, Lurdes Inoue and the BITE team.

My amazing peers, friends and family have made my time as PhD student pleasant and memorable, and have brought me closer to my better self throughout the years. In countless days and nights to come, I will recall the epic(?) talent show and all random and fun discussions, relevant or irrelevant to statistics, that I have had with my cohort. Video and board gaming time with my friends along with our aimless chatters will continue being a precious source of comfort, joy and energy after a long day or week of work. Some of my really talented friends made the best food and desserts that I have ever had, which I believe will continue to be impossible to beat. The humor, open-mindedness and unconditional love from my parents, along with the warm snuggles and funny pictures of my two lovely sisters (aka our cats) will, as always, endow me with the power to undertake whatever challenges waiting ahead of me.

I also feel compelled to thank the gifts of nature, and all my friends who have enjoyed and appreciated them together with me. They (the nature as well as my friends) constantly remind me of the vast world beyond my day-to-day routine, and that there is always more to explore and be excited about. I want to pay a special tribute to the beautiful and peaceful sunsets at Magnuson park and Lake Washington, which I was fortunate enough to live next to during the past five years. I spent lots of time there recharging myself, and they have witnessed and perhaps silently supported me through many of my ups and downs.

The rainy season is starting in Seattle again soon. Now somewhat overwhelmed by the never-ending sunny days in California, I wish to share the plethora of warmth and sunlight I now have with my cherished ones in Seattle and all over the world.

DEDICATION

to my parents and little sisters[1]

[1] technically, our cats

Chapter 1

INTRODUCTION

Statistical machine learning techniques have provided a wide range of versatile tools in various real-world problems, from public health and biology to social science and urban planning. Meanwhile, with recent advances in data collection technologies, practitioners are able to obtain large volumes of data that exhibit intricate structures; examples include high-dimensional geographical information systems (GIS) data (Banerjee, 2017; Goodchild and Haining, 2004), behavioral and clinical information observed from social media or electronic records (Huang et al., 2020; Sharma et al., 2020; Zafarani et al., 2014), and measurements of gene expression with spatial localization information on tissues (Andersson et al., 2021; Ståhl et al., 2016).

The availability of these data presents new opportunities along with challenges, where the intrinsic structure among observations necessitate sophisticated methodologies to uncover scientific insights. A common feature for these data is the dependence between observations induced by their underlying structure such as spatial proximity or the presence of social interactions. For instance, in air pollution modeling, nearby regions often exhibit similar pollutant concentrations despite potential variations in topography or human behavior (Bertazzon et al., 2015; Feng et al., 2019); in vaccine efficacy trials for infectious diseases, vaccinating one individual can potentially benefit others within the same neighborhood (Perez-Heydrich et al., 2014). The presence of such dependence violates some common assumptions (e.g. independent and identically distributed data) required for valid statistical modeling or inference. Consequently, failing to account for this dependence would lead to suboptimal results and limit the full potential of modern datasets.

This dissertation develops machine learning and causal inference methods tailored to data with spatial or network-induced dependence. This chapter presents an overview of the proposed methodologies as well as how they fit into the general theme of this dissertation.

Variable Importance Measure for Spatial Machine Learning Models

Before developing any specific machine learning methodology for dependent data, we first discuss how to maximize the utility of existing ones and gain a deeper understanding of the mechanisms they capture. The interpretation of black-box machine learning prediction models is gaining increasing interest in recent years, yet there are few methodologies well-suited to spatial statistical models.

We address this issue in Chapter 2 with a specific focus on air pollution modeling, though our proposal can be applied to other spatial or correlated data as well. We build our investigation on exposure assessment settings, specifically spatial prediction of air pollution measurements at unobserved locations. This is useful for many health effect studies of air pollution, since the study participants often reside in different locations than the air pollution monitoring sites.

In addition to generating accurate predictions to minimize exposure measurement error, understanding the mechanism captured by the model is another crucial aspect that may not always be straightforward due to the complex nature of machine learning methods, as well as the lack of unifying notions of variable importance. This is further complicated in air pollution modeling by the presence of spatial correlation.

We tackle these challenges in two datasets: sulfur (S) from regulatory United States national PM2.5 sub-species data and ultrafine particles (UFP) from a new Seattle-area traffic-related air pollution dataset. Our key contribution is a leave-one-out approach for variable importance that leads to interpretable and comparable measures for a broad class of models with separable mean and covariance components. We illustrate our approach with several spatial machine learning models, and it clearly highlights the difference in model mechanisms, even for those producing similar predictions. We leverage insights from this variable importance measure to assess the relative utilities of two exposure models that have similar out-of-sample prediction accuracies but appear to draw on different types of spatial information to make predictions.

Dimension Reduction for Dependent Data Balancing Prediction and Approximation Accuracy

Dimension reduction is often a preliminary step in statistical modeling or prediction of multivariate, often high dimensional, data. However, classical dimensional reduction algorithms such as principal

component analysis (PCA) are not appropriate for correlated data, since they do not account for the spatial correlation between observations. Moreover, since this preliminary step is conducted separately from the downstream modeling tasks, the lower-dimensional scores may not contain scientifically meaningful signal or offer ideal performance for subsequent analyses.

In Chapter 3, we formalize the closeness of approximation and the utility of lower-dimensional scores for downstream modeling as two complementary, sometimes conflicting, metrics for dimension reduction. We illustrate how existing methodologies fall into this framework, and propose a flexible dimension reduction algorithm tailored to dependent data that achieves the optimal trade-off. We derive a computationally simple form for our algorithm, and illustrate its performance through simulation studies, as well as two applications in air pollution modeling and spatial transcriptomics.

Computationally Efficient Inference for Doubly-Stochastic Spatial Point Processes

Chapter 4 focuses on the estimation and inference for a specific type of spatial model, namely, spatial point processes. They model the occurrence of events over a spatial domain, where both spatial effects and covariate effects contribute to the probability of events. We are interested in learning about the covariate effects in doubly-stochastic spatial point processes; i.e., inhomogeneous Poisson processes conditioned on the realization of a random intensity function. Doubly-stochastic point processes are flexible tools for capturing spatial heterogeneity and dependence in that the baseline intensity of events is random rather than deterministic.

However, implementations of doubly-stochastic spatial models are computationally demanding, often have limited theoretical guarantees, and/or rely on restrictive assumptions. In Chapter 4, we propose a penalized regression method for estimating covariate effects in doubly-stochastic point processes that is computationally efficient and does not require a parametric form or stationarity of the underlying intensity. In particular, our method is built upon a discretization of the continuously observed spatial data, and solves a mis-specified model with network smoothing/fusion penalties. Despite the mis-specification, we establish the consistency and asymptotic normality of the proposed estimator, and develop a covariance estimator that leads to a conservative statistical inference procedure. This idea reflects the intrinsic connection between the two key components of this dissertation – spatial and network data. A simulation study shows the validity of our approach under less restrictive assumptions on the data generating mechanism, and an application to Seattle

crime data demonstrates better prediction accuracy compared with existing alternatives.

Causal Inference Under Approximate Neighborhood Interference on Random Graphs

Chapter 5 investigates the estimation and inference of causal effects under complexities induced by network structures. To learn about the causal effect of a treatment, a common assumption is that the treatment assigned to one unit should not affect the outcome of other units. However, this can easily be violated in practical settings and particularly for network-linked data, where, for example, the outcome (e.g., whether or not an individual has an infectious disease) could spread between individuals based on some notion of network connectivity (e.g., acquaintance or daily interaction). We leverage advances in graphical modeling to capture the mechanism of such cross-unit interference, and extend upon existing methods by allowing the network to be a realization of a stochastic process rather than deterministic, as well as relaxing the requirements on the sparsity of networks. Moreover, instead of restricting interference to exist only between directly connected individuals, we allow it to exist for other pairs of individuals who are indirectly connected.

Our proposal in Chapter 5 is developed under the penalized regression framework, where we adopt a smoothness penalty on non-parametric nuisance effects, along with a sparsity penalty on parametric, direct and indirect treatment effects. We show the consistency for the estimated treatment effects, and provide a valid statistical inference procedure based on a decorrelated score function.

Chapter 2

VARIABLE IMPORTANCE MEASURE FOR SPATIAL MACHINE LEARNING MODELS WITH APPLICATION TO AIR POLLUTION EXPOSURE PREDICTION

2.1 Introduction

Spatial prediction models are versatile tools that provide deeper understanding of social or natural mechanisms and guide decision making in practice. Examples include crime analysis in sociology (Chainey et al., 2008; Yi et al., 2018; Zhao and Tang, 2017), nature disaster forecasting (Aggarwal et al., 1975; Arnaud et al., 2002; Bui et al., 2018; Karimzadeh et al., 2019; Parker et al., 2017), and exposure assessment in public health (Dias and Tchepel, 2018; Kibria et al., 2002; Kim et al., 2009; Monn, 2001; Xu et al., 2022). The flexibility of machine learning (ML) models make them useful in prediction tasks with potentially complicated underlying mechanisms, but approaches to handling spatial structures in such models are relatively limited, compared to the abundance of ML methods, despite their practical importance.

Du et al. (2020); Kanevski (2009); Li et al. (2011) provided reviews and discussions on the application of ML models in spatial settings. Some approaches incorporate spatial information into the features that are used in vanilla ML models (e.g. Cracknell and Reading, 2014; Hengl et al., 2015, 2018; Kovacevic et al., 2009), which are straightforward to implement but do not provide explicit information on spatial heterogeneity and/or correlation; some combine ML methods and spatial smoothing into two-step models (e.g. Bergen et al., 2013; Blanco, 2021; Chen et al., 2019; Liu et al., 2018), which are flexible but may not partition the heterogeneity attributable to the mean and covariance components in an optimal way; and joint spatial-ML modeling (e.g. Datta et al., 2016; Georganos et al., 2021; Saha et al., 2021; Wai et al., 2020), which are better-suited for spatial prediction, but may lead to more intensive computation and/or less clear theoretical properties.

Such considerations of pros and cons may guide the choice and interpretation of spatial ML models in practice. In addition to prediction accuracy and computational complexity, another cru-

cial consideration is model interpretation, such as quantifying how much each predictor contributes to the predicted outcome (e.g. Masmoudi et al., 2020; Xu et al., 2022). This has been known as a challenge in ML literature partly due to the complexity of the models themselves; furthermore, the abundance of model classes and the wide variety of application disciplines have given rise to diverse and often incomparable measures of variable importance among different models, as noted by Greenwell et al. (2018); Wei et al. (2015); Williamson et al. (2021). Hooker et al. (2021) summarized and described common challenges for different variable importance measures, most of which are intended for non-spatial models. A common class of approaches is based on permuting the values of a covariate and assessing the change in prediction accuracy, as introduced by Breiman (2001). A related permutation-based approach was proposed by Fisher et al. (2019), which considered averaging over all possible permutations. As noted by Strobl et al. (2007, 2008) and Nicodemus et al. (2010), results from these approaches may be questionable when features are correlated, as is often the case for spatial features which commonly come from GIS (geographic information system) data. Friedman (2001); Goldstein et al. (2015) proposed using as a measure of variable importance the average predictions when all entries, or each entry, of a covariate take(s) a specific value. Another type of approach focuses on the change in predictive performance after re-fitting the model with the covariate of interest permuted, removed, or substituted (Candes et al., 2018; Lei et al., 2018; Mentch and Hooker, 2016).

Generalizing the approaches described above to spatial settings is non-trivial. First, for approaches based on permuting or manually setting covariate values and evaluating on the model of interest, manipulating the covariates affects the predicted mean component of the given model, and consequently, the residuals are altered and may no longer be reasonably fitted by the previously-trained covariance model. In addition, the presence of spatial correlation will affect error estimates based on random sample splitting and in turn the validity of permutation-based approaches using out-of-bag observations (Meyer et al., 2019; Ruß and Brenning, 2010). Furthermore, even when a valid variable importance measure can be presented as the change in prediction accuracy, interpretation from such approaches is limited since the exact quantitative contribution of each predictor on the outcome is still unclear.

Recognizing these challenges, we propose a leave-one-out approach based on quantile-level contrasts for variable importance in spatial ML models. This approach assesses the difference in

predicted values for each data point when each predictor is fixed at different quantiles of its distribution. Without refitting the whole model (which is often computationally intensive), the covariance component may not properly account for the change in the predictor values. Our proposed leave-one-out approach is flexible and efficient in that it examines each location individually for the spatial covariance component only, without requiring refitting of the computationally demanding mean model. It provides clear interpretation on how the change in each predictor, between different levels of its distribution, is associated with the difference in the outcome. This variable importance measure can be applied to a wide range of spatial machine learning models with separable mean and variance components, including multi-stage and joint models, and thus provides a standardized comparison between different modeling approaches.

The chapter is organized as follows. We introduce two air pollution datasets in Section 2.2 with a focus on exposure assessment via spatial prediction. In particular, we argue that clearly different models could lead to highly similar patterns in predicted air pollution maps, and therefore additional information such as variable importance is crucial to comprehensive understanding and selection of models. To this end, we introduce a broad class of spatial ML models under a unifying framework in Section 2.3.1, and demonstrate how the models involved in Section 2.2 fit into this framework as concrete examples. Under this modeling framework, we introduce our leave-one-out variable importance measure in Section 2.3.2. Section 2.4 illustrates the proposed approach on a synthetic dataset generated by a known mechanism, and presents variable importance analyses on the previously introduced air pollution studies. We finish our discussion with some concluding remarks in Section 2.5.

2.2 Data Description

In this section, we introduce two air pollution datasets with air pollutant concentrations that are contained within a small and large geographic region, respectively. Each dataset includes annual average concentrations of 5 and 4 air pollutants, along with measurements of 835 and 599 covariates, respectively. We build spatial prediction models for the concentration of each pollutant, and further seek to investigate the contribution of the covariates in each model.

2.2.1 Seattle Mobile Monitoring Data

This study focused on characterizing annual average traffic-related air pollution (TRAP) levels in the greater Seattle area (Blanco et al., 2022b), and leveraged a mobile monitoring (MM) campaign where a vehicle equipped with air monitors repeatedly collected two-minute samples at 309 stationary roadside sites. Approximately 29 measurements were collected from each site during all seasons, times of the week (weekdays, weekends), and most times of the day (5AM to 11PM) between March 2019 and March 2020. Prior work showed that this design provided unbiased annual average pollutant estimates (Blanco et al., 2022a).

Measured pollutants included ultrafine particles (UFP), black carbon (BC), nitrogen dioxide (NO_2), carbon dioxide (CO_2) and fine particulate matter ($PM_{2.5}$). Median 2-minute visit concentrations were trimmed at the site level such that concentrations below the 5^{th} and above the 95^{th} quantile for a given site site were removed. This was done to reduce the influence of large outlier concentrations prior to calculating annual average site concentrations. Figure 2.1 visualizes the resulting annual average UFP concentrations[1]. Annual average concentration for other pollutants are presented in Appendix A.1. We focus our discussion on UFP in the main text, although the results for other pollutants lead to similar conclusions and are included in Appendix A.1. Pollutant concentrations were log-transformed prior to model-fitting, and transformed back to the original scale to assess the R^2 (and variable importance in Section 2.4).

2.2.2 National $PM_{2.5}$ Sub-Species Monitoring Data

This dataset consists of measurements of four $PM_{2.5}$ sub-species, elemental carbon (EC), organic carbon (OC), silicon (Si), and sulfur (S) across the United States, and was collected during 2009 – 2010 by two U.S. Environmental Protection Agency networks: the Interagency Monitoring for Protected Visual Environments (IMPROVE) and Chemical Speciation Network (CSN) (Sacks et al., 2009). Following the approaches in Bergen et al. (2013) and Wai et al. (2020), we only included in our analyses measurements from the monitors that had at least 10 data points per quarter and a maximum of 45 days between measurements. We calculated annual average concentrations of S and

[1]If not noted otherwise, all maps in this chapter were made using the `ggmap` R package (Kahle and Wickham, 2013). Map tiles by Stamen Design, under CC BY 3.0; data by OpenStreetMap, under ODbL.

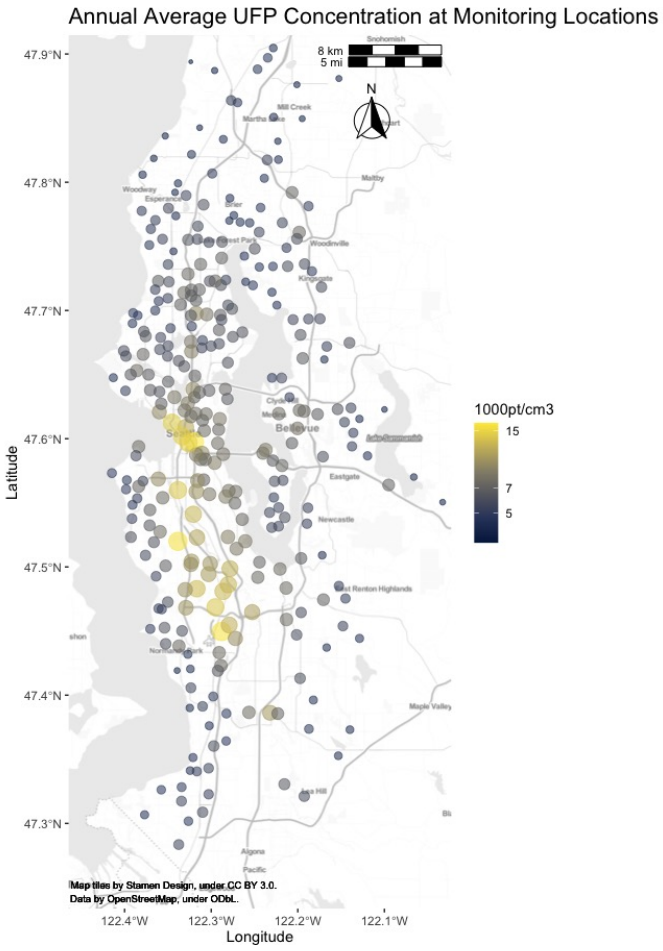


Figure 2.1: (Estimated) annual average concentration of UFP from Seattle mobile monitoring data. The color and size of dots both reflect the magnitude of concentration.

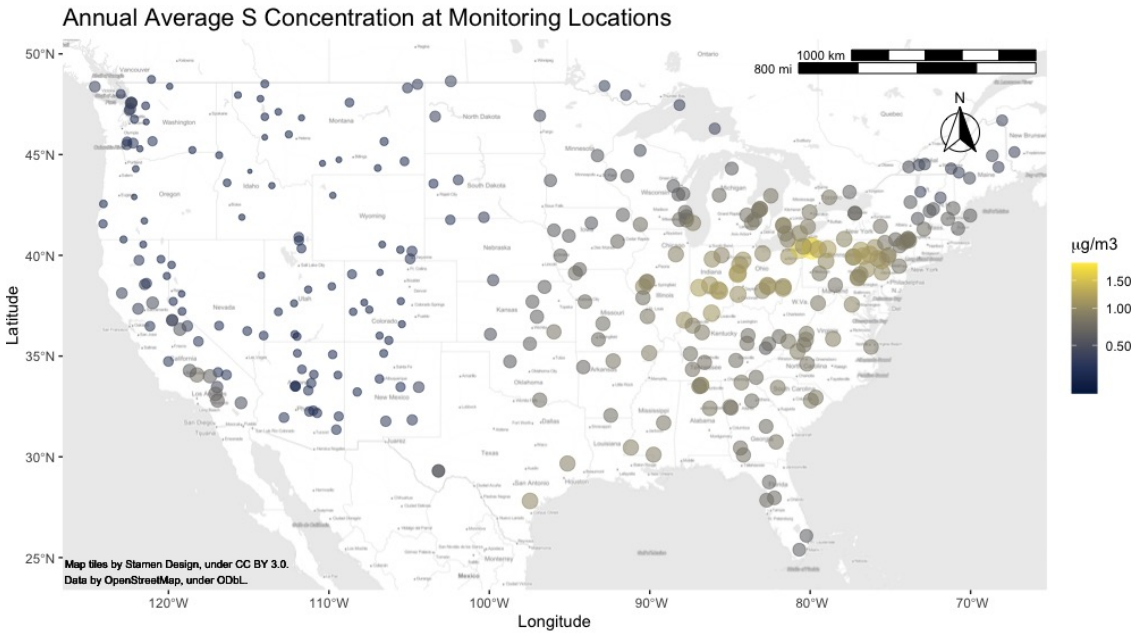


Figure 2.2: (Estimated) annual average concentration of S from national $PM_{2.5}$ monitoring data. The color and size of dots both reflect the magnitude of concentration.

Si from from 323 IMPROVE and CSN monitors over 01/01/2009 – 12/31/2009. For EC and OC, we averaged measurements from 204 IMPROVE and a subset of CSN monitors over 01/01/2009 – 12/31/2009 and from the remaining 51 CSN monitors over 05/01/2009 – 04/30/2010. The later averaging period was used for some of the CSN monitors due to a change in the measurement protocol. See Bergen et al. (2013) for additional details. We focus our discussion on the modeling of S, for which the annual average concentration is plotted in Figure 2.2. Annual averages and results for EC, OC and Si are presented in Appendix A.1. For consistency and comparability to previous analyses of the same data (Bergen et al., 2013; Wai et al., 2020), we square-root transformed the annual averages before modeling, and then transformed them back to the natural scale before presenting results.

2.2.3 Geographic Covariates

Information on 835 geographical covariates for the Seattle TRAP data, and 599 covariates for the national $PM_{2.5}$ data was available from the MESA Air (Multi-Ethnic Study of Atherosclerosis and Air Pollution) Database (MESA Air, 2019a) at all monitor locations within each dataset. Table 2.1 presents the types, details and sources of geographical information on these covariates.

We pre-processed the covariates as described in Keller et al. (2015). Specifically, geocovariates that lacked variability (less than 20% of the data were different from the most common value) or had too many outliers ($> 2\%$ of the sample size) were excluded; proportion land use variables were excluded if the maximum proportion observed in the dataset was less than 10%. This led to a total of 183 and over 480 (482 covariates for EC and OC, and 489 covariates for S and Si, where the difference is because of slightly different monitoring locations for different pollutants) geocovariates for the analysis of the Seattle TRAP and national $PM_{2.5}$ datasets, respectively.

2.2.4 Predicting Pollutant Concentration

As a motivation for our investigation of variable importance, we first discuss the prediction of pollutant concentrations on the Seattle and national air pollution datasets. We briefly describe two main prediction approaches below, and then introduce them rigorously under a unifying framework in Section 2.3.1. We conducted 10-fold cross-validation and characterized the performance of these models via R^2.

- UK-PLS: a two-step procedure that first extracts the top partial least squares (PLS) (Sampson et al., 2011; Wold et al., 1984) components from the covariates (where the number of components is determined by cross-validation within the training set, with the goal of maximizing prediction R^2), and then fits a universal kriging model via maximum likelihood with the selected components as covariates and an exponential covariance structure;

- SpatRF (PL): a spatial random forest algorithm proposed by Wai et al. (2020), where the tree-building algorithm selects each split of the tree adjusted for spatial correlation via thin plate regression splines (TPRS). We adopted the pseudo-likelihood (PL) optimization approach

introduced in Wai et al. (2020), and selected the hyperparameters by grid search via cross-validation.

For the prediction of UFP concentration with the Seattle data, UK-PLS and spatial RF (PL) achieve cross-validated R^2's of 0.81 and 0.78, respectively. Figure 2.3 displays the cross-validated prediction errors of UK-PLS and spatial RF for UFP at all monitoring locations in this study. For the national data where we predicted Sulfur concentration, the R^2 of UK-PLS and spatial RF (PL) are 0.89 and 0.90, respectively; the cross-validated prediction errors are displayed in Figure 2.4.

For the Seattle data, we observe highly similar spatial patterns in the distribution of prediction errors across the monitoring locations produced by UK-PLS and spatial RF, despite their clearly different nature: UK-PLS captures a linear trend in the mean model while spatial RF allows for non-linear effects; UK-PLS is a two-step procedure with explicit dimension reduction followed by spatial smoothing, while spatial RF conducts implicit degree-of-freedom control and jointly accounts for the mean and covariance components. On the other hand, the gridded prediction maps over the Seattle TRAP study region is shown in Figure 2.5, which is based on the evaluation of each model at an additional set of 2815 gridded locations (with higher resolution) within the same study region. The difference map on the third panel reveals that predictions made by UK-PLS and Spatial RF which are highly similar at the mobile monitoring locations, when extrapolated to a higher resolution, could still exhibit different spatial patterns. On the predicted concentrations of S, we see that while both models achieved similar accuracy and produced similar predictions for locations in mid- and western US, their different behaviors in eastern US were reflected by the larger (positive) prediction errors of UK-PLS at a few locations.

All these observations indicate that different spatial prediction models may appear to be highly similar when restricted to certain areas, while the true underlying difference between them may not be observed merely based on their predictions, if evaluations at an additional set of locations (e.g. the gridded locations in addition to the mobile monitoring locations in the Seattle data, or eastern US comparing to mid- and western US in the national data) were unavailable. It would therefore be desirable and also necessary to understand and compare different models by investigating the mechanisms that they capture, beyond just their prediction performance on the training data. Developing a universal and easily interpretable variable importance measure for a diverse class of

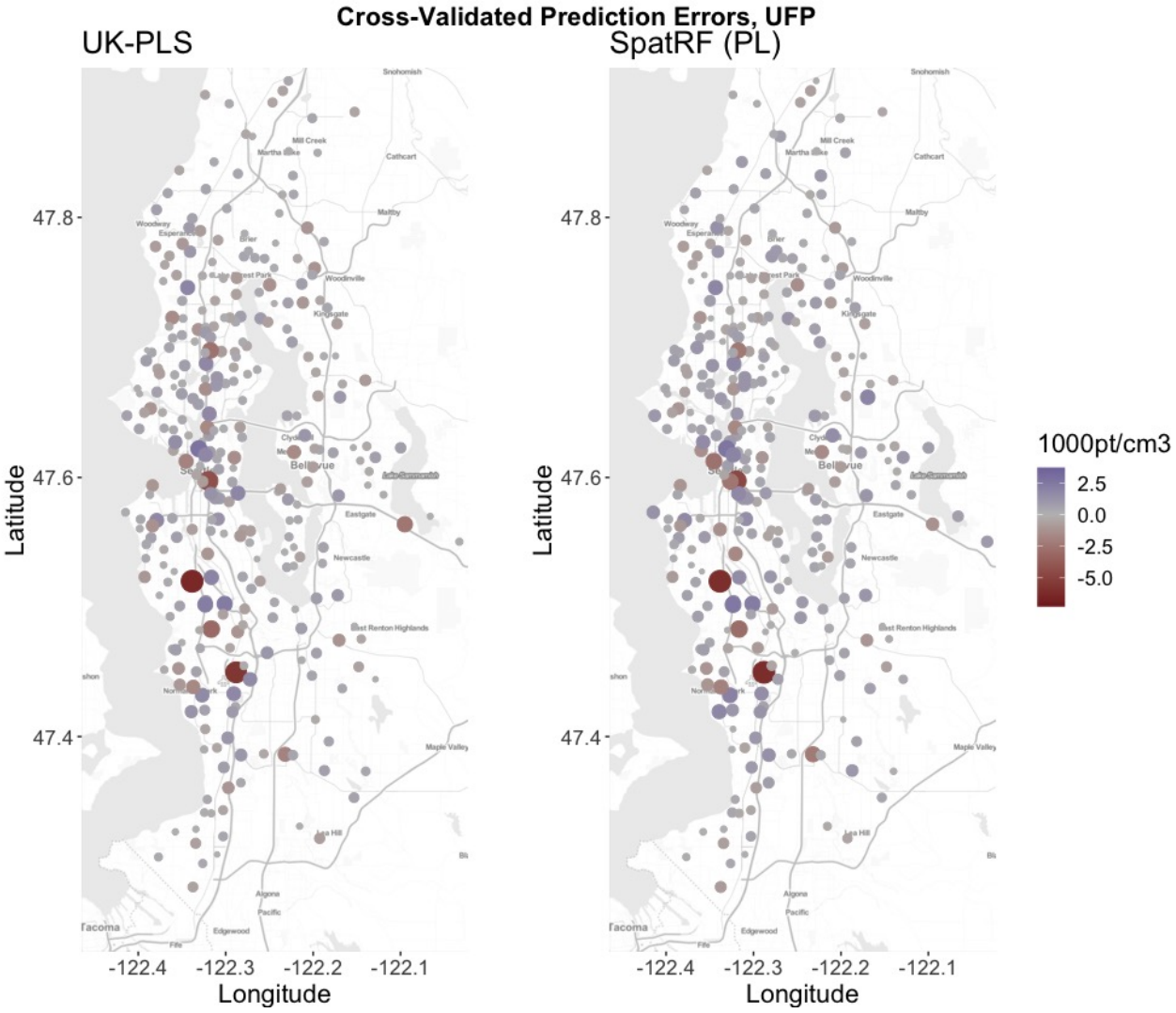


Figure 2.3: Cross-validated prediction errors of UFP for UK-PLS and spatial RF (PL) at each monitoring location of the Seattle study. The shade of color and size of dots both reflect the magnitude of errors.

prediction methods is a key step to facilitate this, and further to aid the selection and interpretation of models.

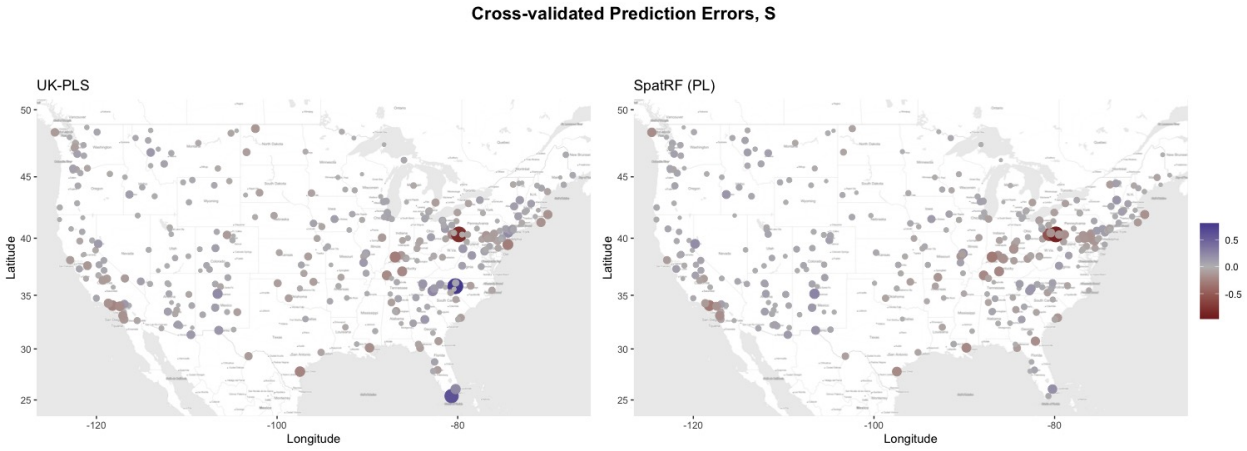


Figure 2.4: Cross-validated prediction errors of S for UK-PLS and spatial RF (PL) at each monitoring location of the national study. The shade of color and size of dots both reflect the magnitude of errors.

2.3 Variable Importance Measure for Spatial ML Models

2.3.1 Spatial Prediction: Setup

Before introducing the proposed variable importance measure, we describe a broad class of spatial prediction models to which such measure is applicable. Consider a class of models where an outcome $Y(s)$, indexed by location $s \in \Omega$, is modeled via an additive mean surface taking the form of

$$g(\mu(s)) = \sum_{k=1}^{K} [f_k(X(s)) + \nu_k(s)] \tag{2.1}$$

where $g(\cdot)$ is a link function, $X(s)$ represents the covariates, each $\nu_k(s)$ represents the correlated error term, and each $f_k(\cdot)$ is an unknown function within some function class $\mathcal{F}$. The indexing k allows for application to ensemble learning methods. As an example, when g is the logarithm link function and $\nu(s)$ is a correlated Gaussian process, $\mu(s)$ is the underlying intensity of a doubly-stochastic Poisson process, also known as the Cox process (Brémaud, 1981; Cox, 1955; Serfozo, 1972); when g is the identity link and $\nu(s)$ is a correlated Gaussian process, $\mu(s)$ models the surface of a continuous outcome, e.g. a universal kriging model if f is linear.

A spatial ML model often learns about each f_k under some assumed restrictions on $\mathcal{F}$, and

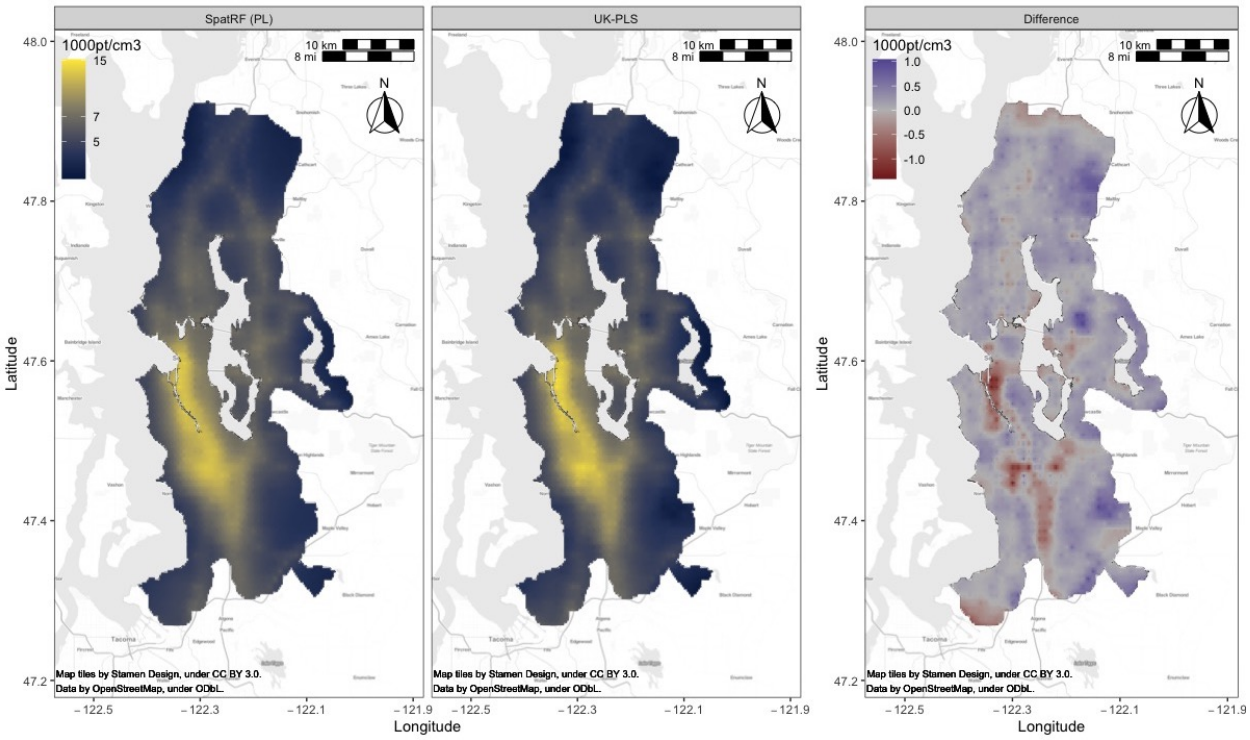


Figure 2.5: Predicted UFP concentration surfaces based on predictions at gridded locations via UK-PLS and Spatial RF (PL) in the Seattle TRAP study region, along with a difference map between them (UK-PLS being the subtrahend)

$\nu_k(s)$ under parametric or structural assumptions. Once a model (2.1) is fitted at a set of training sites s_{trn}, predictions at the test sites s_{tst} can be made via

$$\hat{Y}(s_{\text{tst}}) = g^{-1}\left(\sum_{k=1}^{K}\left\{\hat{f}_k(X(s_{\text{tst}})) + \mathbb{E}_{\hat{\nu}}\left[\nu_k(s_{\text{tst}}) \mid Y(s_{\text{trn}}), \hat{f}_k(s_{\text{trn}})\right]\right\}\right),$$

where the conditional expectation $\mathbb{E}_{\hat{\nu}}\left[\nu_k(s_{\text{tst}}) \mid Y(s_{\text{trn}}), \hat{f}_k(s_{\text{trn}})\right]$ represents the smoothing of residuals via the fitted covariance model $\hat{\nu}_k(s)$ at the training sites. When g is identity and each ν_k is assumed to be a correlated Gaussian random field with covariance $\Sigma(\theta)$, for example, the model predicts

$$\hat{\nu}_k(s_{\text{tst}}) := \mathbb{E}_{\hat{\nu}}\left[\nu_k(s_{\text{tst}}) \mid Y(s_{\text{trn}}), \hat{f}_k(s_{\text{trn}})\right] = \Sigma(\hat{\theta})_{\text{tst,trn}}\left[\Sigma(\hat{\theta})_{\text{trn,trn}}\right]^{-1}\left(Y(s_{\text{trn}}) - \hat{f}_k(s_{\text{trn}})\right).$$

As an illustration and for concreteness of our following discussions, we revisit two the models

introduced in Section 2.2.4 for an observed continuous outcome $Y_{n\times 1}$, with identity $g(\cdot)$ and potentially high-dimensional covariates $X_{n\times p}$, and describe how they fit into this framework. One model is UK-PLS, which has $k = 1$ and first conducts PLS and extracts the first l $(1 \leq l \leq p)$ components of X by finding a decomposition of X

$$T_{n\times l} := X_{n\times p} H_{p\times l}$$

such that the covariance between T and Y is maximized. The number l of components to use can be selected via cross-validation. In the second step, a universal kriging model

$$\begin{aligned} Y_{n\times 1} &= T_{n\times l}\beta_{l\times 1} + \nu_{n\times 1} \\ \nu_{n\times 1} &\sim \text{Normal}(0, \Sigma(\theta)) \end{aligned}$$

where θ are covariance parameters (e.g. the nugget, partial sill and range, see Cressie (2015)) which can be solved jointly with β via maximum likelihood. With $\hat{H}_{p\times l}, \hat{\beta}, \hat{\theta}$ estimated from the model, for m new (test) locations with covariate values $X^*_{m\times p}$, the outcome Y^* can be predicted as

$$\hat{Y}^* = X^*\hat{H}\hat{\beta} + \mathbb{E}_\theta[\nu^* \mid \nu] = X^*\hat{H}\hat{\beta} + \tilde{\Sigma}_{12}(\hat{\theta})\tilde{\Sigma}_{22}^{-1}(\hat{\theta})\left(Y - T\hat{\beta}\right)$$

where $\tilde{\Sigma}_{(m+n)\times(m+n)}(\hat{\theta})$ is the covariance matrix induced by the distances between all training and test locations, partitioned as

$$\tilde{\Sigma}(\hat{\theta}) = \begin{bmatrix} \tilde{\Sigma}_{11}^{(m\times m)} & \tilde{\Sigma}_{12}^{(m\times n)} \\ \tilde{\Sigma}_{21}^{(n\times m)} & \tilde{\Sigma}_{22}^{(n\times n)} \end{bmatrix} \tag{2.2}$$

based on the training and test indices.

The second example is the spatial random forest algorithm proposed by Wai et al. (2020) solved via pseudo-likelihood (SpatRF-PL). It is an ensemble model (i.e. $k > 1$) which specifies

$$\hat{\mu}(s) := \sum_{k=1}^{K} \hat{\mu}_k(s) := \sum_{k=1}^{K} \left[\hat{f}_k(X(s)) + \hat{\nu}_k(s)\right] \tag{2.3}$$

where each f_k is a regression tree, and each ν_k could be modeled via common spatial smoothing methods such as kriging or regression splines (Friedman, 1991; Wood, 2003). With a kriging model, for each k, a spatially adjusted tree can be built by solving the optimization problem resulting from

profile likelihood, assuming normally distributed spatial error terms:

$$\arg\max_{\theta_k}\left[-\frac{1}{2}\log|\Sigma(\theta_k)|-\frac{1}{2}\left(Y-\hat{f}_k(X\mid\Sigma(\theta_k))\right)^\top\Sigma^{-1}(\theta_k)\left(Y-\hat{f}_k(X\mid\Sigma(\theta_k))\right)\right]$$
$$\text{s.t. }\hat{f}_k(X\mid\Sigma(\theta_k))=\arg\min_{f_k(X|\Sigma(\theta_k))}\left(Y-f_k(X\mid\Sigma(\theta_k))\right)^\top\Sigma^{-1}(\theta_k)\left(Y-f_k(X\mid\Sigma(\theta_k))\right). \tag{2.4}$$

And likewise, predictions at test locations can be made via

$$\hat{Y}^*=\sum_{k=1}^{K}\left[\hat{f}_k(X^*)+\mathbb{E}_{\hat{\theta}_k}\left(\nu^*\mid\nu\right)\right]=\sum_{k=1}^{K}\left[\hat{f}_k(X^*)+\tilde{\Sigma}_{12}(\hat{\theta}_k)\tilde{\Sigma}_{22}^{-1}(\hat{\theta}_k)\left(Y-\hat{f}_k(X)\right)\right]$$

with $\tilde{\Sigma}$ defined identically as (2.2).

2.3.2 Leave-One-Out Evaluation of Quantile-Level Contrasts

We now introduce a variable importance measure that is applicable to additive models taking the form of (2.1) as described in Section 2.3.1. This leave-one-out approach is based on the change in predictions across different user-specified quantiles $q_1,\ldots,q_m$ for each covariate, evaluating at each location $s_1,\ldots,s_n$ one at a time. Recall that prediction at the test locations s_{tst} relies on evaluation of the trained covariance model $\hat{\nu}(s)$ via $\hat{\nu}(s_{\text{tst}})=\mathbb{E}[\nu(s_{\text{tst}})\mid\nu(s_{\text{trn}})=Y(s_{\text{trn}})-\hat{f}(s_{\text{trn}})]$. This implies that when we permute or fix the values of covariates X_{tst} as in many common variable importance analyses, the evaluation of the covariance model $\hat{\nu}(s)$ is also implicitly altered, and furthermore, the distribution of residuals $\nu(s_{\text{tst}})$ at the test locations may no longer be well-fitted by $\hat{\nu}(s)$. Therefore, the key idea of the proposed approach is to reuse the trained mean model across all locations, but re-fit the covariance model in a leave-one-out manner. We write $\hat{F}_{X_j}(x_j), j=1,\ldots,p$ as the empirical cumulative distribution function (CDF) of the jth covariate, and $\boldsymbol{s}_{-i}$ as the set of all locations except the ith one.

Suppose we have trained a model

$$g\left(\hat{\mu}(s)\right)=\sum_{i=1}^{K}\left[\hat{f}_k(X(s))+\hat{\nu}_k(s)\right]$$

from observations $\{(X(s_i),Y(s_i))\}_{i=1}^n$. Then for the jth covariate, at the lth quantile q_l of interest and within the kth sub-model $\hat{f}_k$, we replace each $X_j(s_i)$ with the sample q_l-quantile and calculate

the predicted mean as

$$\hat{\zeta}_k^{j,l}(s_i) := \hat{f}_k\left(X_1(s_i), \ldots, \hat{F}_{X_j}^{-1}(q_l), \ldots, X_p(s_i)\right)$$

for location i. In plain words, this is the new predicted mean at s_i with the jth covariate replaced by its q_l-th quantile across $s_1, \ldots, s_n$. Next, we re-fit the kth error component with the new predicted means $\hat{\zeta}_k^{j,l}(\boldsymbol{s}_{-i})$ along with observations $(X(\boldsymbol{s}_{-i}), Y(\boldsymbol{s}_{-i}))$, leaving out the ith site. Denoting the resulting model as $\hat{\nu}_{(-i),k}^{j,l}(s)$, the leave-one-out approach yields the linear predictor

$$\hat{\eta}_k^{j,l}(s_i) := \hat{\zeta}_k^{j,l}(s_i) + \mathbb{E}_{\hat{\nu}_{(-i),k}^{j,l}}\left[\nu_{(-i),k}^{j,l}(s_i) \mid Y(\boldsymbol{s}_{-i}), \hat{f}_k\left(X(\boldsymbol{s}_{-i})\right)\right] \tag{2.5}$$

for location i, which is what the model would predict if the jth covariate of all data points were replaced by the q_l-quantile of its distribution, and if the error component were fitted without the ith data point, while keeping everything else intact. Re-fitting leads to updated covariance model(s) that account for the implicit change in the error distribution caused by manipulating the covariates.

Re-doing this for all i, we obtain a sequence of linear predictors of the form (2.5). Aggregating across each sub-model (each k) and location (each i) finally leads to the averaged leave-one-out predictions at the q_l-th quantile for covariate j:

$$\bar{\mu}_{j,l} := g^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\hat{\eta}_k^{j,l}(s_i)\right). \tag{2.6}$$

For each j, the trajectory $\bar{\mu}_{j,1}, \ldots, \bar{\mu}_{j,m}$ reflects how the predictions, on average, vary across different quantiles of covariate X_j, which serves as an intuitive measure of the contribution of this covariate on the predicted outcome. This procedure could easily be parallelized to facilitate computation. Algorithm 1 presents the described procedure in detail.

2.4 Variable Importance Analyses

2.4.1 Illustration with Synthetic Data

We first illustrate how the proposed variable importance measure performs in comparison to the true mechanism, by presenting a variable importance analysis with synthetic data generated with the same covariates as the national study. We generated a continuous outcome with five active predictors, including distance to A1 road, population density, annual median NDVI (buffer size

1km), land use: mixed urban (buffer size 15km) and land use: residential (buffer size 15km), which were all scaled and centered, except population density and mixed urban land use which were first scaled and then shifted to be non-negative. The mean model was given by

$$\begin{aligned}\mathbb{E}(Y \mid X) = &-0.5 \times \text{distance to A1 road} + 0.2 \times \text{population density}^2 - 1 \times \text{annual median NDVI} \\ &+ 0.5 \times \sqrt{\text{mixed urban land use}} + 0.5 \times \text{residential land use} \\ &- 0.25 \times \text{distance to A1 road} \times \text{annual median NDVI}.\end{aligned}$$

The correlated error term was given by an exponential model with scale and range parameters equal to 4 and 2.5 respectively, where the unit of distance was 1000km. The uncorrelated errors were generated from a standard Gaussian distribution. With this setup, the variances of each component (mean, partial sill and nugget) in the outcome were 2.21, 1.07 and 1.02 respectively, so that we roughly have a 2:1:1 variance decomposition of the outcome. The distribution of this synthetic outcome and variability coming from each component are visualized in Figure A.3 in Appendix A.2.

We trained both UK-PLS and spatial RF to predict this synthetic outcome with all covariates. Although there is an interaction term in the true mechanism, only main effects were included when training each model. UK-PLS and spatial RF achieve cross-validated R^2 0.62 and 0.72 respectively.

Figure 2.6 reflects the fact that both UK-PLS and spatial RF allocate the contribution of the true predictors onto others that are highly correlated with them, and conversely, only the predictors that are highly correlated with the true ones were found to have a meaningful contribution in the model. This plot also suggests that the greedy tree-based algorithm tends to favor a more parsimonious model when autocorrelation is present among the predictors, since spatial RF assigns a close-to-zero contribution to most predictors, and only a few are assigned to have high importance. This aligns with our knowledge that if one of the autocorrelated features is selected into a regression tree model, the remaining ones are less likely to further improve model accuracy and thus less likely to enter the model and be identified as important predictors. The full variable importance plot leads to the same observations, and is presented in Appendix A.2 for completeness.

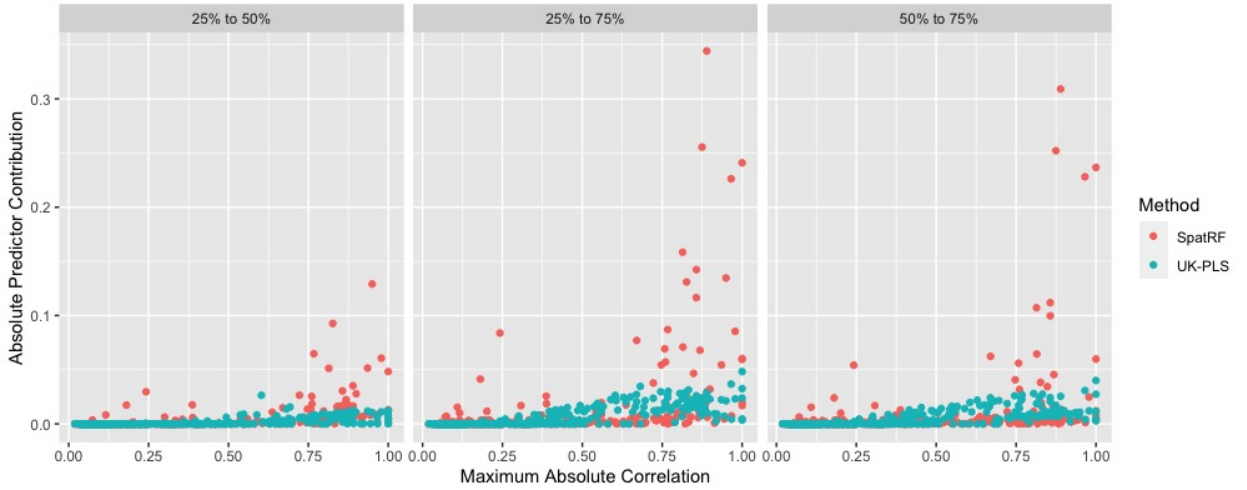


Figure 2.6: Distribution of variable importance versus maximum absolute correlation with any truly active predictors in the synthetic data. The x-axis is the maximum absolute correlation between each predictor across all five active predictors.

2.4.2 Results for Seattle and National Data

We examine the proposed variable importance measure at each quartile of each predictor, i.e. q_1, q_2 and q_3 are $0.25, 0.5$ and 0.75 respectively. We compare UK-PLS and spatial RF models and look at three contrasts, $\bar{\mu}_{j,2} - \bar{\mu}_{j,1}$, $\bar{\mu}_{j,3} - \bar{\mu}_{j,2}$ and $\bar{\mu}_{j,3} - \bar{\mu}_{j,1}$, to evaluate the contribution of each predictor.

Figure 2.7 visualizes the contribution of predictors having the greatest importance in predicted UFP concentration with the Seattle data. Despite similar predicted maps between UK-PLS and spatial RF, the plots highlight the difference in the mechanisms captured by each model. In particular, spatial RF identifies the length of truck routes and closeness to major roads as major contributors to predicted UFP concentration in the Seattle TRAP study, while UK-PLS highlights the distance to large airport as a more significant contributor. As known from prior studies, jet engine exhaust is a significant source of UFP (see e.g. Hudda et al., 2018), which suggests UK-PLS as a more sensible candidate in terms of scientific interpretation. In addition, the UK-PLS model is more consistently influenced by truck traffic and general traffic on large A1 roads than the spatial RF predictions. It is also reasonable that a linear model would perform well in a relatively homogeneous and small area where relationships between sources and pollution levels are consistent

across the domain.

This variable importance measure also reveals the greedy nature of tree building algorithms here. For instance, although both UK-PLS and spatial RF find the length of truck routes within several buffer sizes to be important predictors of UFP concentration, spatial RF highlights only a few of these autocorrelated predictors in contrast to UK-PLS, which highlights all of them, as can be seen in Figure 2.7. Further, the covariate(s) and size of the buffer highlighted varies across quantile contrasts with spatial RF, whereas this is more consistent across quantile contrasts for UK-PLS. This is related to the non-linear property of spatial RF (in contrast to the linear UK-PLS model), namely, the magnitude of effects of the same covariate could differ at different levels of its distribution.

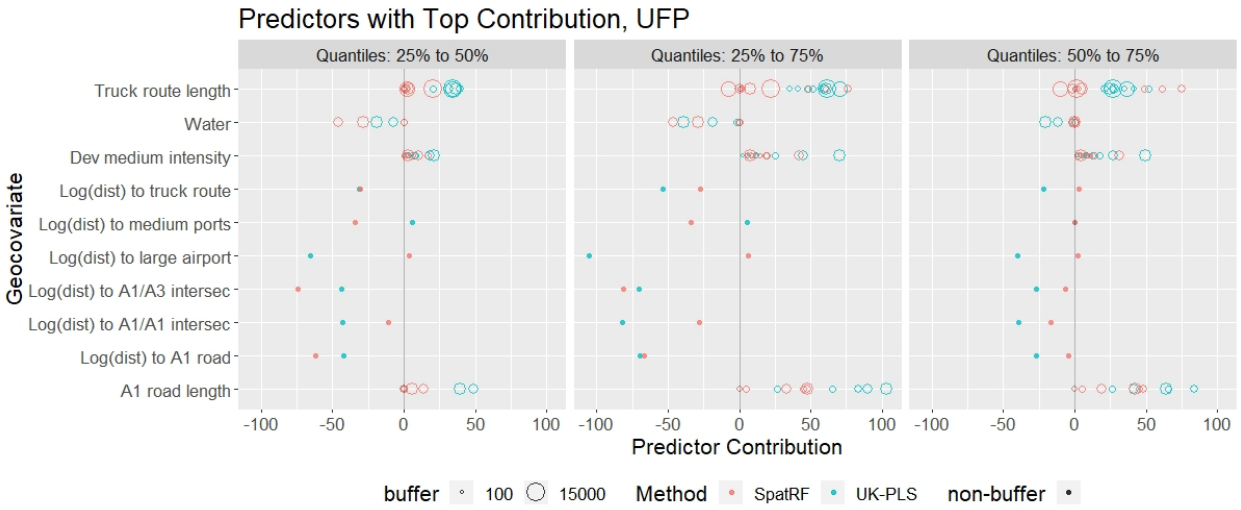


Figure 2.7: Variable importance plot for the prediction of UFP concentration, showing predictors with top 5 contribution for either method for at least one contrast. All buffer sizes are included if one of them is within the top 5 important predictors.

The proposed variable importance measure could also provide insight on how the performance of different fitted models would differ on newly observed data points. Recall that for the Seattle TRAP study, we evaluated the trained UK-PLS and Spatial RF (PL) models on an additional set of 2815 gridded locations (with higher resolution within the same study region) that were not used to train the prediction models. Figure 2.8 shows how the difference in predicted values between

models vary with a set of predictors, for which UK-PLS and spatial RF had the most different variable importance measures as given by the original training data. The analysis based on the original training data finds that spatial RF differs from UK-PLS by -4.11 units when looking at the contribution of NDVI (buffer size 5km) changing from its 25% and 50% quantile, and differs by +41.02 units when looking at the change from 50% to 75% quantiles. Therefore, it is expected that at the higher end in the distribution of NDVI we would observe predictions from a spatial RF model would grow faster, or decline slower, compared with UK-PLS, which is indeed reflected in the plot on the bottom right panel. Similar interpretations can be observed in the trend of evergreen forest land (buffer size 3km), highly-developed land use (buffer size 5km) and truck route length (buffer size 10km, albeit less noticeable), among others. There are also signs of greater variability in the difference between predictions when the distance to A1/A3 intersections, A1/A1 intersections or airports is large, which is consistent with the fact that these predictors were found to play different roles in the two models. Figure A.9 in the Appendix presents a similar comparison, visualizing the differences in predictions at the residential locations of an epidemiological cohort throughout the greater Seattle area, as opposed to the gridded locations in Figure 2.8. We observe that the difference between UK-PLS and spatial RF is less evident at the cohort locations, which are more spatially aligned with and better represented by the monitoring locations, in contrast to the gridded locations which cover less populated areas as well. This further validates our argument that models with similar behaviors on the training data (e.g., the monitoring sites) could have meaningful differences when extrapolated to new locations (e.g., the gridded locations); and with the aid of our variable importance measure, the latter can be anticipated and captured by a variable importance analysis on just the training data.

In contrast to the relatively homogeneous and small area of Seattle, our analyses on the national data demonstrate the use of variable importance measure when greater spatial heterogeneity in the distribution of pollutant concentrations is present. Figure 2.9 reveals that the predictions from spatial RF were greatly driven by proximity to A1 roads and a range of land use features, especially the amount of industrial, agricultural and forest lands within certain buffer sizes. UK-PLS identified a similar set of influential predictors, but all with smaller and more consistent magnitudes across different buffer sizes. While UK-PLS and spatial RF achieved similar prediction accuracy (R^2's of 0.89 and 0.90 respectively), the extreme influence of a few buffer sizes in the spatial RF model raises

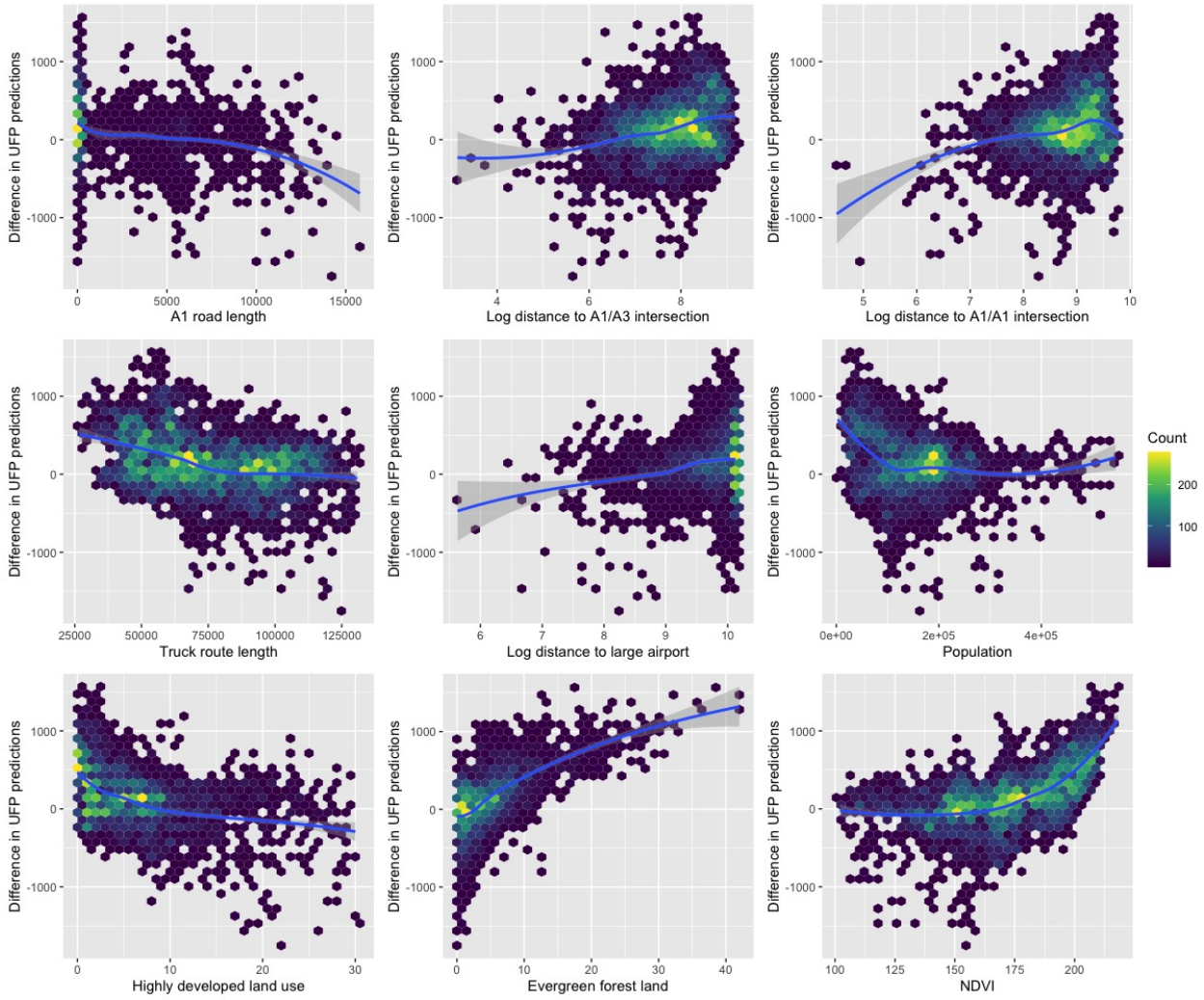


Figure 2.8: Hexagonal bin plot showing the difference between spatial RF (PL) and UK-PLS (the subtrahend) predictions of UFP concentration on gridded locations, versus the distribution of predictors with the greatest difference in variable importance between models. The color reflects the number of points falling to each small region of the plot. Locally weighted scatterplot smoothing (LOESS) curves are added to show the overall trend.

concerns about generating extreme predicted values (potentially at new, unobserved locations), and also brings challenges to the scientific interpretation.

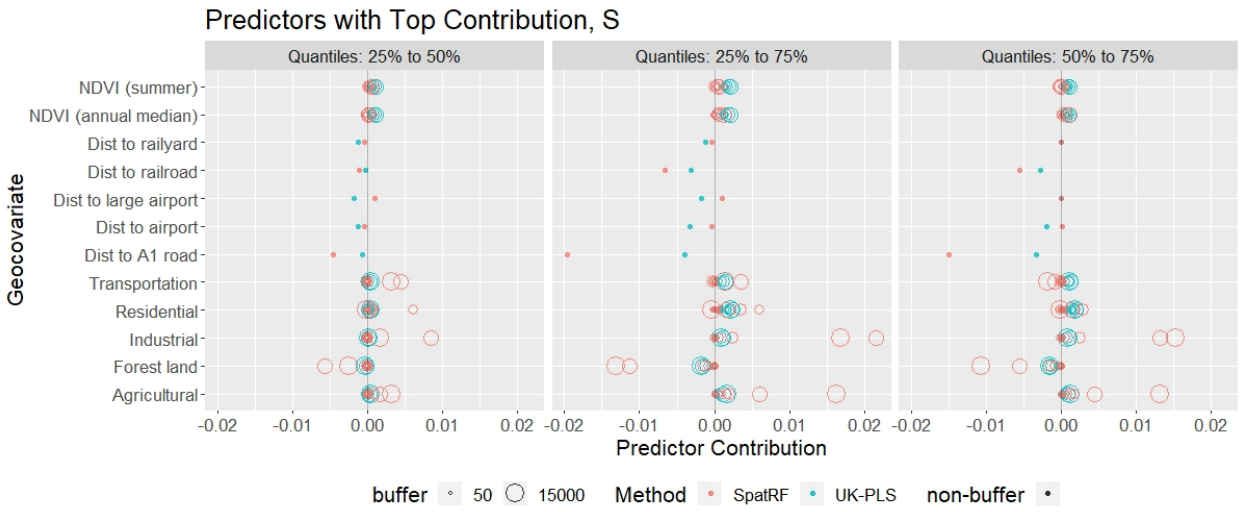


Figure 2.9: Variable importance plot for the prediction of S concentration, showing predictors with top five contributions for either method for at least one contrast. All buffer sizes are included if one of them is within the top five important predictors.

2.5 Discussion

Our investigation starts with the assessment of air pollution exposures in two real-world studies covering a small and large geographic region, respectively. We apply two machine learning algorithms – universal kriging with partial least squares and a spatial random forest – to predict exposure throughout each region. We compare the results of these two approaches, both in terms of standard prediction model performance summaries and a new variable importance metric proposed in this chapter. We found in the small geographic region setting, that although the two models had comparable prediction performances, the variable importance metric indicated that some geographic features, specifically distance to large airport, distance to main roads, and length of truck routes, were differently important contributors to the predictions of UFP concentration from UK-PLS and spatial RF. When a larger geographic region is of interest, we observed UK-PLS and spatial RF both identifying distances to main roads and the amount of different land use as contributors to the prediction of Sulfur concentration, while a few buffer sizes for land use features were assigned extreme influence by spatial RF. Our analysis of the synthetic data showed that the importance assigned by the proposed measure on each predictor was positively associated with its

correlation with the truly active predictors, where spatial RF favored a more parsimonious model with larger magnitude of contributions from each predictor comparing to UK-PLS. Given that epidemiologic cohort studies rely on predicted air pollution exposures for making inference about health effects, use of this variable importance metric has the potential to allow new insights into how seemingly similar exposure metrics may lead to different inferences.

A primary motivation for our work is to improve exposure assessment for air pollution cohort studies where the pollution exposure surface is predicted from a spatial model trained on monitoring data. Recent developments in sensor technology, monitoring study design, and statistical modeling methods have made it possible to construct accurate exposure models for a variety of pollutants at both local and national scales, as illustrated by the data we analyzed in this chapter. This state of affairs introduces a new set of challenges since many choices need to made in designing a particular exposure model, and in some cases, there are already multiple published models to choose from with overlapping spatial and temporal domains. For example, at least three research groups have developed models for $PM_{2.5}$ that cover large portions of the United States (Di et al., 2016; Kirwa et al., 2021; Yanosky et al., 2009).

A typical strategy is to select the model with the smallest out-of-sample prediction error (or highest R^2) as a way of minimizing exposure measurement error. This is generally a sound strategy, although it is now known that the model with the highest R^2 does not always lead to the best health effect inference (Szpiro et al., 2011), in part owing to the complexity of balancing different types of measurement error and interactions between measurement error and covariates in the health model (Bergen et al., 2016; Cefalu and Dominici, 2014; Szpiro and Paciorek, 2013). Given this context, it makes sense to utilize variable importance as we have developed it here as an additional tool to decide between models, giving primacy to those models that are more interpretable in terms of what is known about sources and dispersion of the air pollutant being modeled. An open question that we will consider in future research is how to balance prediction accuracy and variable importance in selecting an exposure model, e.g., when would there be a large enough difference in model performance across modeling approaches that would lead us to focus almost exclusively or entirely on model performance and not put any weight on variable importance?

In some air pollution epidemiology studies, the specific pollutant used as the exposure is regarded as a marker for source-specific pollution. For example, many studies have utilized elemental

carbon (EC), black carbon (BC), oxides of nitrogen (NO, NO_2), and fine particulate matter ($PM_{2.5}$) as markers of traffic-related pollution (TRAP). The strength with which findings about these pollutants implicate traffic as a pollutant source depends on how much of a role traffic played in the exposure model. In a recent overview of health effects of TRAP on a wide variety of health outcomes, systematic but ad-hoc methods were used to determine which exposure models could be regarded as sufficiently traffic-specific (Boogaard et al., 2022), and the selection process would have benefited from availability of a variable importance metric like ours that quantifies the contribution of traffic-related covariates to the predicted concentrations.

The variable importance measure we present is flexible, intuitive, and generally applicable to machine learning models that account for spatial correlation. This leave-one-out approach can be applied to additive models with separable mean and correlation components, including nonlinear, ensemble and/or doubly stochastic spatial models. It provides a unifying notion of variable importance which would otherwise be less comparable between different modeling approaches, and we have demonstrated that meaningful differences in the model structure could be found even for models producing similar predictions. An informative variable importance measure as ours also facilities deeper understanding of complex prediction models in the methodological aspect: our Seattle and national data examples illustrate the greedy nature of tree building algorithms which is already well-known; but such information would not be straightforward to obtain otherwise for more complex black-box models.

Our approach is an example of extrinsic variable importance measure, which is intimately tied to the specific prediction model, see e.g. Breiman (2001); Strobl et al. (2007); on the contrary, intrinsic variable importance is model-agnostic and corresponds to the best possible model (which is often unknown) within a certain class (Lei et al., 2018; Van der Laan, 2006; Williamson et al., 2021). Both types of variable importance measures are meaningful depending on the practical use cases, and extrinsic metrics are useful especially for the interpretation and selection of models.

One interesting extension of our approach would be to estimate the uncertainty of variable importance measures. As a simple and naive solution, sample splitting such as cross-validation or bootstrap can both provide an uncertainty estimate, though more careful treatment is needed with the existence of spatial correlation. Also, as is the case for many variable importance measures, the proposed approach reflects the association between predictors and the outcome captured by a

given model, rather than causal effects. And consequently, when autocorrelation between predictors is present and only a few of them are truly contributing to the outcome, it could be challenging to disentangle them. However, our analysis of synthetic data indicates that only the predictors that are highly correlated with the truly active ones are likely to be identified as important by the proposed measure.

Covariate Category	Data Source	Buffer Sizes	Notes
Airports	National Emissions Inventory Database	–	Distances to airports and large airports
Coastline	TeleAtlas	–	Distance to coastline
Railroads	TeleAtlas	–	Distance to railroads
Railyards	TeleAtlas	–	Distance to railyards
Roads	TeleAtlas	–	Distances to A1, A2 and A3 roads
		50m – 5km	Lengths of A1, A2 and A3 roads within a buffer
Intersections*	TeleAtlas	–	Distances to A1/A2/A3 intersections
		500m, 1km, 3km	Number of A1/A2/A3 intersections within a buffer
Population	US Census Bureau	500m – 15km	Population within a buffer
Land use	MRLC 2006 National Landcover Dataset & USGS historical source	50m – 15km	Land use (e.g. commercial, residential, urban, cropland, mixed forest, streams, beaches) within a buffer
	USGS historical source	–	Distance to commercial and services land use
Ports	National Geospatial Intelligence Agency	–	Distances to small, medium, large ports
Emission Sources	National Emissions Inventory Database	3km, 15km, 30km	Sum of major emissions NOx, SO_2, $PM_{2.5}$, CO and PM_{10} within a buffer
Truck routes*	Bureau of Transportation Statistics	–	Distance to truck routes
		50m – 15km	Length of truck routes within a buffer
Impervious Surface*	National Landcover Dataset	50m – 5km	Percentage of an area covered with an impervious surface (e.g. pavement, concrete) within a buffer
Elevation*	National Elevation Dataset	–	Elevation above sea level
		1km, 5km	Relative elevation: counts of points within a buffer that is less/more than 20m/50m uphill/downhill of the location
Normalized difference vegetation index (NDVI)	University of Maryland	250m – 10km	Measures the level of vegetation in a monitor's vicinity; summarized at: the 25th/50th/75th percentiles annually; median of summer (Apr to Sept) and winter (Jan to Mar and Oct to Dec)

Table 2.1: Summary of available geographical information. Distances to spatial features were truncated at 25km in the Seattle TRAP data, and at 10km in the national data. All these geocovariates were available for the Seattle mobile monitoring locations, while those marked with asterisks (*) were not available at the IMPROVE and CSN monitoring locations, and thus not included in the national study.

Algorithm 1: Leave-one-out variable importance

Input: data $(X_{n\times p}, Y_{n\times 1})$, quantile levels $q_1, \ldots, q_m \in [0,1]$ of interest, and a trained model

$$g\left(\hat{\mu}(s)\right) = \sum_{k=1}^{K}\left[\hat{f}_k(X(s)) + \hat{\nu}_k(s)\right]$$

for $j = 1, \ldots, p$ **do**

  for $l = 1, \ldots, m$ **do**

    for $k = 1, \ldots, K$ **do**

      for $i' = 1, \ldots, n$ **do**

        Replace the i'th observation of the jth covariate, $X_{i'j}$, with the q_l-th sample quantile $\hat{F}^{-1}_{X_j}(q_l)$:

$$\tilde{X}_{i'\cdot} := \left(X_1(s_{i'}), \ldots, \hat{F}^{-1}_{X_j}(q_l), \ldots, X_p(s_{i'})\right)$$

        Calculate the new predicted mean as $\hat{\zeta}^{j,l}_k(s'_i) := \hat{f}_k(\tilde{X}_{i'\cdot})$

      end

      for $i = 1, \ldots, n$ **do**

        Re-fit a covariance model on $(X_{-i,\cdot}, Y_{-i})$ with the updated residuals $Y(\boldsymbol{s}_{-i}) - \hat{\zeta}^{j,l}_k(\boldsymbol{s}_{-i})$, denoted as $\hat{\nu}^{j,l}_{(-i),k}(s)$

        Evaluate the covariance term for location i from the re-fitted covariance model $\hat{\nu}^{j,l}_k(s_i) := \mathbb{E}_{\hat{\nu}^{j,l}_{(-i),k}}\left[\nu^{j,l}_{(-i),k}(s_i) \mid Y(\boldsymbol{s}_{-i}), \hat{\zeta}^{j,l}_k(\boldsymbol{s}_{-i})\right]$

        Calculate the linear predictor for location i as $\hat{\eta}^{j,l}_k(s_i) := \hat{\zeta}^{j,l}_k(s_i) + \hat{\nu}^{j,l}_k(s_i)$

      end

    end

    Calculate the averaged leave-one-out predictions $\bar{\mu}_{j,l} := g^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\hat{\eta}^{j,l}_k(s_i)\right)$

  end

end

Result: Output averaged leave-one-out predictions $\bar{\mu}_{j,l}$ for each $j = 1, \ldots, p$, and $l = 1, \ldots, m$.

Chapter 3

PRINCIPAL COMPONENT ANALYSIS BALANCING PREDICTION AND APPROXIMATION ACCURACY FOR SPATIAL DATA

3.1 Introduction

Statistical modeling of multivariate data is a common task in many areas of research. Environmental epidemiologists often seek to learn the relationship between some health outcome and a complex mixture of multiple air pollutants, where the health effects of such mixture could vary depending on its chemical composition rather than the overall quantity alone (Achilleos et al., 2017; Dai et al., 2014; Goldberg, 2007; Lippmann et al., 2013). Spatial transcriptomics is another area involving the analysis of multivariate, and most often, high-dimensional measurements, where researchers analyze gene expressions on tissues with spatial localization information (Hu et al., 2021; Zhao et al., 2021a).

Despite the different domain areas, there are several common challenges for such tasks: First, the (potential) high dimensionality and strong autocorrelation of the multivariate measurements would require additional data handling, such as dimension reduction, as a preliminary step; see e.g., Dominici et al. (2003) for epidemiological and Kiselev et al. (2017); Sun et al. (2019) for genomics studies. Also, the spatial characteristics of the measurements should be accounted for in dimension reduction, which is not always guaranteed by existing methods.

The second complexity is that dimension reduction is often executed independently from the subsequent modeling steps, which would cause the processed data uninterpretable, or of sub-optimal quality, for downstream analyses. One example is the study of health effects of air pollution, where there is often spatial misalignment between the air pollution monitoring sites and locations where health outcomes are available (Bergen et al., 2013; Chan et al., 2015; Özkaynak et al., 2013). Therefore, after lower-dimensional scores are obtained, they still need to be extrapolated to the latter locations; but these lower-dimensional scores may be noisy and not spatially predictable. This problem is also present in spatial gene expression analyses, where the reduced-dimensional

gene expression matrix may not preserve all biologically meaningful patterns. When the processed gene expression data further undergo spatial imputation due to incomplete profiling (Li et al., 2021; Pierson and Yau, 2015; Prabhakaran et al., 2016), or are clustered into spatial domains based on spatial and other biological information (commonly termed *spatial domain detection* in spatial transcriptomics) (Long et al., 2023; Zhao et al., 2021a), the statistical performance may be affected by loss of information that seems unimportant in dimension reduction, but is significant for downstream modeling (Liu et al., 2022).

Principal component analysis (PCA) (Jolliffe, 2002) is a classical dimension-reduction technique leading to independent, lower-dimensional scores (called principal component scores, or PC scores) approximating the multivariate measurements. Existing methodologies that tackle some of the aforementioned challenges are often based on extensions of PCA. Jandarov et al. (2017) proposed a spatially predictive PCA algorithm, where the PC scores were constrained to be linear combinations of some covariates and spatial basis, so that they can be more accurately predicted at new locations. Bose et al. (2018) extended this predictive PCA approach and enabled adaptive selection of covariates to be included for each PC. Vu et al. (2020, 2022) proposed a probabilistic version of predictive PCA along with a low-rank matrix completion algorithm, which offers improved performance when there is spatially informative missing data. Keller et al. (2017) developed a predictive k-means algorithm where dimension reduction was conducted by clustering.

In spatial transcriptomics applications such as domain detection (see detailed discussion in Section 3.5.2), Zhao et al. (2021a) processed gene expression information using PCA and conducted downstream clustering via a Bayesian approach, where the spatial arrangement of the measured spots were modeled using the PC scores. Shang and Zhou (2022) proposed a probabilistic PCA algorithm, where spatial information was incorporated into a latent factor model for gene expression. Liu et al. (2022) developed a joint approach that simultaneously conducted dimension reduction and clustering, and used a latent hidden Markov random field model to enforce spatial structures of the clusters.

When dimension reduction is used as an intermediate step before downstream analyses such as prediction, statistical inference and clustering, there are two key considerations for the quality of dimension reduction: The first is *representability*, which is how well the lower dimensional components approximate the original measurements. The second, often conflicting, goal is *predictability*

of the resulting components, so that they preserve meaningful scientific and spatial information and are of high quality for subsequent modeling. Even in analytical tasks that do not primarily focus on prediction (e.g., domain detection in spatial transcriptomics), the predictability of the PC scores is still desirable and implicitly considered, since it enforces the interpretability and spatial structure of the PCs — see Section 3.5.2 for further illustration.

Among existing methodologies, classical PCA solely focuses on representability, while predictive PCA (Jandarov et al., 2017) and its variants (Bose et al., 2018) prioritize predictability and optimize representability only after the former is guaranteed. Probabilistic approaches such as Shang and Zhou (2022), Liu et al. (2022) and Vu et al. (2020) implicitly incorporated both tasks into the latent factor models, though their performance depends on the validity of the parametric model assumptions, and there is no explicit interpretation or optimality guarantee on either property or the trade-off between them.

In this chapter, we propose a flexible dimension reduction algorithm, termed *representative and predictive PCA* (RapPCA), that explicitly minimizes a combination of representation and prediction errors and finds the optimal balance between them. We further allow the underlying lower-dimensional scores to have complex spatial structure and non-linear relationships with external covariates (if any). We show that the optimization problem involved can be solved by eigendecomposition on transformed data, enabling simple and efficient computation.

We start in Section 3.2 by briefly reviewing related methods, introducing notations, and defining various performance metrics of dimension reduction. Section 3.3 describes our proposed approach, and establishes the optimality of the proposed solution. We compare the performance of our method with existing variants of PCA via simulation studies in Section 3.4, and demonstrate its application in epidemiological exposure assessment as well as spatial transcriptomics in Section 3.5. Section 3.6 concludes this chapter with a discussion of our methodology and related alternatives.

3.2 Setting

3.2.1 Classical and Predictive PCA

Let $Y \in \mathbb{R}^{n\times p}$ represent p-dimensional spatially-structured variables of interest measured at n locations. Examples include the concentrations of a mixture of p pollutants at n monitoring sites,

or the expression of p genes at n spots on tissues. In addition, we observe the spatial coordinates $\{(s_{i1}, s_{i2})\}_{i=1}^n$ of the n locations, along with (potential) observations of d covariates, $X \in \mathbb{R}^{n\times d}$. These could be, for example, population density and land use information for air pollution studies, or histology information from images for spatial transcriptomics.

Our ultimate goal is to conduct prediction, clustering or other modeling tasks on Y based on the spatial coordinates and covariates. Since the columns of Y may be strongly correlated and/or noisy (Rao et al., 2021), dimension reduction can be used to to extract scientifically meaningful signals from the original data. The classical PCA achieves this by maximizing the proportion of variability in Y that is explained by the lower dimensional PCs. We briefly introduce the classical PCA algorithm under the equivalent formulation of Shen and Huang (2008) to highlight its connection with related techniques.

Suppose the columns of Y are properly centered and scaled. To find an r-dimensional representation of Y, the classical PCA can be expressed as a sequence of rank-1 optimization problems (Shen and Huang, 2008) for $l = 1, \ldots, r$:

$$\min_{u_l, v_l} \left\| Y^{(l)} - u_l v_l^\top \right\|_F^2 \quad \text{s.t. } \|v_l\|_2 = 1 \tag{3.1}$$

where the 1-dimensional PC score u_l is an n-vector, the loading v_l is $p \times 1$, and $\|\cdot\|_F$ and $\|\cdot\|_2$ represent Frobenius norm for matrices and ℓ_2 norm for vectors, respectively. $Y^{(l)}$ is the residual from approximation after each iteration. In other words, denoting $\tilde{u}_l, \tilde{v}_l$ as the solution to (3.1), we define $Y^{(l)} = Y^{(l-1)} - \tilde{u}_{l-1}\tilde{v}_{l-1}^\top$ for $2 \le l \le r$ and $Y^{(1)} = Y$.

Combining the optimal solutions as $\tilde{U} = [\tilde{u}_1, \ldots, \tilde{u}_r]$, we obtain the r-dimensional PC scores $\tilde{U}$ as an approximation of the p-dimensional data Y. However, while each of the u_l's explain the greatest variability in Y, there is no guarantee that they are scientifically interpretable, nor do they explicitly account for the spatial structures underlying the observations of Y. A natural idea is to constrain each PC score u_l within certain model space based on the spatial coordinates and covariates X (if any), which motivates the predictive PCA (Jandarov et al., 2017) algorithm.

Predictive PCA builds upon the expression in (3.1) and solves

$$\min_{\alpha_l, v_l} \left\| Y^{(l)} - \left(\frac{Z\alpha_l}{\|Z\alpha_l\|_2} \right) v_l^\top \right\|_F^2 \tag{3.2}$$

for each $l = 1, \ldots, r$. Here, $Z = [X \ B]$ where $B \in \mathbb{R}^{n\times m}$ contains m thin-plate spline functions

capturing the spatial effects. $Y^{(l)}$ is also defined as the residual after approximation, i.e. $Y^{(l)} = Y^{(l-1)} - Y^{(l-1)}\frac{\tilde{v}_{l-1}\tilde{v}_{l-1}^\top}{\|\tilde{v}_{l-1}\|_2}$ for $2 \leq l \leq r$ and $Y^{(1)} = Y$.

After solving either (3.1) or (3.2) for all r PCs, the loadings are ℓ_2-normalized ($\tilde{v}_r \leftarrow \tilde{v}_r/\|\tilde{v}_r\|_2$) if not already done, and they are concatenated as $\tilde{V}_{p\times r} := [\tilde{v}_1 \ \ldots \ \tilde{v}_r]$. The PC scores are then defined as $\tilde{U}_{n\times r} := Y\tilde{V}$.

By restricting the induced PC score to fall within the linear span of Z, the solution to (3.2) enforces the resulting PC score to have spatial smoothness and contain information from the covariates X. This solution is therefore more predictable by X along with spatial smoothing, compared to classical PCA.

As noted earlier, classical PCA aims to achieve optimal representability of the PC scores $\tilde{u}$, but these $\tilde{u}$'s may not retain meaningful signals in X or important spatial patterns to be well predictable. In contrast, predictive PCA and its variants (Bose et al., 2018; Jandarov et al., 2017) minimize the approximation gap, which will be formally defined in Section 3.2, after constraining each $\tilde{u}$ to lie within a specific model space; but when these PC scores $\tilde{u}$ are predicted at new locations and transformed back to the higher dimensional space of Y, closeness in the space of $\tilde{u}$ may not necessarily translate to the original space of observations Y due to reduced approximation accuracy. Our proposed method is based on an interpolation between classical and predictive PCA, and encourages $\tilde{u}$ to be close to, but not exactly within, some model space while balancing the quality of representation. This proposal will be introduced in detail in Section 3.3.

3.2.2 Evaluation Metrics for Dimension Reduction

Before introducing our proposal, it is helpful to formalize different aspects of dimension reduction performances, such as representability and predictability, and to characterize the overall balance between them. Though the ultimate goal may not exactly be the prediction of Y, we quantify the predictability of the PC scores in the training-test setting because desirable predictability indicates important scientific information (based on the external covariates X) and spatial structure being preserved in the resulting PC scores.

Suppose the PC scores $\tilde{U}_{\text{trn}} = [\tilde{u}_1 \ \ldots \tilde{u}_r]$ and loadings $\tilde{V}$ are obtained on a set of training data, and $\tilde{u}_1, \ldots, \tilde{u}_r$ are then predicted at the test locations with a spatial prediction model as

$\hat{U}_{\text{tst}} := [\hat{u}_1 \ \ldots \ \hat{u}_r]$. We define the *mean-squared prediction error* (MSPE) of this procedure as

$$\text{MSPE} = n^{-1} \left\| (\hat{U}_{\text{tst}} - U^*)\tilde{V}^\top \right\|_F^2$$

where $U^* = \arg\min_U \|Y_{\text{tst}} - U\tilde{V}^\top\|_F^2 = Y_{\text{tst}}\tilde{V}$. Thus, U^* is what the actual PC scores on the test set would be according to the loadings $\tilde{V}$, if Y_{tst} were known. Consequently, MSPE characterizes the gap between the predicted and true PC scores and reflects the predictability of the PCs.

The *mean-squared representation error* (MSRE) is defined as

$$\text{MSRE} = n^{-1} \left\| Y_{\text{tst}} - U^*\tilde{V}^\top \right\|_F^2,$$

which measures the gap between the true data, Y_{tst}, and the closest approximation possible based on $\tilde{V}$. Since the quality of representation alone, without considering predictive performance, is of more interest for the training than the test data, we instead examine MSRE-trn as the metric for representation: MSRE-trn $= n^{-1}\|Y_{\text{trn}} - \tilde{U}_{\text{trn}}\tilde{V}^\top\|_F^2$. It can be seen that MSRE-trn exactly matches the objective function for PCA in (3.1).

We finally define the *total mean-squared error* (TMSE) resulting from both dimension reduction and prediction as

$$\text{TMSE} = n^{-1} \left\| Y_{\text{tst}} - \hat{U}_{\text{tst}}\tilde{V}^\top \right\|_F^2.$$

The TMSE measures the discrepancy between the true data, Y_{tst}, and the predicted scores that are transformed back to the original, higher-dimensional space. Also, we note that MSPE and MSRE can be viewed informally as a decomposition of the overall TMSE, where the prediction and approximation gaps, i.e., $(U^* - \hat{U}_{\text{tst}})\tilde{V}^\top$ and $Y_{\text{tst}} - U^*\tilde{V}^\top$, add up to the overall error $Y_{\text{tst}} - \hat{U}_{\text{tst}}\tilde{V}^\top$.

3.3 Method

Predictive PCA restricts the PC scores $\tilde{u}$ within some model space — specifically, span(Z) — to enforce predictability while trading off representability to some extent. A natural idea with more flexibility would be relaxing this constraint, and instead letting $\tilde{u}$ be close to, but not exactly within, the model space. By choosing the magnitude of penalties data-adaptively for the representation error versus the distance between $\tilde{u}$ and the model space, we aim to achieve the optimal balance between representation and prediction leading to the smallest overall error.

Noting that the model space can be generalized from the linear span in (3.2) to incorporate more flexible mechanisms underlying each PC $\tilde{u}$, we adopt kernel method (Evgeniou et al., 2000; Hastie et al., 2009) to capture non-linear covariate effects; specifically, we use smoothing splines (Wahba, 1990; Wood, 2003) to enforce spatial structures.

Combining these ideas, we propose to solve the following sequence of optimization problems, for $l = 1, \ldots, r$, to extract r PCs from Y:

$$\begin{aligned} &\min_{u,v,\alpha,\beta} \; f_{\gamma,\lambda_1,\lambda_2}(u, v, \alpha, \beta) := \left\| Y^{(l)} - uv^\top \right\|_F^2 + \gamma \left\| u - (K\alpha + B\beta) \right\|_2^2 + \lambda_1 \alpha^\top K \alpha + \lambda_2 \beta^\top Q \beta \\ &\text{s.t. } u = Y^{(l)} v, \; v^\top v = 1; \end{aligned}$$

here, $Y^{(l)}$ is defined as the residual $Y^{(l-1)} - \tilde{u}_{l-1} v_{l-1}^\top$ for $2 \leq l \leq r$, and set to be Y when $l = 1$. Moreover, K is the kernel matrix such that $K_{ij} = k(X_{i\cdot}, X_{j\cdot})$ for some kernel function $k(\cdot, \cdot)$, with $X_{i\cdot}$ being the ith observation (row) of X. The columns of B are evaluations of m spline basis functions at the spatial coordinates $\{(s_{i1}, s_{i2})\}_{i=1}^n$, and Q is the penalty matrix for the spline basis. In practice, we replace K and Q with $\tilde{K} := K + \delta I_n$ and $\tilde{Q} := Q + \delta I_m$ for a small constant δ and identity matrices I_n, I_m in the penalty terms, to avoid near-singularity of the penalties in computation, which leads to

$$\begin{aligned} &\min_{u,v,\alpha,\beta} \; f_{\gamma,\lambda_1,\lambda_2}(u, v, \alpha, \beta) := \left\| Y^{(l)} - uv^\top \right\|_F^2 + \gamma \left\| u - (K\alpha + B\beta) \right\|_2^2 + \lambda_1 \alpha^\top \tilde{K} \alpha + \lambda_2 \beta^\top \tilde{Q} \beta \\ &\text{s.t. } u = Y^{(l)} v, \; v^\top v = 1. \end{aligned} \tag{3.3}$$

In (3.3), the first term $\left\| Y^{(l)} - uv^\top \right\|_F^2$ measures the approximation gap of the score u and loading v, similar to the objective of minimization for classical PCA. The second term reflects the distance between the score u and the model space specified by K and B, and encourages u to be predictable based on the covariates and spatial coordinates. In particular, the kernel method allows for non-linear relationships between X and u, and is more flexible than most existing, predictive variants of PCA (e.g., Bose et al., 2018; Jandarov et al., 2017; Shang and Zhou, 2022; Vu et al., 2020) which specify linear relationships. The constraint $u = Y^{(l)} v$ ensures a similar relationship between the PC scores and loadings as in classical PCA. The tuning parameter γ controls the trade-off between predictability and representability, with larger γ imposing higher penalty on unpredictable scores u. The ℓ_2 penalty terms in (3.3) enforces the identifiability of the model parameters.

Though optimizing u, v, α, β iteratively via coordinate descent is straightforward, the optimization problem in (3.3) is biconvex (Gorski et al., 2007) and such an algorithm is not guaranteed to converge to the global minimizer. Instead, we propose an analytical solution to (3.3) that attains the miminum despite the non-convexity of the objective function $f_{\gamma,\lambda_1,\lambda_2}(u, v, \alpha, \beta)$ and/or the constraints. We describe the proposed algorithm in Theorem 3.1 and provide a proof in Appendix B.1. We also present examples in Appendix B.2 that numerically verify the optimality of the solution in Theorem 3.1.

Theorem 3.1. *Denote the singular value decomposition of* $Y^{(l)}$ *as* $Y^{(l)} = UDV^\top$. *Let* $\eta := \left[\alpha^\top \quad \sqrt{\frac{\lambda_2}{\lambda_1}}\beta^\top\right]^\top$, $Z := \left[K \quad \sqrt{\frac{\lambda_2}{\lambda_1}}B\right]$, *and let the combined penalty matrix be*

$$P := \begin{bmatrix} \tilde{K} & 0 \\ 0 & \frac{\lambda_2}{\lambda_1}\tilde{Q} \end{bmatrix}.$$

Then the optimization problem (3.3) has a unique solution given by

- $\tilde{v} = Vq$ *where* q *is the (normalized) first eigenvector of*

$$-(\gamma - 1)D^2 + \gamma^2 DU^\top Z(\gamma Z^\top Z + \lambda_1 P)^{-1} Z^\top UD;$$

- $\tilde{u} = Y^{(l)}\tilde{v}$;

- $\tilde{\eta} = (\gamma Z^\top Z + \lambda_1 P)^{-1}(\gamma Z^\top UDq)$, *with* q *defined above.*

The computational complexity of the procedure described above is $O\left((n+m)^3 + n^2p + np^2\right)$, where we recall that n is the sample size, p is the dimensionality of Y, and m is the number of spline basis functions included in B. This algorithm is computationally simple in that it obtains the optimal solution with an explicit expression, as opposed to iterative numerical optimization procedures.

3.4 Simulations

In this section, we illustrate the performance of our proposed method, RapPCA, with simulated data, in comparison to classical and predictive PCA. In particular, we investigate three scenarios

with different trade-offs between predictability and representability resulting from different data generating mechanisms.

In all settings, we randomly generate $n = 200$ locations $\{(s_{i1}, s_{i2})\}_{i=1}^{n}$ within $[0,1] \times [0,1]$. For each location, we generate $d = 10$ covariates x_{ij} ($i = 1, \ldots, n$ and $j = 1, \ldots, d$) from independent Uniform$[-1, 1]$ distributions and calculate 6 independently distributed PCs as

$$\mathrm{PC}^{(l)} = f_l(X) + \epsilon_l,\ l = 1, \ldots, 6,$$

where $\epsilon_l \sim \mathrm{Normal}(0, \sigma_l^2 \Sigma)$ with σ_l^2 controlling the signal-to-noise ratio and hence the predictability of each PC; $f_l(\cdot)$ is a specified mean function, and the covariance matrix Σ has an exponential structure with $\Sigma_{ii'} = 0.5 \exp\left(-\frac{(s_{i1}-s_{i'1})^2+(s_{i2}-s_{i'2})^2}{0.5}\right) + 0.5$. The outcome Y represents concentrations of $p = 15$ pollutants, and is given by

$$Y_{n\times p} = \mathrm{PC}_{n\times 6} M_{6\times p} + \varepsilon_{n\times p}$$

where M is adjusted differently in each scenario to control the contribution from different (e.g., predictable versus unpredictable) PCs, and the entries of the noise terms ε are drawn from independent Normal$(0, 0.1)$ distributions.

We examine the metrics of interest including TMSE, MSPE and MSRE-trn for 100 replicates of data, in terms of their overall magnitudes as well as the breakdown by each PC and/or γ (the tuning parameter). The optimal γ is selected by minimizing TMSE via cross-validation. More specifically, we conduct 10-fold cross-validation and extract $r = 3$ PCs sequentially with PCA, predictive PCA and RapPCA on each training dataset, as described in Sections 3.2 and 3.3. Prediction is done by a two-step procedure, where we first train a random forest model for each of the extracted PCs $\tilde{u}_1, \tilde{u}_2, \tilde{u}_3$, and then conduct spatial smoothing on the residuals with thin-plate regression splines (TPRS) (Wood, 2003).

3.4.1 Scenario 1: PCs with Equal Contribution

We first consider the setting with 3 (out of 6) well-predictable PCs and 3 PCs mainly consisting of structured and unstructured Gaussian errors. More specifically, we let $f_l(X) = X\beta_l$ where the entries of β_l are drawn independently from Uniform$[-1, 1]$, and $\sigma_l^2 = 0$, for $l = 1, 2, 3$; we let $f_l = 0$ and $\sigma_l^2 = 1$ for $l = 4, 5, 6$. M is set to be $\Lambda\tilde{M}$, where Λ is diagonal with $\Lambda_{ll} = 1/\mathrm{sd}(\mathrm{PC}^{(l)})$ and the

entries of $\tilde{M}$ are independent Uniform$[-1, 1]$. In other words, we first scale the 6 PCs to have equal variability, and then assign random but overall comparable weights to them.

It is expected in this scenario that predictable PCA would outperform classical PCA in prediction without severely compromising approximation accuracy, since the predictable PCs explain a similar amount of variability compared to the unpredictable ones. Consequently, predictable PCA would achieve better overall performance as reflected by TMSE. RapPCA flexibly interpolates between classical and predictable PCA, and is expected to have comparable or improved performance than predictable PCA.

Panel A of Figure 3.1 presents the breakdown of MSPE, MSRE-trn (which is MSRE on the training set) and TMSE for the first PC by γ, the main tuning parameter for RapPCA, in comparison to PCA and predictive PCA for this scenario. We only present this breakdown for the first PC but not the subsequent ones because different PCA algorithms will have different residuals $Y^{(l)}$ (recall Equations 3.1, 3.2 and 3.3) starting from the second PC, and the PC-specific breakdown of metrics are consequently no longer comparable. We observe that MSPE decreases and training MSRE increases as we increase γ for RapPCA, since this imposes a greater penalty on prediction errors in optimization.

Figure 3.2 compares the prediction MSEs for each PC, as well as the overall TMSE, MSPE and MSRE-trn for different methods. As expected for this scenario, we observe in Panel A that predictive PCA achieves lower prediction errors than PCA, for each PC as well as the overall MSPE. Such advantage does not cost a higher approximation gap (MSRE), since the predictable components in the outcome Y explain a similar amount of variability as the unpredictable ones. Our method achieves comparable but improved prediction accuracy than predictive PCA, because it is able to adjust the penalties on the covariate versus spatial effect terms adaptively, especially when there are a large number of spatial basis terms (recall the separate tuning parameters λ_1 and λ_2 in Equation 3.3), as opposed to predictive PCA which is restricted to a low-dimensional combination of covariates and spatial basis functions (recall the term Z in Equation 3.2). Consequently, predictive PCA achieves better overall performance (reflected by smaller TMSE) than PCA, while RapPCA further improves the prediction and overall performance due to its flexibility in capturing the covariate and spatial effects.

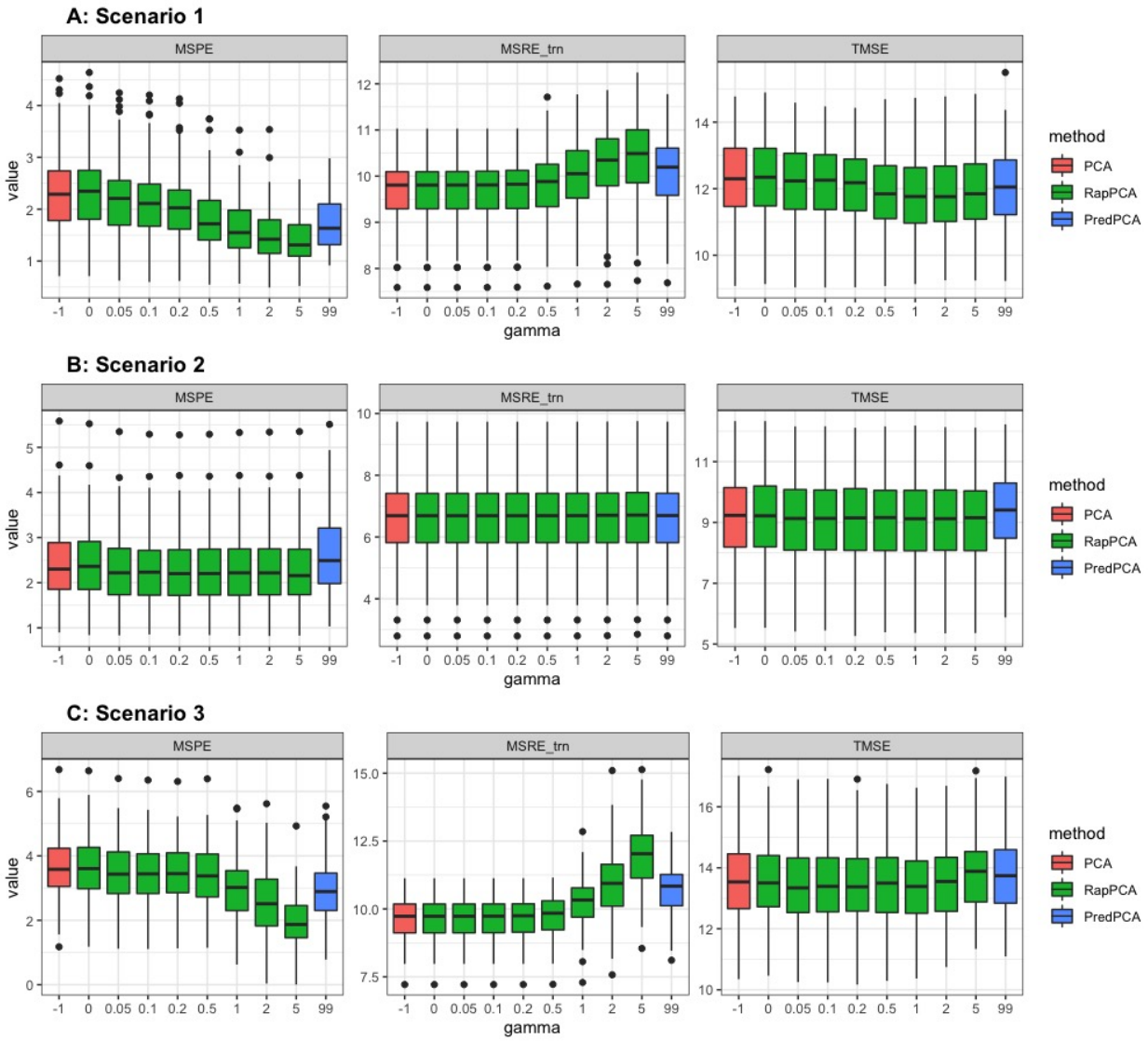


Figure 3.1: Breakdown of MSPE, MSRE-trn (which is MSRE on the training set) and TMSE for the first PC by γ arcoss 100 replicates of data, with classical PCA (coded as $\gamma = -1$) and predictive PCA (`PredPCA`, coded as $\gamma = 99$) results presented for reference

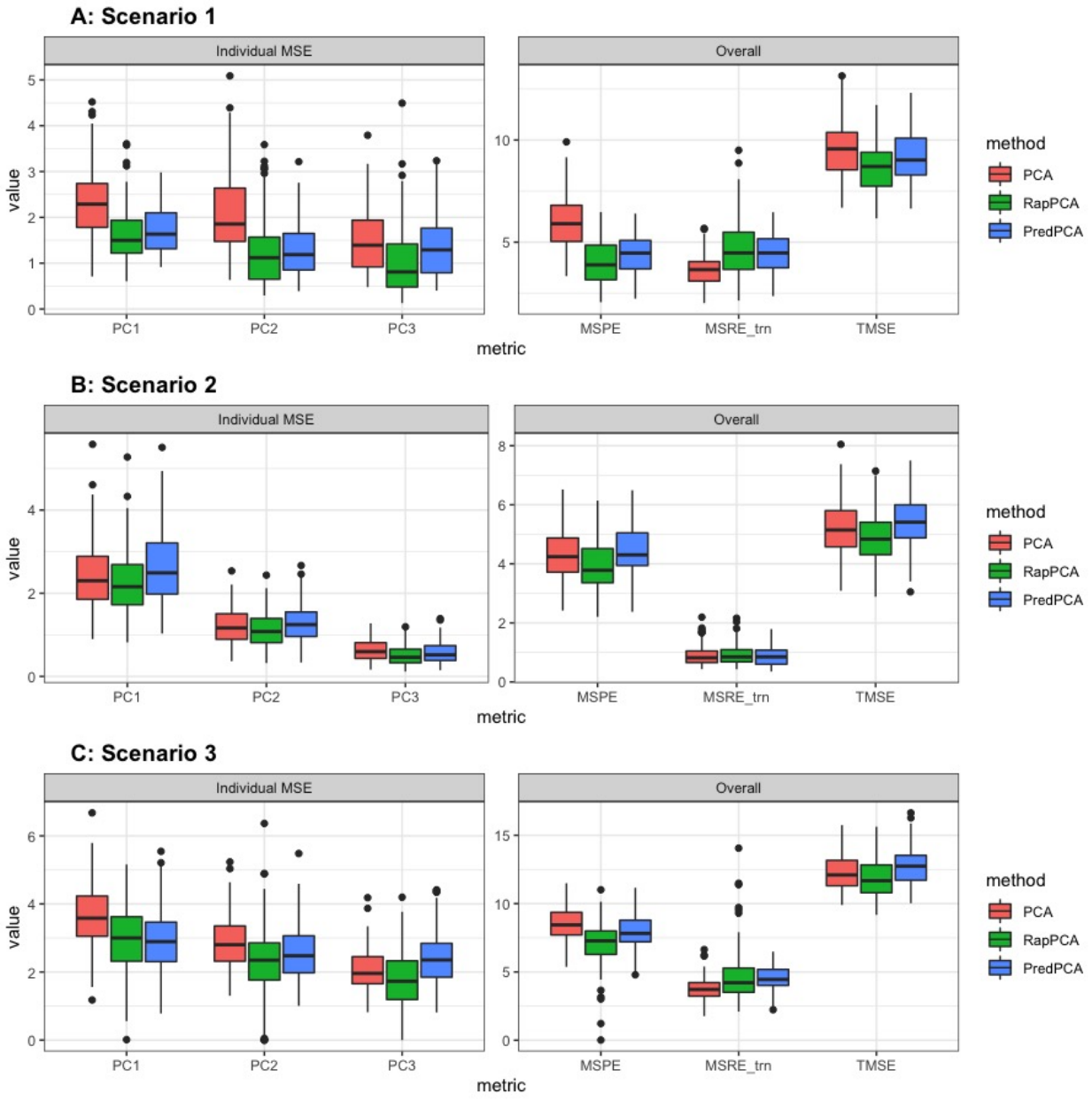


Figure 3.2: Individual prediction MSEs for each PC (left) and overall metrics for all PCs (right) arcoss 100 replicates of data

3.4.2 Scenario 2: PCs with Unequal Contribution

We now consider a scenario where the predictable PCs explain a higher proportion of variability than the unpredictable PCs. In particular, the simulation setting is similar to Scenario 1, except that the entries of $\tilde{M}$ are first drawn from independent Uniform$[-1, 1]$ distributions, and then the rows are scaled to have decaying norm (recall that the first 3 PCs are more predictable than the last 3 PCs). In contrast to Scenario 1, we expect in this case that PCA and predictive PCA would have comparable prediction, approximation and overall accuracy, since the well-predictable components of Y coincide with those that explain the greatest amount of variance.

For the same reason, adjusting γ in this case would not lead to steep changes in MSPE, MSRE or TMSE for RapPCA, as reflected by the flat curves of each metric in Panel B of Figure 3.1. Despite the similar behaviors of PCA and predictive PCA, Panel B of Figure 3.2 reveals that RapPCA is still able to improve the prediction and overall performance by more flexibly exploiting the covariate and spatial information when identifying the PCs.

3.4.3 Scenario 3: PCs with Non-Linear Mean

In our last simulation, we investigate the setting where the relationship between the true PCs and the covariates is non-linear. We modify the setting in Scenario 1 so that

$$f_l(X) = X^{\odot 2}\beta_l + 2\sum_{j=1}^{\lfloor d/2 \rfloor}(X_{2j-1} \odot X_{2j})\alpha_{lj},$$

for $l = 1, 2, 3$, where $\odot$ and $\odot^2$ denote element-wise multiplication and square, and $\lfloor x \rfloor$ denotes the largest integer not exceeding x. In other words, we specify the mean function to be a combination of squared and interaction terms of the covariates, where interactions exist between pairs of consecutive covariates ($X_{\cdot 1}$ and $X_{\cdot 2}$; $X_{\cdot 3}$ and $X_{\cdot 4}$, etc).

Panel C of Figure 3.1 indicates that the flexibility of handling non-linear effects enables a clear advantage in prediction accuracy for RapPCA compared to both predictive and classical PCA in this scenario, when we impose a large enough penalty γ on prediction errors in optimization. It is not surprising that putting an emphasis on prediction leads to some loss in approximation accuracy; however, we could observe from Panel C of Figure 3.2 that by controlling such balance in a data-adaptive way, RapPCA is able to achieve meaningful improvement in MSPE while maintaining

comparable, if not superior, overall performance.

3.4.4 Remarks on the Predictability-Representability Trade-Off

In all of our simulation studies, we choose the key tuning parameter γ of RapPCA data-adaptively by minimizing TMSE in cross-validation. While TMSE is a comprehensive and balanced metric to use for tuning, alternative choices can be made with considerations on the practical use cases.

For example, in health effect studies of air pollution with spatial misalignment between the locations of measured health outcomes and those of measured pollutant concentrations, the accuracy of such spatial extrapolation is arguably more important than the closeness of approximation for the original pollution measurements (Jandarov et al., 2017). In this case, researchers may examine the breakdown of metrics in cross-validation similar to Figure 3.1 and manually select a value of γ leading to satisfying prediction accuracy (MSPE) without significantly compromising MSRE or TMSE. We also note from the curvature of TMSE versus γ in these figures that for many practical cases, the overall performance of RapPCA is not too sensitive to the choice of γ, or the metric(s) used to select it. In other words, manually adjusting the predictability-representability trade-off with RapPCA leads to consistently improved predictability and better, or at least similar, overall accuracy (TMSE) compared with existing methods such as PCA and predictive PCA.

3.5 Applications

We illustrate the utility of RapPCA using real datasets in two different domain areas — environmental epidemiology and spatial transcriptomics. The first application of dimension reduction in Section 3.5.1 represents the case where practitioners seek to improve predictability (MSPE) and the overall performance (TMSE), and hence RapPCA is able to explicitly optimize these metrics as desired. The second example in Section 3.5.2 further includes the scenario where tasks other than spatial prediction (e.g., clustering) are of interest, and we demonstrate that RapPCA, though not directly incorporating the ultimate goal into optimization, is still able to capture meaningful biological and spatial information in the extracted PCs by enforcing predictability, and thus achieves desirable model performance.

3.5.1 Analysis of Seattle Traffic-Related Air Pollution Data

The performance of RapPCA in comparison to common existing methods, including classical and predictive PCA, is first illustrated with the multivariate traffic-related air pollution (TRAP) data in Seattle (Blanco et al., 2022b). The data came from the same study as in Section 2.2 where a vehicle equipped with air monitors repeatedly collected two-minute samples at $n = 309$ stationary roadside sites in the greater Seattle area. Repeated samples for the concentrations of 6 types of pollutants were obtained, including ultrafine particles (UFP), black carbon (BC), nitrogen dioxide (NO_2), carbon monoxide (CO), carbon dioxide (CO_2) and fine particulate matter ($PM_{2.5}$). For UFP and BC, the concentrations were measured with multiple instruments corresponding to different measurement ranges or units. In particular, 13 NanoScan instruments with different bin sizes, as well as DiSCmini and PTRAK 8525 (with and without diffusion screen) instruments measured the concentration of UFP in terms of the counts of particles with different sizes, along with the median size and overall count of particles. Black carbon was assessed by micro-aethalometer (MA200), which measures the concentration of particles with different light absorbing properties, corresponding to 5 different ranges of wavelengths.

The median 2-minute visit concentrations were winsorized at the site level such that concentrations below the 5^{th} and above the 95^{th} quantile for a given site site were set to those thresholds, respectively. This was done to reduce the influence of outliers in the measurements. These winsorized medians were then averaged for each site, leading to $p = 27$ annual average measurements of pollutant concentrations in total, including 18 for UFP, 5 for BC, and one for each of NO_2, CO, CO_2 and $PM_{2.5}$. We remove the concentration of $PM_{2.5}$ from our analysis, and instead use it to normalize all other variables except the sizes of UFP. All variables are centered and scaled before running each PCA algorithm.

We are interested in extrapolating the pollutant measurements to unobserved locations in the same study region. Because of the autocorrelation between these measurements, predicting each of them separately would not lead to sensible results; instead, dimension reduction is needed before building spatial prediction models. As in Section 3.4, we compare the performance of RapPCA with classical and predictive PCA, each extracting the top 3 PCs from the 27 measurements of air pollutants. Due to the high dimensionality of geographical covariates, PCA is run on these 189

covariates and the top 15 PCs are used as predictors for predictive PCA. We examine the overall metrics as well as the individual prediction MSEs for each PC via 10-fold cross-validation, where the same spatial prediction approach as Section 3.4 is followed, namely, a random forest model followed by spatial smoothing via TPRS on the residuals.

Table 3.1 compares each dimension reduction algorithm in terms of cross-validated TMSE, MSPE and MSRE (on the training data) as well as prediction MSEs for each PC. Consistent with the findings from simulation studies, RapPCA achieves the lowest prediction and overall errors, while PCA has the smallest approximation gap on the training data. The advantage in predictability of RapPCA is reflected in both the overall MSPE and the individual MSEs for each PC. Predictive PCA is not guaranteed to show an advantage in predictability, especially in settings with high-dimensional predictors (covariates and/or spatial basis), since it requires a separate processing step on these predictors instead of selecting the information to use for prediction data-adaptively as RapPCA does.

	Overall Metrics			Individual MSEs		
	TMSE	MSPE	MSRE-trn	PC1	PC2	PC3
PCA	13.97	7.55	5.83	5.23	1.69	0.62
PredPCA	14.15	7.74	6.20	5.63	1.47	0.63
RapPCA	13.15	6.93	6.14	4.93	1.38	0.61

Table 3.1: Comparison of overall metrics and individual prediction MSEs for each PC, assessed by 10-fold cross-validation on the Seattle traffic-related air pollution data

We then run dimension reduction with each of the PCA algorithms on the whole dataset to assess the spatial distribution of top 3 PC scores, where the tuning parameters leading to the optimal TMSE is selected for RapPCA. Next, we train spatial prediction models for each of the PC scores, and these models are evaluated on a finer grid of 5,040 locations over the study region to obtain smoothed plots of each PC score. Figure 3.3 visualizes the smoothed PC scores obtained by each method across the study region, and Figure 3.4 presents the PC loadings reflecting the contribution of each pollutant on the PC scores. We observe similar spatial patterns in the distribution of the PC scores across different methods, except for the south end of the study region for the third PC where

PCA identifies a stronger signal than RapPCA and PredPCA. In particular, all methods highlight regions near the airport and/or around major roads (the dark red area) for the first PC, indicating aircraft and road traffic emissions as a major source of overall pollution level. This is consistent with the large contributions of BC as well as UFP with small or moderate sizes (Bendtsen et al., 2021; Zhang et al., 2019), as reflected by the loadings of UFP (with relatively small particle sizes) and BC in Figure 3.4.

We note that despite the similarity of PC scores or loadings for RapPCA as compared with standard or predictive PCA, significant advantages in the predictability of PCs could still be achieved by just slightly varying the loadings based on our algorithm (see Table 3.1). This corroborates the implications of simulations that by optimizing the predictability-representability trade-off, it is typically possible to obtain PCs that are better explained by the covariates, without affecting the overall or approximation errors from dimension reduction.

3.5.2 Analysis of HER2-Positive Breast Tumor Spatial Transcriptomics

In this section, we present another use case of RapPCA by analyzing spatial transcriptomics data. While dimension reduction is also a common data handling step in these applications, there are a variety of downstream modeling tasks where spatial prediction, as in the Seattle TRAP study, may not be the primary focus. We will nevertheless illustrate that optimizing the balance between predictability and representability with RapPCA still leads to desirable performance in these analytical tasks.

We analyze the HER2-positive (human epidermal growth factor receptor 2) breast tumor data (Andersson et al., 2021), which includes expression measures of genes in HER2-positive tumors from eight individuals. We focus on the first tumor section of the last individual, coded as sample H1 in Andersson et al. (2021), where the data consists of expression counts of $15,030$ genes on 613 spatial locations. Following the same filtering steps as in Shang and Zhou (2022), we removed genes with non-zero expression at less than 20 locations and those confounded by technical artifacts (Andersson et al., 2021), as well as locations with non-zero expressions for less than 20 genes. The filtered set of data contains $10,053$ genes measured on 607 spots, which is then normalized via regularized negative binomial regression as implemented by the `Seurat R` package (Hafemeister

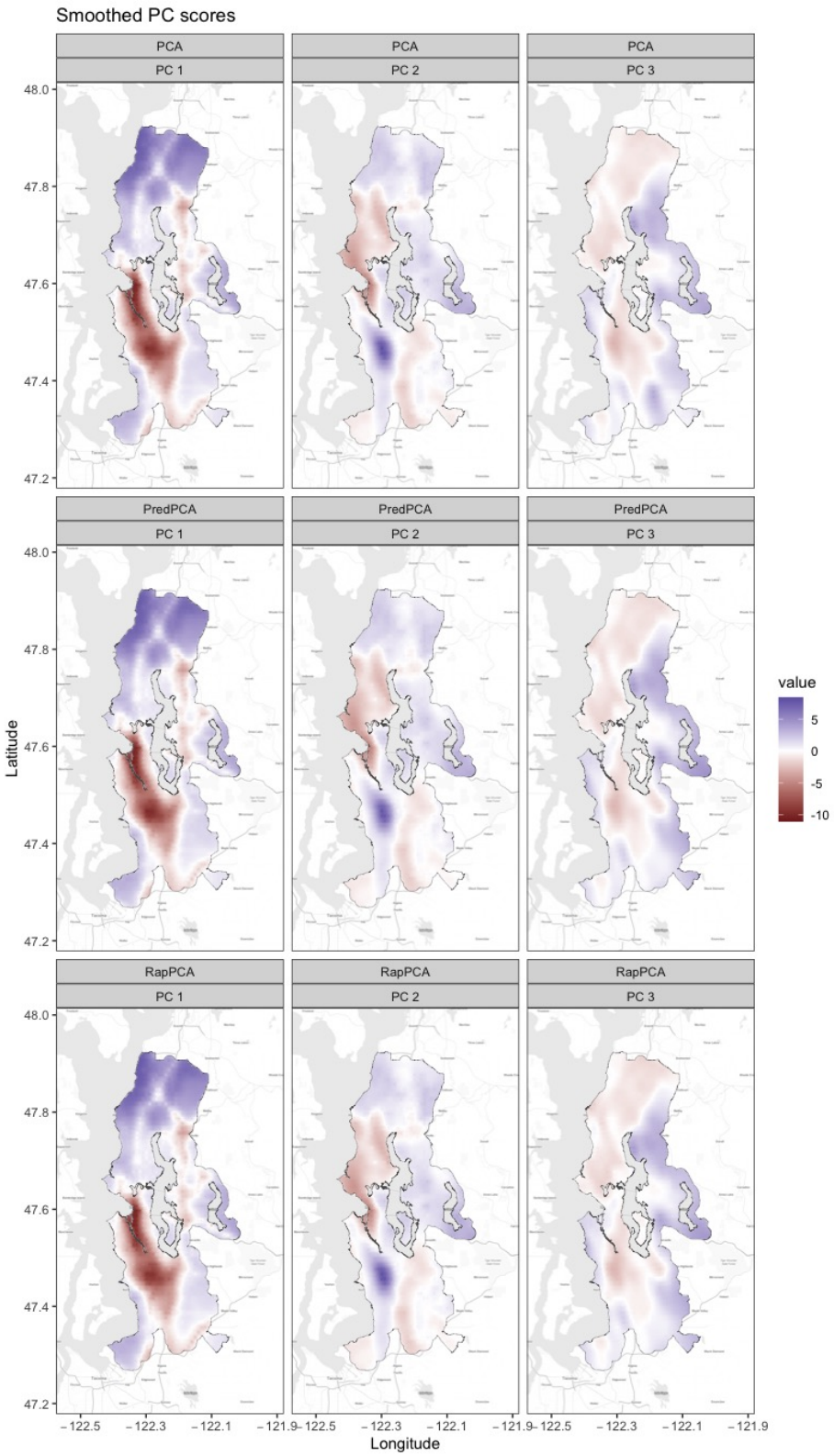


Figure 3.3: Smoothed PC scores of pollutant concentrations from the Seattle TRAP data

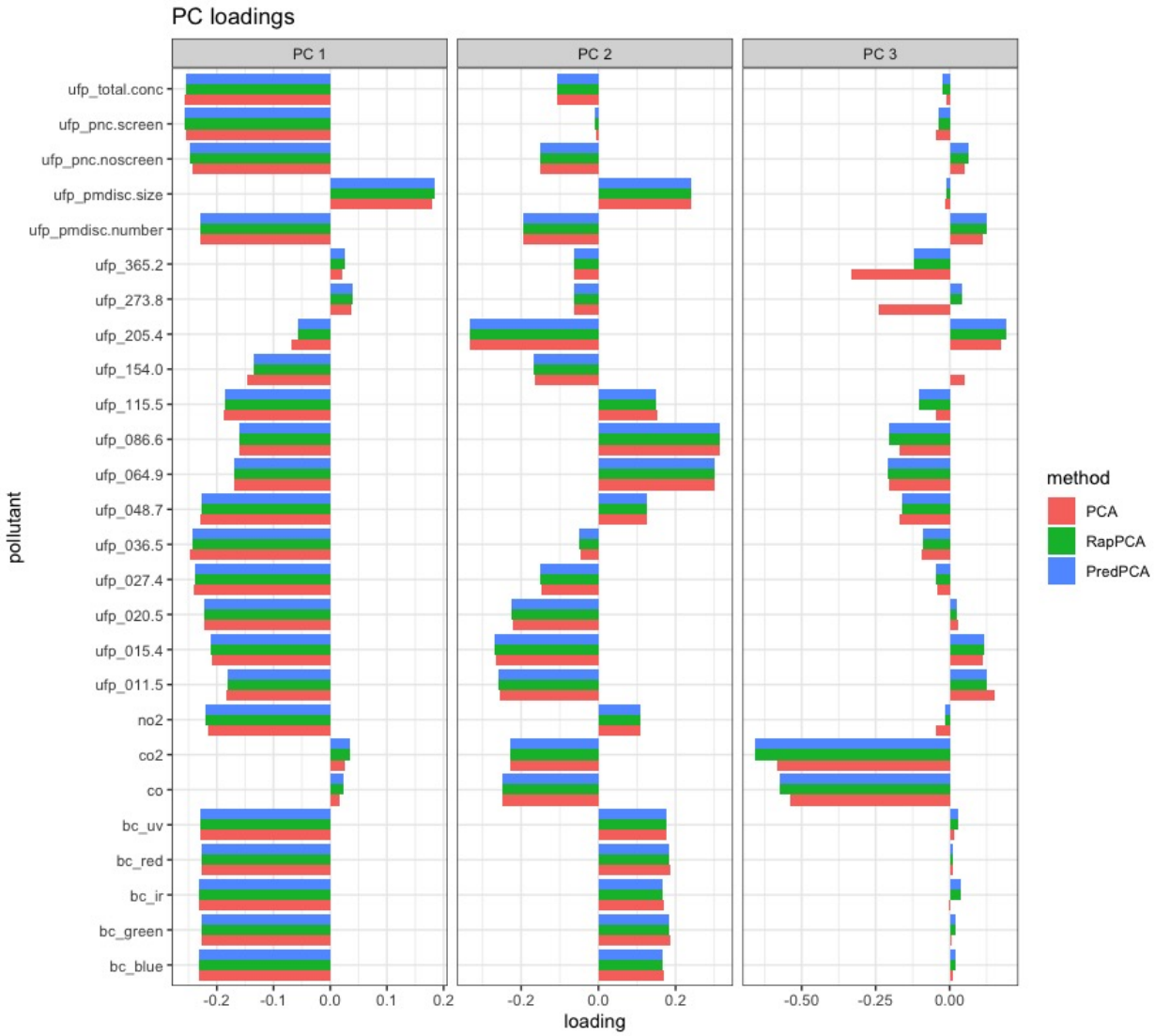


Figure 3.4: PC loadings for each pollutant. There are 6 types of pollutants, where the suffix, if applicable, represents the properties of, or the instruments used to measure, each pollutant. In particular, the numeric suffix after "ufp" represents the range of sizes for the particles.

and Satija, 2019).

Due to the large number of genes along with the noise present in the expression measurements (Rao et al., 2021), dimension reduction is commonly done before the main analytic tasks on genomic data to extract biologically meaningful signals as well as to facilitate computation; see Shang and Zhou (2022); Sun et al. (2019); Zhao et al. (2021a) for related examples. We focus on two modeling tasks on the HER2-positive breast tumor data: spatial extrapolation and domain detection. In addition to the three PCA methods (PCA, predictive PCA, RapPCA) that are investigated in previous sections, we also include spatial PCA (Shang and Zhou, 2022) for comparison, which is a probabilistic PCA algorithm for spatial transcriptomics data incorporating spatial proximity information in the modeling of PC scores.

The first application is more similar to the analysis in Section 3.5.1, where we are interested in predicting the PC scores (and hence gene expression) at tissue locations with no measurements. This prediction problem is involved, for example, in the reconstruction of high-resolution spatial maps of gene expression. For each dimension reduction method, we extract the top 20 PCs from the expression of all genes, and conduct 10-fold cross-validation to assess TMSE, MSPE and MSRE-trn, where spatial prediction is done by random forest followed by TPRS smoothing as in previous sections. For spatial PCA, we also include results from its built-in spatial prediction function (from the `SpatialPCA R` package) for completeness. Table 3.2 summarizes the overall metrics along with individual prediction MSEs for the top 3 PCs. RapPCA demonstrates significant advantage in overall prediction accuracy (MSPE) compared to all other methods, followed by predictive PCA which also has the lowest per-PC MSEs for the top 3 PCs. Also, consistent with findings from previous sections, RapPCA achieves the optimal trade-off between prediction and approximation gaps, as reflected by TMSE.

Figure 3.5 shows the top 3 PC scores from the expressions of all genes. We observe that spatial PCA and predictive PCA produce spatially smoother surfaces of PC scores compared to the other two methods. This is natural for predictive PCA since it prioritizes spatial predictability of the PC scores which results in stronger spatial smoothing; while for spatial PCA, the prediction-representation balance is implicit and less straightforward to interpret. The gap in prediction performance comparing spatial or predictive PCA with RapPCA is likely the consequence of over-smoothing, which also indicates the effectiveness of explicit optimization for the

	Overall Metrics			Individual MSEs		
	TMSE	MSPE	MSRE	PC1	PC2	PC3
PCA	272.21	32.24	239.97	6.64	5.09	3.18
PredPCA	278.08	31.85	246.15	3.93	4.12	2.26
SpatialPCA	292.78	46.09	246.70	6.84	5.89	3.69
SpatialPCA: built-in	296.59	49.90	246.70	6.81	5.96	3.81
RapPCA	271.35	20.56	250.80	5.59	4.86	3.79

Table 3.2: Comparison of overall metrics and individual prediction MSEs for each of the top 3 PCs, assessed by 10-fold cross-validation on the breast tumor data

prediction-representation balance achieved by (3.3).

As a second example, we investigate the problem of domain detection following the dimension reduction step. Domain detection is the task where different sections of tissues are identified as various spatially coherent and functionally distinct regions (Dong and Zhang, 2022; Shang and Zhou, 2022), providing helpful insights on the biological function of tissues. To this end, we conduct domain detection via walktrap clustering algorithm (Pons and Latapy, 2005) using the top 20 PCs extracted by each variant of PCA. The number of clusters is set to align with the "ground truth" labels based on pathologist annotations of tissue regions in the original study (Andersson et al., 2021).

Figure 3.6 visualizes the detected spatial domains based on PCs extracted from different dimension reduction procedures, compared to the ground truth labels. In general, the relative performance of each approach differs across spatial domains. For example, PCA fails to detect the adipose tissue region and produces noisier results for regions surrounding invasive cancer cells. Predictive PCA and RapPCA both mis-classify part of the cancer in situ region as invasive cancer, where predictive PCA results are noisier for the top left region, and RapPCA shows larger uncertainty regarding adipose tissue. Spatial PCA, on the other hand, infers part of the connective tissue to be cancerous.

To make a more in-depth comparison, we examine the breakdown of clustering accuracy by ground truth labels, in other words, whether or not each approach correctly classifies the spots belonging to each spatial domain, with the undetermined cluster removed from comparison. Fig-

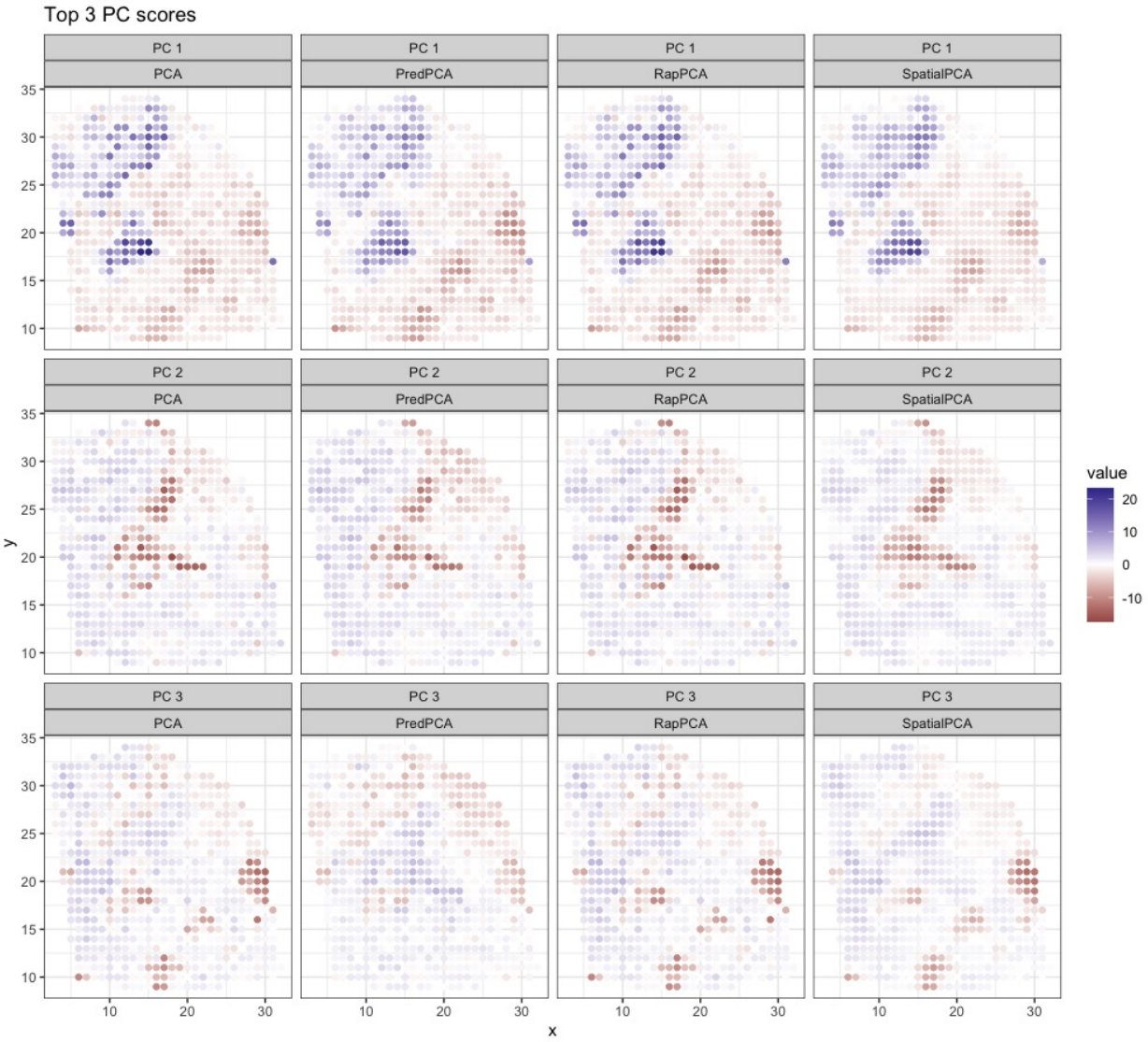


Figure 3.5: Top 3 PC scores based on gene expression in the HER2-positive breast tumor data

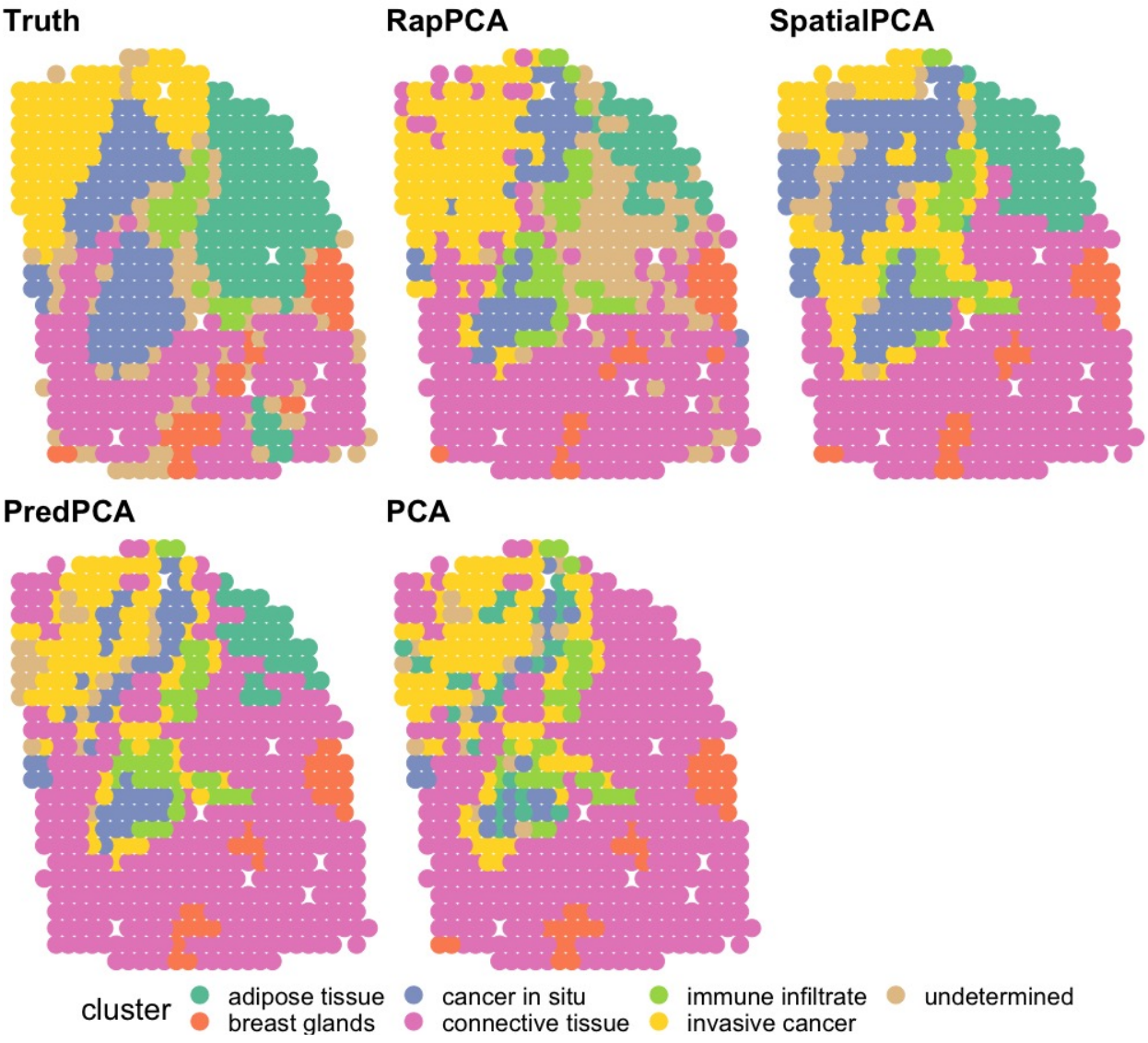


Figure 3.6: Detected spatial domains and annotated ground truth

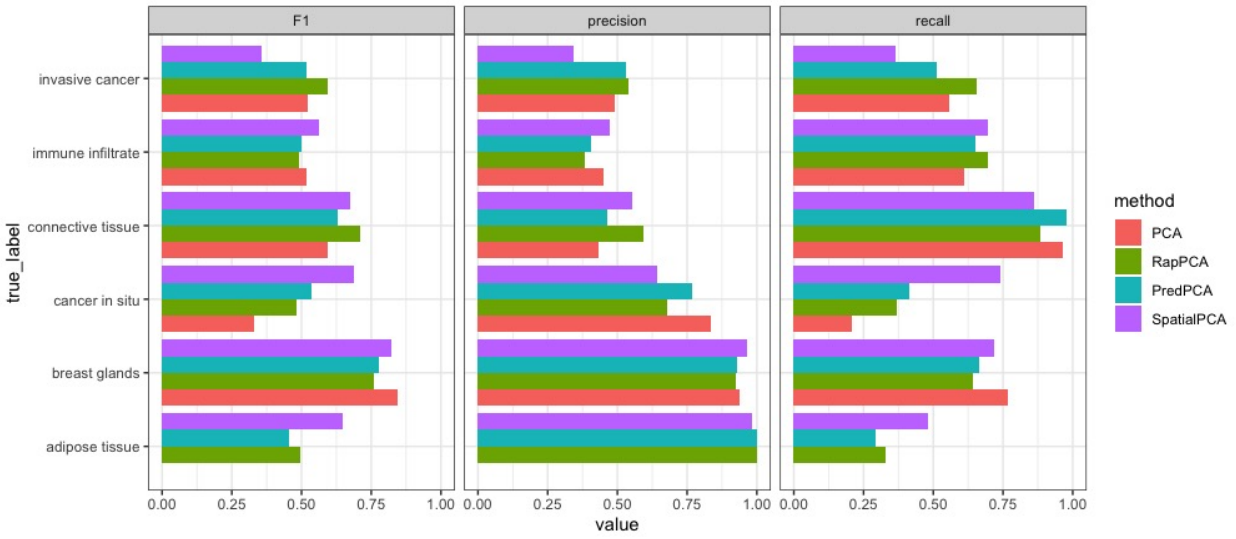


Figure 3.7: Breakdown of domain detection accuracy by true label. Note that the metrics for the adipose tissue region are not well-defined for PCA because it fails to detect any spot in this region.

ure 3.7 presents this breakdown for each method in terms of precision, recall and F1 score as a composite measure. As observed from the spatial domain plots, the relative accuracy for each method differs by region, and in particular, RapPCA has the best performance in both precision and recall on the identification of invasive cancer regions, whereas spatial PCA shows an advantage in recall for cancer in situ.

3.6 Discussion

When dimension reduction is conducted as an intermediate step before a spatial modeling task of interest, there are two typically conflicting considerations guiding the choice of approaches. The more obvious one is how closely the lower dimensional score represents the original variables, which we refer to as representability. Another important aspect is the predictability of the resulting scores, i.e., how well they could be expressed or predicted by available covariates and/or spatial effects.

We discussed how existing dimension reduction algorithms fit into this general framework of optimizing either criterion, and proposed a flexible interpolation between them that achieves the optimal representability-predictability trade-off. Our proposal, called RapPCA, can also handle high-dimensional covariates and non-linear relationships. Simulation studies under different sce-

narios illustrate the gain in downstream prediction accuracy by our method, which also achieves smaller overall errors when both predictability and representability are taken into account. Applications to real datasets in different scientific areas further demonstrate the utility of our proposed method, even in analytic tasks that do not explicitly involve prediction.

The balance between prediction and representation performance can be viewed as a generalized form of bias-variance trade-off. While methods such as classical PCA minimize the representation gap specifically for the training data at hand, such representation may capture excessive noise if it explains a large amount of variability in the data. In contrast, the formulation of predictive PCA restricts the PC scores to fall within a certain model space (e.g. the linear span of the covariates), which is a form of regularization enforcing the smoothness of the PCs. Many probabilistic dimension reduction approaches implicitly address both aspects, while our proposal seeks the optimal balance in an explicit and interpretable way.

While we used TMSE to guide the choice of tuning parameters in our examples, such criteria can be driven by the specific analytic goal in practice. Researchers could examine the trend of prediction, representation and total errors by each tuning parameter, and in particular γ in (3.3), on a set of test data, and choose the combination leading to a desired trade-off. Our empirical evaluations suggest that striking such balance with our proposed method would most often improve, or at least maintain similar, overall errors compared to existing alternatives.

Though Theorem 3.1 establishes an exact solution to the optimization problem (3.3) that in theory does not contain numerical errors, computational inaccuracy may occur in scenarios with large sample size or high dimensionality, since the quality of eigen decomposition could suffer when the condition number of the matrix involved is large. Alternative numerical techniques could improve the accuracy and stability of the proposed dimension reduction procedure.

Chapter 4

A PENALIZED POISSON LIKELIHOOD APPROACH TO HIGH-DIMENSIONAL SEMI-PARAMETRIC INFERENCE FOR DOUBLY-STOCHASTIC POINT PROCESSES

4.1 Introduction

Spatial point process models (Chiu et al., 2013; Diggle, 2003; Illian et al., 2008; Møller and Waagepetersen, 2003) are used in many application areas to capture observed patterns of events over a region. Examples include modeling disease prevalence in epidemiology (Best et al., 2005; Franch-Pardo et al., 2020), crime incidence in sociology, (Ferreira et al., 2012; Leong and Sung, 2015) and species abundance in ecology (Law et al., 2009; Renner et al., 2015). Two key features of observed events in these and many other applications are *spatial heterogeneity* and *spatial correlation* (Anselin, 1988; Plotkin et al., 2000; Vinatier et al., 2011). Spatial heterogeneity refers to the variation of the underlying intensity of events across the space, which may come from individual characteristics (captured by covariates) and/or purely spatial effects (variation in baseline intensity). It is often described by first-order properties (e.g., the intensity function) of a spatial point process. Spatial correlation, on the other hand, reflects the similarity of event rates in close-by areas, which is captured by second-order properties of the underlying process.

A wide range of point process models are variants of Poisson processes that capture one or both of these characteristics. Møller and Waagepetersen (2007, 2017) provide a review of modeling choices and computational methods for spatial point processes. As a basis for more flexible models, consistency and asymptotic normality of maximum likelihood estimates (MLE) for Poisson processes are studied by Brillinger (1975) and Rathbun and Cressie (1994). Jensen (1993) and Dereudre and Lavancier (2017) investigate asymptotic properties of the MLE for Gibbs point process models, which allow for aggregation—or, positive spatial correlation—and repulsion—or, negative spatial correlation—via interactions between points. The specification of the Gibbs point process via an interaction function leads to analytically intractable intensities (Baddeley and Nair, 2012), giving rise to challenges for simulation and theoretical analysis. For example, the theories in Jensen (1993)

rely on strong restrictions on the parameter space, while Dereudre and Lavancier (2017) do not discuss statistical inference.

Another class of models, namely, doubly-stochastic Poisson processes, also known as Cox processes, (Cox, 1955), specify random intensity functions for conditionally Poisson processes, that, in turn, capture spatial correlation. As suggested by Møller and Waagepetersen (2007), this approach provides more flexibility by separate modeling of the first order heterogeneity and spatial correlation. Møller et al. (1998) and Diggle et al. (2013) provide overviews of log-Gaussian Cox processes (LGCP)—which are conditionally Poisson processes depending on the realization of a Gaussian random field—and related approaches to inference, including moment-based, likelihood-based and Bayesian methods. Moment-based methods, such as minimal contrast estimation (e.g. Diggle, 2003; Møller and Waagepetersen, 2003), minimize the discrepancy between theoretical and empirical summary statistics of the process. These methods are computationally simple but rely on somewhat arbitrary specification of a tuning parameter. General statistical theory on properties of such estimators is also lacking (Cressie, 2015). Furthermore, as noted in Møller and Waagepetersen (2003) and Guan (2006), there is in general no closed form for the likelihood of a Cox process, and the unobserved, infinite-dimensional random intensity needs to be approximated by truncation or discretization.

The above limitations make maximum likelihood estimation of Cox processes computationally challenging. One alternative is to conduct Bayesian inference under discretization (Møller and Waagepetersen, 2003; Møller et al., 1998); see also Teng et al. (2017) for a review of related approximation methods. Waagepetersen (2004) discusses the convergence of posterior for LGCPs under discretization when the cell sizes tend to zero. In general, Markov chain Monte Carlo (MCMC) computation for the posterior, without any additional approximation, is time consuming for moderate sample sizes (see Sections 4.4 and 4.5), while limited theoretical guarantees are available for computationally-tractable approximations, such as variational Bayes and integrated nested Laplace approximation (INLA) (Rue et al., 2009). Wang and Blei (2019) present general results for variational approximation and show that the variational Bayes posterior converges to the Kullback-Leibler minimizer of a normal distribution centered at the truth; however, variational Bayes optimization is typically non-convex and the optimization loss surface is not well characterized. Simpson et al. (2016) propose a basis function approximation of the random field underlying

the LGCP, and show the convergence of such an approximation as well as the discrete approximation of the likelihood. However, the convergence of the full posterior, which is needed for inference, remains to be investigated.

In this chapter, we focus on doubly stochastic spatial models. Current approaches and theories, if any, generally rely on stationarity or a parametric form, or at least a known second-order intensity function, of the latent process. Table C.1 in Appendix C.1 provides a summary of related models and their limitations. For instance, under the frequentist paradigm, Guan (2006) proposes a composite likelihood method (Lindsay, 1988) for parameters in stationary spatial point processes with consistency and asymptotic normality guarantees; however, the generalization of this framework to non-stationary settings relies on knowledge of the second-order properties of the process. Guan (2008) develops a nonparametric estimation method and establishes its consistency for inhomogeneous point processes, but the method does not handle inference for covariate effects. Schoenberg (2005) advocates the use of the Poisson likelihood or weighted sum of squares as estimating functions for covariate effects, and shows the consistency of the resulting estimator even for non-Poisson data; however, inference for such estimates is not investigated. Waagepetersen (2007) suggests a two-step estimation procedure for both the covariate effects and clustering parameters of inhomogeneous Neyman-Scott processes, and proves the asymptotic normality of the former. Waagepetersen and Guan (2009) propose a two-step procedure that leads to asymptotically normal estimates for the covariate effects along with correlation parameters. Dvořák et al. (2019) extend composite likelihood methods to non-stationary settings by applying a three-step procedure, but without investigating the theoretical properties of this approach.

Our goal is to generalize the applicability of existing approaches and theoretical analyses, which generally rely on stationarity or restrictive parametric forms, or at least a known second-order intensity function, of the latent process. These assumptions may not be straightforward to test or justify in practice, and therefore, more flexible methods with less stringent assumptions are greatly needed for spatial point processes. Moreover, existing approaches do not facilitate estimation and inference in high-dimensional covariate settings, which is an increasingly common scenario in practice given the development of data collection techniques such as geographic information systems (GIS) (Cai and Maiti, 2020; Gonella et al., 2022). To address these needs, we develop a penalized estimation and inference framework for semi-parametric, non-stationary Cox processes with high-

dimensional covariates. In addition to appealing theoretical properties, the proposed method also offers significant computational advantages over existing approaches.

Key to our proposal, presented in Section 4.2, is a discretization of the observation window, which allows us to adopt the idea of Poisson maximum likelihood estimation (PMLE)—as in Schoenberg (2005)—to explicitly model the realization of the random intensity function together with potentially high-dimensional covariate effects. We justify this discretization by showing that consistent estimates and valid inferences for high-dimensional parameters corresponding to model covariates can be obtained despite the misspecification of the random intensity through discretization and the fact that the random field is ignored in the Poisson likelihood. Building on this observation, in Section 4.3 we establish the consistency of the regression parameter estimates under less restrictive assumptions than existing approaches. These results are obtained without assuming a parametric distribution or stationarity for the random component of the intensity. We also establish the asymptotic normality of de-biased estimates of regression parameters under a few additional assumptions. Performance of our approach is illustrated and compared with common Bayesian approaches via a simulation study in Section 4.4, as well as an application to Seattle crime data in Section 4.5.

4.2 Penalized Poisson Maximum Likelihood Estimation (PMLE)

4.2.1 Model

Consider a Cox process $\mathcal{Y}(s) : s \in \Omega$ over an observation window Ω. That is, $\mathcal{Y}(s)$ is an inhomogeneous Poisson process with intensity $\lambda(s)$, which is a realization of the random intensity $\Lambda(s)$ modeled as

$$\log \Lambda(s) = \log P(s) + \alpha^0(s) + X(s)\boldsymbol{\beta}^0 + \varepsilon(s), \tag{4.1}$$

where $P(s)$ is the offset, $\alpha^0(s)$ is the baseline intensity, X(s) is a p-dimensional vector-valued function representing the distribution of p covariates over Ω; here, $\boldsymbol{\beta}^0 \in \mathbb{R}^p$ denote the true parameters of interest, and $\varepsilon(s)$ is a mean zero, latent random field of errors. For example, if $\varepsilon(s)$ is a Gaussian random field, then (4.1) corresponds to a LGCP.

Schoenberg (2005) shows that maximizing a Poisson log-likelihood for certain low-dimensional, parametric, non-Poisson point processes leads to consistent parameter estimates. Adapting this idea

to Cox processes with high-dimensional covariates would greatly simplify the optmization problem which would otherwise have a less tractable form. Following Schoenberg (2005), we denote by $\lambda(\cdot)$ the *conditional intensity*, and refer to its expectation, $\mathbb{E}_0\left[\lambda(\cdot)\right]$, taken pointwise with respect to the data generating mechanism giving rise to $\varepsilon(\cdot)$, the *unconditional intensity*. By Fubini's Theorem—which holds under conditions discussed in Section 4.3.1—the unconditional intensity at any location $s \in \Omega$ is determined by the moment generating function of $\varepsilon(s)$ via

$$\begin{aligned}\mathbb{E}_0\left[\lambda(s)\right] = \mathbb{E}_0 \int_{\Omega_i} P(s) \exp\left[\alpha^0(s) + X(s)\boldsymbol{\beta}^0 + \varepsilon(s)\right] \, \mathrm{d}s &= \int_{\Omega_i} P(s) \exp\left[\alpha^0(s) + X(s)\boldsymbol{\beta}^0\right] \mathbb{E}_0\left[\exp \varepsilon(s)\right] \, \mathrm{d}s \\ &:= \int_{\Omega_i} P(s) \exp\left[\alpha^0(s) + X(s)\boldsymbol{\beta}^0 + \phi(s)\right] \mathrm{d}s,\end{aligned} \tag{4.2}$$

where $\phi(s) = \log \mathbb{E}_0[\exp \varepsilon(s)]$. In Section 4.3.1 we shall see that the unconditional intensity is a key quantity for establishing the relationship between the simple Poisson log-likelihood and parameters underlying a more complex Cox process model.

In practice, even when the data arise from a spatially continuous point process, it is common that the events are discretely observed as counts aggregated over small regions; see, e.g., Li et al. (2012) and Taylor et al. (2018) for additional examples and discussion. Likewise, the offset and covariates are also commonly observed as, and (perhaps implicitly) assumed to be, piecewise constant where each small region is associated with a common value. This is specially the case for many epidemiological studies of disease prevalence, where the resolution of observations is constrained by confidentiality issues, as well as analyses leveraging both spatial and non-spatial, individual-level data (see, e.g., the example in Section 4.5, and Diggle et al., 2010). A realistic approach, therefore, would be to assume continuous $\alpha^0(\cdot)$ and $\varepsilon(\cdot)$, while treating the discretely observed quantities $P(\cdot)$ and $X(\cdot)$ as *piecewise constant* based on the discretization for which data is available. For example, $P(\cdot)$ and $X(\cdot)$ may be aggregated by census tract or zip code if they are obtained from census data. When $\mathcal{Y}(\cdot)$, $P(\cdot)$ and $X(\cdot)$ are observed with different resolutions, we can simply take the finest partition available.

Under a discretization $\Omega = \Omega_1 \cup \ldots \cup \Omega_n$, we assume that the observed data is generated by

$$\begin{aligned} Y_i \mid \lambda_i &\sim \text{Poisson}(\lambda_i),\ i = 1, \ldots, n \\ \lambda_i \mid X_i, \varepsilon(\cdot) &= P_i \exp\left(X_i \boldsymbol{\beta}^0\right) \int_{\Omega_i} \exp\left[\alpha^0(s) + \varepsilon(s)\right]\ \mathrm{d}s, \end{aligned} \tag{4.3}$$

where Y_i is the case count within Ω_i, $\alpha^0(\cdot)$ and $\varepsilon(\cdot)$ are the same as in (4.1) and $X_i \in \mathbb{R}^p$ and P_i are the covariate values and offset shared by all locations within Ω_i. To account for potential non-stationarity, we aim to conduct estimation and inference on the regression parameters $\boldsymbol{\beta}^0 \in \mathbb{R}^p$ based on observed $\mathbf{Y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\mathbf{P} \in \mathbb{R}^n$ under minimal assumptions on the latent random field $\varepsilon(\cdot)$, while allowing for flexibility in the unknown baseline $\alpha^0(\cdot)$.

To describe our PMLE approach, suppose first that data are generated from a Poisson process with $\varepsilon(s) \equiv 0$ on Ω. Our first idea is to approximate the true intensity function $\alpha^0(\cdot)$ with a piecewise constant function, which takes constant values at each discretized region and is thus expressed as n-dimensional vector $\tilde{\boldsymbol{\alpha}} \in \mathbb{R}^n$; the notation $\tilde{\boldsymbol{\alpha}}$ underscores the use of a vector resulting from discretization as apposed to the true baseline intensity function, $\alpha^0(s)$. Then, using the same discretization for $\mathbf{X}$ and $\mathbf{Y}$, we obtain the simple Poisson log-likelihood

$$\ell(\tilde{\boldsymbol{\alpha}}, \boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{n} y_i \left(\log P_i + \tilde{\alpha}_i + X_i \boldsymbol{\beta}\right) - \sum_{i=1}^{n} |\Omega_i| P_i \exp\left(\tilde{\alpha}_i + X_i \boldsymbol{\beta}\right), \tag{4.4}$$

where $|\Omega_i|$ is the area of Ω_i. This formulation is similar to a mixed effects Poisson model, except that we do not impose a parametric assumption on the distribution of the random effects $\tilde{\alpha}_i$, but instead model their realization as (region-specific) fixed parameters, allowing for valid estimation for a broader class of point processes. The second idea in our PMLE approach is to ignore the latent random field $\varepsilon(s)$, and approximate the Cox process with the Poisson process specified in (4.4). This approximation is motivated by Schoenberg (2005), which we extend to high-dimensional and semi-parametric Cox processes in Section 4.3.1. In particular, we show that the gradient of (4.4) yields a valid estimating equation for $\boldsymbol{\beta}$ despite the misspecified discrete form of $\tilde{\boldsymbol{\alpha}}$ and the ignored randomness arising from $\varepsilon(\cdot)$.

Due to the high dimensionality of both $\tilde{\boldsymbol{\alpha}}$ and $\boldsymbol{\beta}$, we impose penalties on these parameters to ensure identifiability for this over-parametrized model. We impose an ℓ_1 sparsity penalty on $\boldsymbol{\beta}$ (Tibshirani, 1996), and an additional ℓ_1 (Tibshirani et al., 2005) or ℓ_2 (Li et al., 2019; Zhao and Shojaie, 2016a) fusion penalty on $\tilde{\boldsymbol{\alpha}}$. More specifically, the partition of Ω induces a graph

$\mathcal{G}_n = (V_n, E_n)$, where the set of vertices $V_n = \{\Omega_1, \ldots, \Omega_n\}$ correspond to the small regions under such partition, and the set of edges $E_n \subseteq V_n \times V_n$ consists of unordered pairs (Ω_i, Ω_j) such that Ω_i and Ω_j are adjacent. Given the spatially continuous nature of $\mathcal{Y}$, the edges of graph $\mathcal{G}_n$ could further be weighted by distances between centroids, or other notions of (dis)similarity, of adjacent regions.

Let W_n be the weighted or unweighted adjacency matrix and $D_n = \text{diag}(d_1, \ldots, d_n)$ where $d_i = \sum_{j \in V_n} w_{ij}$. The edge incidence matrix $B_n \in \mathbb{R}^{|E_n| \times |V_n|}$ is defined such that its kth row corresponds to the kth edge of $\mathcal{G}_n$, say (Ω_i, Ω_j) where $i < j$, given by $b_{ki} = \sqrt{w_{ij}}$ and $b_{kj} = -\sqrt{w_{ij}}$. The graph Laplacian (Chung, 1997), $L_n = D_n - W_n$, satisfies $L_n = B_n^\top B_n$. It can be seen that $L_n \mathbf{1} = 0$, where $\mathbf{1}$ is a vector of all ones. The singularity of L_n could bring numerical instability to the optimization. As proposed by Li et al. (2019), we replace L_n with $\tilde{L}_n := L_n + \delta I_n$ where δ is a small positive constant and I_n is the identity matrix. The fusion penalty term for $\tilde{\boldsymbol{\alpha}}$ then takes the form

$$R(\tilde{\boldsymbol{\alpha}}; \mathcal{G}_n) = \begin{cases} \|B_n \tilde{\boldsymbol{\alpha}}\|_1 = \sum_{(\Omega_i, \Omega_j) \in E_n} \sqrt{w_{ij}} |\tilde{\alpha}_i - \tilde{\alpha}_j| & (\ell_1) \\ \frac{1}{2} \tilde{\boldsymbol{\alpha}}^\top \tilde{L}_n \tilde{\boldsymbol{\alpha}} = \frac{1}{2} \sum_{(\Omega_i, \Omega_j) \in E_n} w_{ij} (\tilde{\alpha}_i - \tilde{\alpha}_j)^2 + \frac{\delta}{2} \sum_{i=1}^n \tilde{\alpha}_i^2 & (\ell_2) \end{cases}.$$

The ℓ_1 fusion penalty is a form of generalized Lasso penalty (Tibshirani and Taylor, 2011) and encourages a piecewise constant baseline intensity surface where most connected regions have exactly equal $\tilde{\alpha}$'s. The ℓ_2 fusion penalty, on the other hand, encourages the baseline intensities between connected regions to be similar, but not exactly equal.

The penalized PMLE is given by the solution to the optimization problem

$$\hat{\boldsymbol{\theta}} := \left(\hat{\boldsymbol{\alpha}}^\top, \hat{\boldsymbol{\beta}}^\top \right)^\top = \underset{\boldsymbol{\theta} := (\tilde{\boldsymbol{\alpha}}, \boldsymbol{\beta})}{\text{argmin}} -\ell(\tilde{\boldsymbol{\alpha}}, \boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{Y}) + \gamma_n R(\tilde{\boldsymbol{\alpha}}; \mathcal{G}_n) + \tau_n \|\boldsymbol{\beta}\|_1, \tag{4.5}$$

where γ_n and τ_n are tuning parameters to be determined, for example, via cross-validation. Strategies for prediction and cross-validation in the presence of dependence between regions are discussed in Section 4.2.3.

The penalized PMLE is related to Bayesian spatial models with intrinsic conditional autoregressive (ICAR) priors, first introduced by Besag (1974). For instance, under an ℓ_2 fusion penalty, the PMLE is similar to the maximum a posteriori (MAP) estimate of the Besag-York-Mollié (BYM) model (Besag et al., 1991) which specifies a pair of random effects per region. The first set of

random effects captures spatially correlated errors from a Gaussian Markov random field (GMRF): $\tilde{\alpha}_i \mid \tilde{\alpha}_{-i} \sim N(\sum_{j\sim i} \tilde{\alpha}_j/d_i, \sigma^2/d_i)$ with d_i being the number of neighbors of region i, and $j \sim i$ indicating that regions i and j are connected. The second set reflects non-spatial heterogeneity and are modeled as independent normal random effects. In particular, the GMRF prior takes a similar quadratic form in the posterior distribution of $(\tilde{\boldsymbol{\alpha}}, \boldsymbol{\beta})$ as our ℓ_2 fusion penalty in the objective function (4.5).

4.2.2 Computation

We start our discussion of computational algorithms with the ℓ_2 fusion penalty. Defining the soft-thresholding operator $S_\tau(x) := \text{sign}(x)\max\{|x| - \tau, 0\}$, the optimization problem can be solved by a proximal gradient descent algorithm; see Algorithm 2. The step size is set adaptively via backtracking line search (Armijo, 1966; Boyd et al., 2004). Lines 2 through 11 in Algorithm 2 can be replaced by coordinate-wise gradient descent, where $\tilde{\boldsymbol{\alpha}}$ and $\boldsymbol{\beta}$ are optimized iteratively, instead of jointly. This could make the tuning of γ_n, τ_n more efficient.

With the ℓ_1 fusion penalty, $R(\tilde{\boldsymbol{\alpha}}; \mathcal{G}_n)$ is nonseparable with respect to $\tilde{\boldsymbol{\alpha}}$. This nonseparability introduces challenges in optimization for nonlinear models, such as the Poisson model. To overcome these challenges, we follow the proposal of Chen et al. (2012) and adopt a smooth ℓ_∞ approximation for the ℓ_1 fusion penalty,

$$\gamma_n \|B_n^\top \tilde{\boldsymbol{\alpha}}\|_1 \approx h_\xi(\tilde{\boldsymbol{\alpha}}) := \gamma_n \max_{\|\nu\|_\infty \le 1} \left[\nu^\top B_n \tilde{\boldsymbol{\alpha}} - \frac{\xi}{2}\|\nu\|_2^2 \right]. \tag{4.6}$$

The parameter ξ controls the amount of smooth relaxation to the original problem, with $\xi = 0$ recovering the original ℓ_1 fusion penalty. The gradient of h_ξ can simply be calculated as

$$\nabla h_\xi(\tilde{\boldsymbol{\alpha}}) = B_n^\top S_\infty\left(\frac{\gamma_n B_n \tilde{\boldsymbol{\alpha}}}{\xi}\right),$$

where $S_\infty(\cdot)$ is the element-wise projection operator onto the ℓ_∞ ball:

$$S_\infty(x) = \begin{cases} -1, & x \le -1 \\ x, & -1 < x \le 1 \\ 1, & x > 1 \end{cases}.$$

Algorithm 2: Proximal gradient descent for penalized PMLE

Set tolerance *tol* as well as (small) positive constants a, b for backtracking line search.

Initialize $\boldsymbol{\theta}^{(0)} = (\tilde{\boldsymbol{\alpha}}^{(0)}, \boldsymbol{\beta}^{(0)})$ and calculate the objective function

$$f(\boldsymbol{\theta}^{(0)}) = \mathcal{L}(\boldsymbol{\theta}^{(0)}) + \tau_n \|\boldsymbol{\beta}^{(0)}\|_1 := -\ell(\boldsymbol{\theta}^{(0)}) + \gamma_n R(\tilde{\boldsymbol{\alpha}}^{(0)}) + \tau_n \|\boldsymbol{\beta}^{(0)}\|_1$$

for $t = 0, 1, \ldots$ *until convergence* **do**

Evaluate the gradient $\nabla\mathcal{L}(\boldsymbol{\theta}^{(t)}) := -\nabla\ell(\boldsymbol{\theta}^{(t)}) + \gamma_n \tilde{L}_n \tilde{\boldsymbol{\alpha}}^{(t)}$

Line search: set the initial step size $\eta^{(t)} := 1$

while $\mathcal{L}\left(\boldsymbol{\theta}^{(t)} - \eta^{(t)} \nabla\mathcal{L}(\boldsymbol{\theta}^{(t)})\right) - \mathcal{L}(\boldsymbol{\theta}^{(t)}) \geq -a\|\boldsymbol{\theta}^{(t)}\|_2^2$ **do**

$\eta^{(t)} \leftarrow b\eta^{(t)}$

end

Gradient step: $\boldsymbol{\theta}^{\dagger} := (\tilde{\boldsymbol{\alpha}}^{\dagger}, \boldsymbol{\beta}^{\dagger}) \leftarrow \boldsymbol{\theta}^{(t)} - \eta^{(t)} \nabla\mathcal{L}(\boldsymbol{\theta}^{(t)})$

Proximal step: $\boldsymbol{\theta}^{(t+1)} \leftarrow \left(\tilde{\boldsymbol{\alpha}}^{\dagger}, S_\tau(\boldsymbol{\beta}^{\dagger})\right)$ where $S_\tau(\cdot)$ is applied element-wise on $\boldsymbol{\beta}^{\dagger}$

Convergence criterion: Calculate $f(\boldsymbol{\theta}^{(t+1)})$ and convergence is achieved if $\left|f(\boldsymbol{\theta}^{(t+1)}) - f(\boldsymbol{\theta}^{(t)})\right| < tol \cdot \left|f(\boldsymbol{\theta}^{(t)})\right|$

end

Result: Output $\boldsymbol{\theta}^{(t+1)}$

Incorporating the smooth approximation (4.6) into the optimization leads to a slightly modified version of Algorithm 2 where we replace $\gamma_n R(\tilde{\boldsymbol{\alpha}})$ with $h_\xi(\tilde{\boldsymbol{\alpha}})$, and likewise for the corresponding gradients. Chen et al. (2012) show that with $\xi = \epsilon/|E_n|$, the approximation gap $|\gamma_n \|B_n \tilde{\boldsymbol{\alpha}}\|_1 - h_\xi(\tilde{\boldsymbol{\alpha}})| \leq \epsilon$ is guaranteed within $O(\sqrt{|E_n|}/\epsilon)$ iterations.

4.2.3 Prediction

Making out-of-sample predictions is of interest when the goal is to learn about new regions with newly observed data or areas in which data are missing, or to evaluate the model's performance, e.g., in cross-validation. Because $\alpha(\cdot)$ is approximated with discretized region-specific baselines $\tilde{\boldsymbol{\alpha}}$, predicted individual baselines are required for such task. To obtain such predictions, we use the ℓ_2

cohesion approach of Li et al. (2019). Suppose there are n_1 training and n_2 test samples, and the Laplacian of the entire graph connecting the $n := n_1 + n_2$ regions is rearranged and partitioned as

$$L_n = \begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix},$$

where L_{11} and L_{22} correspond to the training and test samples respectively. Likewise, $\tilde{\boldsymbol{\alpha}}$ is partitioned as $(\tilde{\boldsymbol{\alpha}}_1, \tilde{\boldsymbol{\alpha}}_2)$. Setting $\tilde{\boldsymbol{\alpha}}_1$ to its estimate $\hat{\boldsymbol{\alpha}}_1$ obtained from model-fitting, $\tilde{\boldsymbol{\alpha}}_2$ can be predicted via

$$\hat{\boldsymbol{\alpha}}_2 = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \left(\hat{\boldsymbol{\alpha}}_1, \boldsymbol{\alpha}\right)^\top L_n \left(\hat{\boldsymbol{\alpha}}_1, \boldsymbol{\alpha}\right) = -L_{22}^{-1} L_{21} \hat{\boldsymbol{\alpha}}_1;$$

observe that when regions in the training and test sets are not connected, $\tilde{\boldsymbol{\alpha}}_2$ is predicted to be 0.

As noted by Li et al. (2019), it may not be straightforward and fully justified to split dependent samples from a connected graph into training and test sets. However, Li et al. (2019) find that in practice the described procedure performs reasonably well for cross-validation.

4.3 Theoretical Guarantees

In this section, we establish theoretical properties of the penalized PMLE with ℓ_1 sparsity and ℓ_1 or ℓ_2 fusion penalties given in (4.5). However, before focusing on the sparsity or fusion penalties, we will first discuss the relationship between the target parameter, i.e., the minimizer of the expected negative Poisson log-likelihood $-\mathbb{P}_0 \ell(\tilde{\boldsymbol{\alpha}}, \boldsymbol{\beta})$, and the true slope parameter $\boldsymbol{\beta}^0$ along with the intensity function $\alpha^0(\cdot)$ underlying the Cox process. In particular, we show that the Poisson likelihood yields an unbiased estimating equation for $\boldsymbol{\beta}^0$ despite the ignored error random field and the misspecification of $\alpha^0(\cdot)$. We then use empirical process arguments to show the convergence of the penalized PMLE to the target parameters. Furthermore, we define a de-biased estimator of $\boldsymbol{\beta}^0$, establish its asymptotic normality and provide an estimate for the covariance, accounting for the doubly stochastic nature of the process not explicitly captured by the PMLE. We end by deriving the asymptotic distribution of the de-biased estimator, providing a valid statistical inference procedure for $\boldsymbol{\beta}^0$.

4.3.1 Consistency

The discussion of consistency for spatial processes relies on the specification of an asymptotic regime. While the definition of an "increasing n" scenario may be straightforward under independent sampling, there are multiple asymptotic regimes for spatial data under which the same estimator could have drastically different behaviors, as noted by Stein (1999) and Zhang and Zimmerman (2005). For clarity, we define the asymptotic regime of interest below. This notion is related to the classical increasing domain asymptotics in the spatial literature.

Definition 4.1 (Asymptotic regime)**.** *Let the observation window Ω be implicitly indexed by n, and let its size $|\Omega| \to \infty$ as $n \to \infty$. The partition $\Omega = \Omega_1 \cup \ldots \cup \Omega_n$ satisfies $0 < a_0 \leq \liminf_{n\to\infty} \min_{i=1,\ldots,n} |\Omega_i| \leq \limsup_{n\to\infty} \max_{i=1,\ldots,n} |\Omega_i| \leq A_0 < \infty$ and the offset satisfies $0 < p_0 \leq \liminf_{n\to\infty} \min_{i=1,\ldots,n} P_i \leq \limsup_{n\to\infty} \max_{i=1,\ldots,n} P_i \leq P_0 < \infty$, where a_0, A_0, p_0, P_0 are constants not depending on n.*

In words, the observation window expands and incorporates new, unobserved regions as n grows. Correspondingly, the partition includes more and more regions, while maintaining a constant rate of granularity. This requirement is not restrictive given that we allow $\alpha^0(\cdot)$ and $\varepsilon(\cdot)$ to be non-constant within each cell. Note that the domain of $\alpha^0(\cdot)$ and $\varepsilon(\cdot)$, the range of region-specific covariates $\boldsymbol{X}$, and the graph $\mathcal{G}_n = (V_n, E_n)$ induced by the partition all depend on Ω and n. Requirements on their behavior as n increases are stated under our full set of assumptions for consistency, which we now present along with some interpretations.

Assumption 4.1 (Regularity conditions)**.**

i) The partition $\Omega = \Omega_1 \cup \ldots \cup \Omega_n$ is such that each Ω_i is bounded and connected, and the true baseline function $\alpha^0(\cdot)$ is continuous on each Ω_i.

ii) The function $\phi(s) := \log \mathbb{E}_0 [\exp \varepsilon(s)]$ as defined in (4.2) is continuous on each Ω_i.

iii) Let $\mathcal{F}$ be a σ-algebra over Ω, $\mu(\cdot)$ be a measure (e.g. the Lebesgue measure) defined on $(\Omega, \mathcal{F})$ and $\mathbb{P}_\varepsilon$ be the probability measure of the random field $\varepsilon(\cdot)$ defined on $(\Omega_\varepsilon, \mathcal{F}_\varepsilon)$. Then there exists a product measure $\rho(\cdot)$ on $(\Omega \times \Omega_\varepsilon, \mathcal{F} \times \mathcal{F}_\varepsilon)$ such that for every $A \in \mathcal{F}$ and $A_\varepsilon \in \mathcal{F}_\varepsilon$,

$\rho(A \times A_\varepsilon) = \mu(A)\mathbb{P}_\varepsilon(A_\varepsilon)$. *We assume that*

$$\limsup_{n\to\infty} \max_{i=1,\ldots,n} \int_{\Omega_i \times \Omega_\varepsilon} P_i \exp\left[\alpha^0(s) + X_i\boldsymbol{\beta}^0 + \varepsilon(s)\right] d\rho(s,\varepsilon) < \infty.$$

Condition iii) of Assumption 4.1 enables the application of Fubini's Theorem over each Ω_i, so that we only need to learn about some functionals of the error random field evaluated in a pointwise manner, without explicitly handling the integral involving the realization of $\varepsilon(\cdot)$. Combined with conditions i) and ii), iii) further guarantees the existence of one point within each Ω_i at which the local unconditional intensity given by (4.2) is representative of the average regional mean. This ensures the convergence of the discretized solution to some summary statistics for the continuous function within each region. However, we still need the magnitude of penalty terms to scale appropriately in order to complete this argument, as stated in the next assumption for both ℓ_1 and ℓ_2 fusion penalties.

Assumption 4.2 (Rates of tuning parameters)**.** *For any set of n locations* $\boldsymbol{s} := (s_1, \ldots, s_n) \in \Omega$*, denote the vectorized form of the true intensity* $\alpha^0(\cdot)$ *as* $\tilde{\boldsymbol{\alpha}}^0(\boldsymbol{s}) = (\alpha^0(s_1), \ldots, \alpha^0(s_n)) \in \mathbb{R}^n$*. Also, let* $\boldsymbol{\alpha}^\dagger(\boldsymbol{s}) := \tilde{\boldsymbol{\alpha}}^0(\boldsymbol{s}) + \phi(\boldsymbol{s})$ *for* ϕ *defined in Assumption 4.1. Then,*

i) $\tau_n = O_P\left(\sqrt{\frac{\log p}{n}}\right)$*;*

ii) under the partition $\Omega = \Omega_1 \cup \ldots \cup \Omega_n$*, we have, for the* ℓ_2 *smoothing penalty,*

$$\gamma_n \sup_{\boldsymbol{s}: s_1 \in \Omega_1, \ldots, s_n \in \Omega_n} \|\tilde{L}_n \boldsymbol{\alpha}^\dagger(\boldsymbol{s})\|_2 = O_p(n^c),$$

where $c \in (0, 1/2)$*, and we recall* $\tilde{L}_n = L_n + \delta I_n$*; alternatively,* $\gamma_n^2 \max_i d_i = o_P(1)$ *and*

$$\gamma_n^2 \sup_{\boldsymbol{s}: s_1 \in \Omega_1, \ldots, s_n \in \Omega_n} \|B_n \boldsymbol{\alpha}^\dagger(\boldsymbol{s})\|_1 = O_P(n^c)$$

for the ℓ_1 *fusion penalty;*

The rate in Assumption 4.2 i) is common in high-dimensional estimation literature (Hastie et al., 2019; Negahban et al., 2012). Condition ii) reflects that the fusion penalty for $\tilde{\boldsymbol{\alpha}}$ takes into account the similarity of both $\alpha^0(\cdot)$, the baseline intensity, and $\phi(\cdot)$, the error random field, between close-by

regions. Such penalty, however, need not be fully informative, and our belief on the closeness of $\tilde{\boldsymbol{\alpha}}^0$ between connected regions imposed by $R(\tilde{\boldsymbol{\alpha}}; \mathcal{G}_n)$ need not align perfectly with the truth. When such similarity does exist in the true data generating mechanism, $\|B_n \boldsymbol{\alpha}^\dagger(\boldsymbol{s})\|_1$ or $\|\tilde{L}_n \boldsymbol{\alpha}^\dagger(\boldsymbol{s})\|_2$ is small and we in turn allow for a larger γ_n to enforce such structure. On the contrary, if the fusion term does not represent the truth closely, γ_n is forced to be small and the regularization is thus weaker. Also, we write $\delta = O(n^{-1/2})$ instead of $O_P(n^{-1/2})$ to reflect that the choice of δ need not be data-driven. A user-specified choice of small δ suffices for computational purposes.

Consider, for the moment, the low-dimensional $\boldsymbol{\beta}^0$ without the ℓ_1 sparsity penalty. With the assumptions introduced above, we are now ready to examine the minimizer of the combination of the loss function along with the fusion penalty, and investigate its relationship with the true baseline intensity $\alpha^0(\cdot)$ and regression parameters $\boldsymbol{\beta}^0$.

Lemma 4.1 (Validity of PMLE in low dimensions)**.** *Under i)-iii) of Assumption 4.1, there exists $s_1 \in \Omega_1, \ldots, s_n \in \Omega_n$ such that letting $\boldsymbol{\alpha}^\dagger := (\alpha^0(s_1) + \phi(s_1), \ldots, \alpha^0(s_n) + \phi(s_n))$ and $\boldsymbol{\beta}^\dagger := \boldsymbol{\beta}^0$, we have*

$$-\nabla_{(\tilde{\boldsymbol{\alpha}}, \boldsymbol{\beta})} \mathbb{P}_0 \ell(\tilde{\boldsymbol{\alpha}}, \boldsymbol{\beta})\big|_{(\boldsymbol{\alpha}^\dagger, \boldsymbol{\beta}^\dagger)} = 0,$$

where $\mathbb{P}_0$ denotes expectation under the true data generating mechanism. Furthermore, denoting

$$(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) := \underset{\tilde{\boldsymbol{\alpha}}, \boldsymbol{\beta}}{\operatorname{argmin}} -\mathbb{P}_0 \ell(\tilde{\boldsymbol{\alpha}}, \boldsymbol{\beta}) + \gamma_n R(\tilde{\boldsymbol{\alpha}}; \mathcal{G}_n),$$

it holds under Assumption 4.2 that $\boldsymbol{\beta}^* = \boldsymbol{\beta}^0$, and

$$\left\| \boldsymbol{\alpha}^* - \boldsymbol{\alpha}^\dagger \right\|_2 = O_P\left(\gamma_n \sup_{\boldsymbol{s}: s_1 \in \Omega_1, \ldots, s_n \in \Omega_n} \|\tilde{L}_n \boldsymbol{\alpha}^\dagger(\boldsymbol{s})\|_2 \right)$$

for the ℓ_2 smoothing penalty, or

$$\left\| \boldsymbol{\alpha}^* - \boldsymbol{\alpha}^\dagger \right\|_1 = O_P\left(\gamma_n^2 \sup_{\boldsymbol{s}: s_1 \in \Omega_1, \ldots, s_n \in \Omega_n} \|B_n \boldsymbol{\alpha}^\dagger(\boldsymbol{s})\|_1 \right)$$

for the ℓ_1 fusion penalty.

Proofs of Lemma 4.1 and all other theoretical results in this chapter are provided in Appendix C.3. We call $\boldsymbol{\theta}^\dagger := ({\boldsymbol{\alpha}^\dagger}^\top, {\boldsymbol{\beta}^\dagger}^\top)^\top$ as defined in Lemma 4.1 the *target parameter*, since it is what the loss function (on population level), without any penalty, would lead us to find. Lemma 4.1

states that the target slope parameter $\boldsymbol{\beta}^\dagger$ associated with the loss function is equal to the true slope $\boldsymbol{\beta}^0$ when using either fusion penalty, even though the loss function ignores the stochasticity in the intensity as well as the continuous (rather than discrete) nature of the baseline intensity function $\alpha^0(\cdot)$. The ignored stochasticity translates to a systematic bias in the target intercepts (comparing to the discretized true baseline $\tilde{\boldsymbol{\alpha}}^0$), determined only by the distribution of the errors at a finite set of locations, instead of the whole error random field.

We impose a soft constraint on the structure of $\alpha^0(\cdot)$, reflecting the belief that the intensities at close-by regions are similar. Lemma 4.1 provides a bound on the change in the solution when such constraint is incorporated into optimization. When this belief is violated by the true mechanism, under Assumption 4.2, the ℓ_2 or ℓ_1 norm of the difference is $o_P(n^{1/2})$, which is on average decaying when examining each entry of $\boldsymbol{\alpha}^*$ element-wise. In contrast, when our structure assumption holds and the total variation of $\alpha^0(\cdot) + \varepsilon(\cdot)$ is bounded, $\|\tilde{L}_n \boldsymbol{\alpha}^\dagger(\boldsymbol{s})\|_2$ and $\|B_n \boldsymbol{\alpha}^\dagger(\boldsymbol{s})\|_2$ are small and thus the gap between $\boldsymbol{\alpha}^*$ and $\boldsymbol{\alpha}^\dagger$ resulting from the smoothness penalty is negligible.

An additional set of conditions on the tail behavior of the process and the scale of and the structure of the design matrix are needed for our consistency result.

Assumption 4.3 (Compatibility condition)**.** *Given the true support $\mathcal{S} \subset \{1, \ldots, p\}$ of $\boldsymbol{\beta}^0$ such that $|\mathcal{S}| = s$, define*

$$\mathcal{C}(\mathcal{S}) := \{\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}) : \|\boldsymbol{\beta}_{\mathcal{S}^C}\|_1 \leq \|\boldsymbol{\alpha}\|_1 + 3\|\boldsymbol{\beta}_{\mathcal{S}}\|_1\}.$$

Then, for any $\boldsymbol{\theta} \in \mathcal{C}(\mathcal{S})$,

$$\frac{\|\boldsymbol{\alpha}\|_1}{2} + \|\boldsymbol{\beta}_{\mathcal{S}}\|_1 \leq \frac{\|\boldsymbol{\theta}\|_2 \sqrt{s}}{\varphi_s}$$

for some constant $\varphi_s > 0$ only depending on the sparsity s.

Assumption 4.4 (Bounded intensity)**.** $0 < \psi_{\alpha,\beta,\phi} \leq \inf_{s \in \Omega} \exp\left[\alpha^0(s) - \|\boldsymbol{X}\boldsymbol{\beta}^0\|_\infty + \phi(s)\right]$*, and* $\sup_{s \in \Omega} \exp\left[\alpha^0(s) + \|\boldsymbol{X}\boldsymbol{\beta}^0\|_\infty + \phi(s)\right] < \Psi_{\alpha,\beta,\phi} < \infty$ *for some $\psi_{\alpha,\beta,\phi}$ and $\Psi_{\alpha,\beta,\phi}$.*

The compatibility condition is common in high-dimensional literature (Bühlmann and van de Geer, 2011). Error bounds for high-dimensional models are often established by assuming sub-Gaussian or sub-exponential tails. However, the validity of these assumptions is not automatically clear for our setting, since the stochasticity in intensity leads to a heavier tail than the conditional

Poisson distribution. The upper bound on intensity and the asymptotic regime given in Definition 4.1 guarantee that the case counts have bounded finite moments, uniformly across $\Omega_1, \ldots, \Omega_n$, which suffices for our proof of consistency. The lower bound is required in combination with Assumption 4.6 below to ensure sufficient curvature near the target parameter $\boldsymbol{\theta}^\dagger$, which is a form of restricted strong convexity (Negahban et al., 2012), a common condition required for high-dimensional M-estimators. Sufficient curvature of the loss function around the target parameter guarantees that a small difference in the loss function translates to a small estimation error.

Assumption 4.5 (Sparsity of $\boldsymbol{\beta}^0$)**.** *The true slope* $\boldsymbol{\beta}^0$ *satisfies* $\|\boldsymbol{\beta}^0\|_0 = s$ *with* $s = o\left(\sqrt{\frac{n}{\log p}}\right)$.

Assumption 4.6 (Design matrix)**.** *The design matrix* $\boldsymbol{X}$ *satisfies* $\max_i \max_j |X_{ij}| \leq R$ *for some* $R < \infty$. *Also the restricted eigenvalue condition (Bickel et al., 2009),* $\frac{1}{n}\|\boldsymbol{X}\Delta\|^2 \geq \kappa\|\Delta\|^2$, *holds over* $\mathcal{B} := \left\{\Delta : \|\Delta\|_1 \leq \frac{4\tau_n^2 s}{c\rho\varphi_s^2}\right\}$ *for* $c > 0$, $\varphi_s > 0$ *and* $\rho = O\left(\sqrt{\frac{\log p}{n}}\right)$.

We can now present our consistency result.

Theorem 4.1 (Consistency of penalized PMLE)**.** *Under Assumptions 4.1–4.6, the solution* $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\alpha}}^\top, \hat{\boldsymbol{\beta}}^\top)^\top$ *of (4.5) satisfies*

$$\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_1 \leq C\sqrt{s\frac{\log p}{n}}$$

for some constant $C > 0$ *with probability converging to 1 under the asymptotic regime in Definition 4.1.*

4.3.2 Inference

In this section, we introduce a procedure for constructing confidence intervals for each β_j^0, $j = 1, \ldots, p$. The same result can easily be generalized to contrasts, i.e., linear combinations of multiple β's. It is known that solutions to penalized estimation problems are in general biased (Voorman et al., 2014), and it is not straightforward to analytically characterize their uncertainty (Zhao et al., 2021b). We adopt the idea of a de-biasing approach proposed by Javanmard and Montanari (2014), with two key differences from the original procedure: we generalize to non-Gaussian models, and account for the extra randomness from the error random field via a conservative sandwich covariance estimator.

A general de-biasied estimator takes the form $\hat{\boldsymbol{b}} = \hat{\boldsymbol{\beta}} + n^{-1} M \nabla_{\boldsymbol{\beta}} \ell(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$, where the choice matrix of M determines how well the bias and variance are controlled by the inference procedure. In our setting, such an estimator is given by

$$\hat{\boldsymbol{b}} = \hat{\boldsymbol{\beta}} + \frac{1}{n} M \boldsymbol{X}^\top \left[\boldsymbol{Y} - \boldsymbol{B} \odot \exp\left(\hat{\boldsymbol{\alpha}} + \boldsymbol{X} \hat{\boldsymbol{\beta}} \right) \right],$$

where, recalling the notation in Definition 4.1i),$\boldsymbol{B} = (|\Omega_1| P_1, \ldots, |\Omega_n| P_n)$ and $\odot$ denotes element-wise multiplication. Our choice of M is based on two quantities, the empirical Hessian of the negative Poisson log-likelihood,

$$\hat{\boldsymbol{H}} = -\frac{1}{n} \sum_{i=1}^{n} \nabla^2_{\beta} \ell(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}; x_i, y_i),$$

and an estimated covariance $\hat{\boldsymbol{\Sigma}}$ of the gradient $\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\alpha}^\dagger, \boldsymbol{\beta}^0)$, where $\boldsymbol{\alpha}^\dagger(\boldsymbol{s}) := \tilde{\boldsymbol{\alpha}}^0(\boldsymbol{s}) + \phi(\boldsymbol{s})$ was defined in Assumption 4.2. Note that simply using a plug-in estimate $\hat{\boldsymbol{H}}$ to derive $\hat{\boldsymbol{\Sigma}}$ would underestimate the variability, due to the stochasticity of the baseline intensity. Instead, we propose a (conservative) covariance estimate

$$\hat{\boldsymbol{\Sigma}} := \frac{2}{n} \sum_{i=1}^{n} X_i^\top X_i \left[\left(Y_i - |\Omega_i| P_i \exp(\hat{\alpha}_i + X_i \hat{\boldsymbol{\beta}}) \right)^2 + \left(|\Omega_i| P_i \exp(\hat{\alpha}_i + X_i \hat{\boldsymbol{\beta}}) - \bar{\mu} \right)^2 \right], \tag{4.7}$$

where $\bar{\mu} := n^{-1} \sum_i |\Omega_i| P_i \exp(\hat{\alpha}_i + X_i \hat{\boldsymbol{\beta}})$. The first term in (4.7), without the multiplier 2, is a natural estimator for Poisson (not doubly-stochastic) data, and the added terms capture the additional stochasticity in the latent intensity.

Finally, M is defined such that its jth row, m_j is the solution of

$$\min_m m \hat{\boldsymbol{\Sigma}} m^\top \quad \text{s.t.} \quad \| \hat{\boldsymbol{H}} m^\top - e_j \|_\infty \leq \eta \tag{4.8}$$

with e_j being the vector with one at the jth entry and zero everywhere else, and η being a small tolerance parameter. Extending Javanmard and Montanari (2014), the optimization problem (4.8) aims to control two quantities: $\max_{i,j} |(\hat{\boldsymbol{H}} M - I_p)_{ij}|$ corresponding to the non-Gaussianity and bias of $\hat{\boldsymbol{b}}$, and $(M \hat{\boldsymbol{\Sigma}} M)_{ii}$ relating to the variance of $\hat{\boldsymbol{b}}$. However, (4.8) differs from the original optimization problem proposed by Javanmard and Montanari (2014) in that the bias and variance are captured separately by $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{H}}$ in our setting. This is expected since the first-order properties of the penalized PMLE are determined by the Poisson log-likelihood, while the doubly-stochastic nature of the true process needs to be accounted for when characterizing second-order properties.

The following theorem establishes the asymptotic normality of each $\hat{b}_j$, from which valid statistical inference can be conducted.

Theorem 4.2 (Asymptotic normality)**.** *Let* $\sigma_j := \left[M\mathbb{E}_0\nabla_\beta\ell(\boldsymbol{\alpha}^\dagger,\boldsymbol{\beta}^0)\nabla_\beta\ell(\boldsymbol{\alpha}^\dagger,\boldsymbol{\beta}^0)^\top M^\top\right]_{jj}$. *Under Assumptions 4.1–4.5 and further assuming that*

i) η *in (4.8) is set to be* $o(1/\sqrt{s\log p})$*;*

ii) There exists a small neighborhood $\mathcal{N}(\delta_{\boldsymbol{\alpha}},\delta_{\boldsymbol{\beta}})$ *around 0 such that for any* $(\delta_{\boldsymbol{\alpha}},\delta_{\boldsymbol{\beta}})\in\mathcal{N}(\delta_{\boldsymbol{\alpha}},\delta_{\boldsymbol{\beta}})$

a) $\max_j\left\|\frac{1}{n}\nabla^2_{\boldsymbol{\beta}}\ell\left(\hat{\boldsymbol{\alpha}}+\delta_{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}}+\delta_{\boldsymbol{\beta}}\right)m_j^\top-e_j\right\|_\infty=o_P\left(1/\sqrt{s\log p}\right)$*;*

b) $\left\|M\nabla^2_{\boldsymbol{\beta},\boldsymbol{\alpha}}\ell(\hat{\boldsymbol{\alpha}}+\delta_{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})\right\|_\infty=o_P(n/\sqrt{s\log p})$ *and* $\left\|M\nabla^2_{\boldsymbol{\beta},\boldsymbol{\alpha}}\ell\left(\boldsymbol{\alpha}^\dagger+\delta_{\boldsymbol{\alpha}},\boldsymbol{\beta}^0\right)\right\|_2=O_P(1)$,

we have, for each $j=1,\ldots,p$,

$$\frac{\sqrt{n}\left(\hat{b}_j-\beta_j^0\right)}{\sigma_j}\xrightarrow{d}N(0,1).$$

We show in Appendix C.3 that $[M\hat{\Sigma}M^\top]_{jj}$ as defined in (4.7) serves as a conservative estimator of σ_j, and thus leads to a conservative inference procedure. The inference procedure above does not rely on a known form of the error distribution. When such knowledge is available, however, we could obtain a more efficient covariance estimate. In particular, we would be able to derive the expression of the population-level quantity $\mathbb{E}_0\left[\nabla_{\boldsymbol{\beta}}\ell(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})\right]\left[\nabla_{\boldsymbol{\beta}}\ell(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})\right]^\top$ depending on some variance parameters. For example, when the error random field is independent, stationary and Gaussian with variance σ^2, a covariance estimator is given by

$$\tilde{\Sigma}:=\frac{1}{n}\sum_{i=1}^n X_i^\top X_i\left[P_i\exp(\hat{\alpha}_i+X_i\hat{\beta})+\left(\exp(\hat{\sigma}^2)-1\right)P_i^2\exp(\hat{\alpha}_i+X_i\hat{\beta})^2\right]. \tag{4.9}$$

Calculating (4.9) requires an estimate for σ^2. However, the entire term $\zeta:=\exp(\hat{\sigma}^2)-1$ can be estimated via a method of moment approach:

$$\hat{\zeta}:=\frac{1}{n}\sum_{i=1}^n\frac{\left(Y_i-P_i\exp(\hat{\alpha}_i+X_i\hat{\beta})\right)^2-P_i\exp(\hat{\alpha}_i+X_i\hat{\beta})}{P_i^2\exp\left(\hat{\alpha}_i+X_i\hat{\beta}\right)^2}. \tag{4.10}$$

A small σ^2 may lead to negative estimates of ζ. To avoid this, the summand in (4.10) can be replaced with its positive part $(\cdot)_+:=\max\{\cdot,0\}$. This leads to slightly conservative confidence intervals for β_j's in the worst case scenario.

4.4 Simulations

In this section, we illustrate the performance of the penalized PMLE approach in comparison to Bayesian methods for LGCP. We simulate 100 replicates from an LGCP on $\Omega = [0, m] \times [0, m]$, partitioning Ω into $n = m^2$ cells of unit squares. The baseline intensity is given by $\alpha^0(s) = \frac{1}{4m}\sqrt{s_1^2 + s_2^2}$ for $(s_1, s_2) \in \Omega$. The random error $\varepsilon(\cdot)$ consists of a spatially structured component along with an unstructured component. The structured component is generated from a Gaussian random field having zero mean and an exponential covariance with range parameter $0.2m$; the unstructured component is generated on a fine (60×60) grid where the error is constant on each small cell, drawn from independent Gaussian distributions with unequal variances to induce non-stationarity. In particular, the variances are simulated from inverse Gamma distribution with shape parameter 2 and rate parameter 1. Though a Gaussian random field is continuous, it is typically discretized and simulated on fine grids in practice, as is our case for $\alpha^0(s)$ and $\varepsilon(s)$. Each entry of the p-dimensional covariate X is drawn from Uniform$[-0.5, 0.5]$, and locations within the same cell share the same covariate values. The offset P is set to 2 for all m^2 cells. We consider two settings: (i) a low-dimensional setting where $p = 10$, with $\beta_1 = \beta_2 = -1$, $\beta_3 = \beta_4 = 1$, and $\beta_5 = \cdots = \beta_{10} = 0$; and (ii) a high-dimensional setting where $p = 100$, with $\beta_1 = \cdots = \beta_5 = -1$, $\beta_6 = \cdots = \beta_{10} = 1$, and all remaining entries being 0. We investigate a sequence of sample sizes, $n = 5^2, 10^2, 20^2, 30^2$, and define the graph $\mathcal{G}_n$ of cells as unweighted, where two cells are connected if they are adjacent (from the left, right, top or bottom).

We run PMLE with ℓ_1 and ℓ_2 fusion penalties, where tuning parameters γ_n and τ_n are jointly selected via 5-fold cross-validation, and compare our results with two discretized Bayesian LGCP models:

- LGCP specifying Gaussian random errors with exponential covariance fitted via `RStan`, based on 1000 posterior MCMC samples. The slope parameters are assigned Normal$(0, 10)$ priors, and the covariance parameters are assigned truncated Normal$(0, 5)$ priors;

- LGCP specifying a correlated error component via a two-dimensional random walk (RW2D) model on lattice grids, as well as an uncorrelated error component, fitted via `R-INLA`. Normal$(0, 10)$ priors are assigned for the slopes and an inverse Gamma$(1, 0.01)$ prior, which is the default

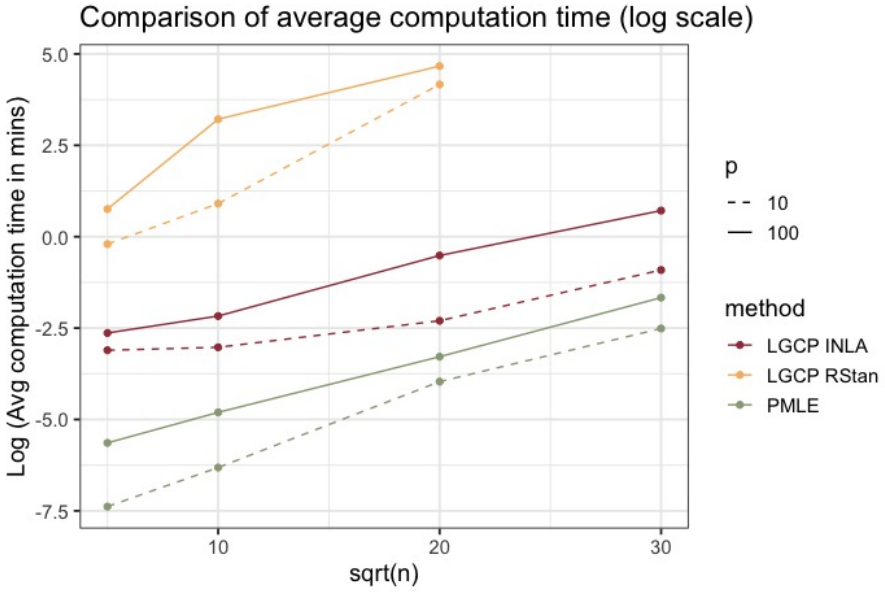


Figure 4.1: Average computation time for a single replicate of data in minutes, plotted on log scale, over 100 replicates for penalized PMLE and Bayesian LGCP model run via `RStan` and `R-INLA`.

prior implemented by the `INLA R` package, is adopted for the variance parameter.

Figure 4.1 presents the average computation time of penalized PMLE and Bayesian LGCP. It can be seen that PMLE and INLA scale well as the dimensionality and sample size increase, and PMLE is slightly faster than INLA in both settings. In contrast, MCMC sampling via `RStan` is time-consuming for large p and/or large n. For this reason, the simulation setting $n = 30^2$ is not examined for LGCP fitted via `RStan`.

Figure 4.2 compares our inference procedure with Bayesian inference, in terms of the coverage of confidence intervals, type I error rate and power. All three metrics are averaged across all relevant (e.g., non-zero for power) entries of $\boldsymbol{\beta}$. In low dimensions, Bayesian model fitted via `INLA` performs well, with power approaching 1 and well-controlled type I error rate along with valid 95% coverage. Penalized PMLE achieves similar accuracy as well, but requires more samples. The reduced power is not surprising, given the over-parameterized nature of penalized PMLE, and the fact that it does not require or make use of the parametric distribution of $\varepsilon(\cdot)$. `RStan` fails to control the type I error and provide proper coverage, at least for the given amount of data and MCMC samples.

With higher dimensions, Bayesian LGCP methods are not guaranteed to achieve the nominal 95% coverage or control type I error within 0.05, and we observe a trend of decreasing coverage for `INLA` as m increases. In contrast, the penalized PMLE controls type I error rate within 0.05 and still maintains reasonable power despite being slightly conservative.

In the high-dimensional setting, the observation that `INLA` has good power and acceptable type I error rate but decreasing coverage could be explained by its non-decaying estimation bias for the non-zero parameters. Figure 4.3 visualizes the element-wise estimation errors for `INLA` and PMLE with different sample sizes ($n = 10^2$ and 30^2, respectively). The variability of estimation errors shrinks faster for `INLA`, reflecting higher efficiency due to its parametric nature. Estimates for the non-zero entries (index 1 through 10) are attenuated for both methods, but this bias is decreasing for PMLE with more samples, while increasing for `INLA`. This issue occurs because the RW2D model for `INLA` assumes constant baseline risk on each observed cell, as well as a stationary error random field, both of which are violated in this simulation setting. Also, it is not clear how well `INLA` can handle high-dimensional covariates, as the current choice of priors on $\boldsymbol{\beta}$ does not induce shrinkage or regularization to handle high dimensionality. Shrinkage priors, such as horseshoe (Carvalho et al., 2009, 2010), may alleviate this issue but can be more computationally demanding.

4.5 Application: Seattle Crime Data

We analyze the Seattle crime data[1] to further demonstrate the performance of our approach in comparison with a wider range of alternative methods. We focus on crimes against persons that were reported to the Seattle Police Department in Spring 2021 (April 1 through June 30). Crime cases are recorded as point incidences (with blurred location) over the Seattle map, which we aggregate to the level of census tracts, since this is the finest resolution of covariates available. The population size of each census tract is used as offset. Covariates are obtained from King County GIS Open Data[2] and include:

- Demographic and socioeconomic information: age distribution (proportion of residents in four age groups: 18-29, 30-44, 45-59, 60 and above); race/ethnicity distribution (proportion

[1] `https://www.seattle.gov/police/information-and-data/crime-dashboard`

[2] `https://www.kingcounty.gov/services/gis/GISData.aspx`

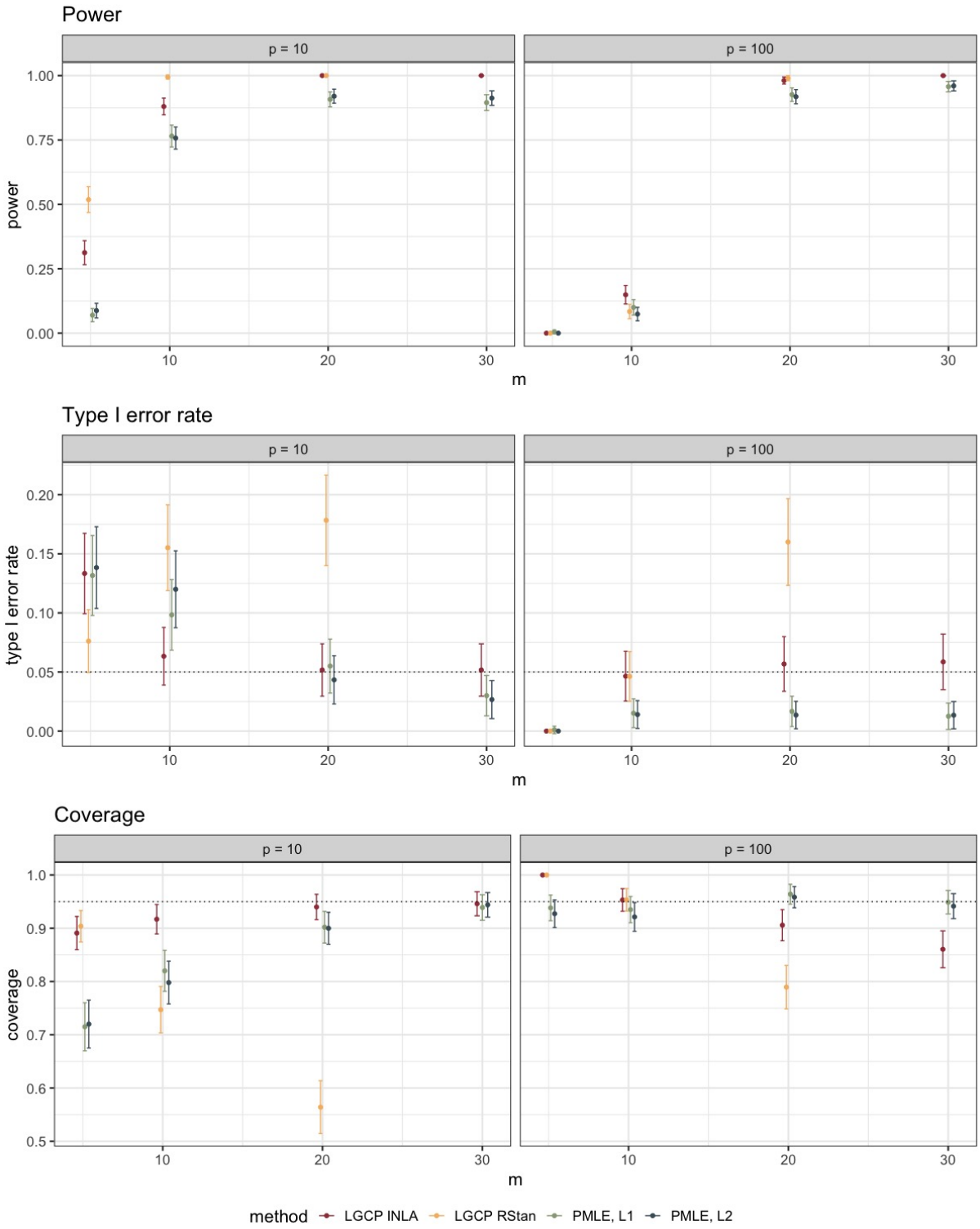


Figure 4.2: Comparison of coverage, type I error rate and power for penalized PMLE and Bayesian LGCP methods, with standard error bars.

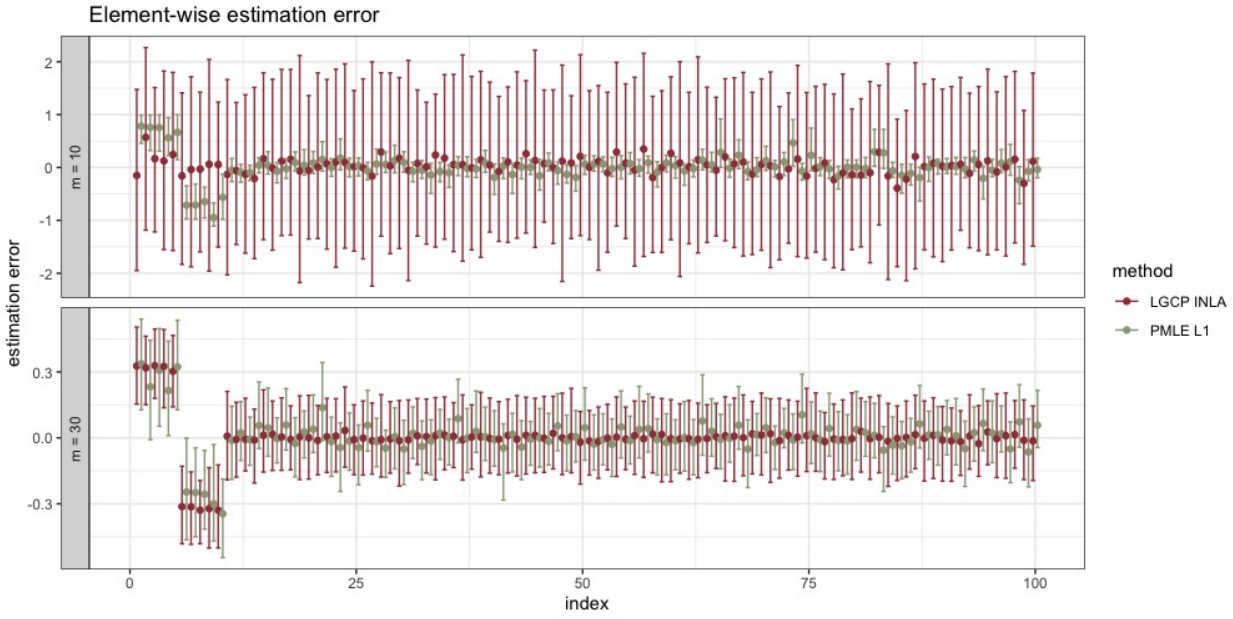


Figure 4.3: Average element-wise estimation errors along with the 5% and 95% percentiles of errors for β in the high-dimensional ($p = 100$) setting, with and $n = 5^2$ (top) and 30^2 (bottom) cells, respectively.

of Asian, Black, Hispanic, White populations, and populations with two or more races); median household income, education status (proportion of residents with college degree or above); and proportion of residents with medical insurance.

- Public facilities: number of hospitals; transit stops; fire stations; police stations; food facilities; schools; solid waste facilities; farmers' markets;

- Environmental information: area of region; proportion of medium and high basins.

We purposely choose a wide range of covariates, including those that are not known as good predictors of crime rate. Covariates are all summarized by census tract. For covariates characterized by proportion of different groups (e.g. age, race/ethnicity, medium/high basins), we omit one category as the reference level and adopt the additive log ratio transformation (Aitchison, 1982) to alleviate the spurious correlation in such compositional data. The spatial domain is modeled as an unweighted graph, where two regions are connected if they share a common border.

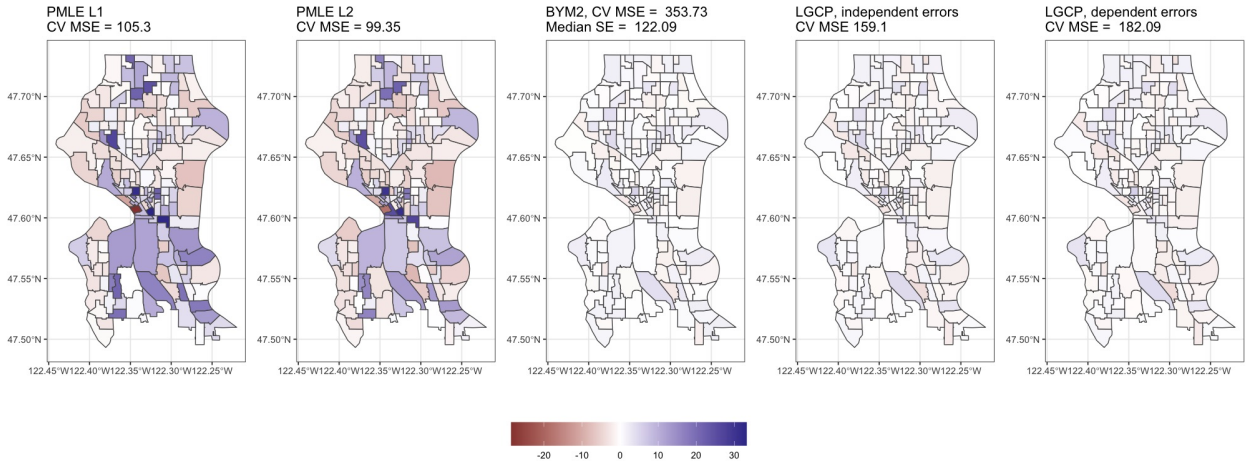


Figure 4.4: Residuals from each model, with cross-validated MSEs reported in the titles. Due to the large variability of prediction errors for the BYM2 model, we also report its median prediction SE for reference.

We compare the penalized PMLE with ℓ_1 and ℓ_2 fusion penalties with the following Bayesian models, implemented in `INLA`. The default penalized complexity (PC) priors (Simpson et al., 2017) in the `INLA R` package are used for the variance, range (for LGCP) and mixing (for BYM2) parameters.

- The BYM2 model (Riebler et al., 2016) which specifies the linear predictor to be a sum of covariate effects, with spatially correlated errors induced by connectivity, and independent, non-spatial heterogeneity. The mixing parameter (which is between 0 and 1 and modeled on the logit scale) controls how much variance comes from the independent versus spatially dependent random effects.

- The LGCP model with independent Gaussian error random field.

- The LGCP model with Gaussian error random field having exponential covariance.

The predictive performance for each model is evaluated using 5-fold cross-validation. We use prediction MSE as our key metric, recognizing that metrics such as conditional predictive ordinate

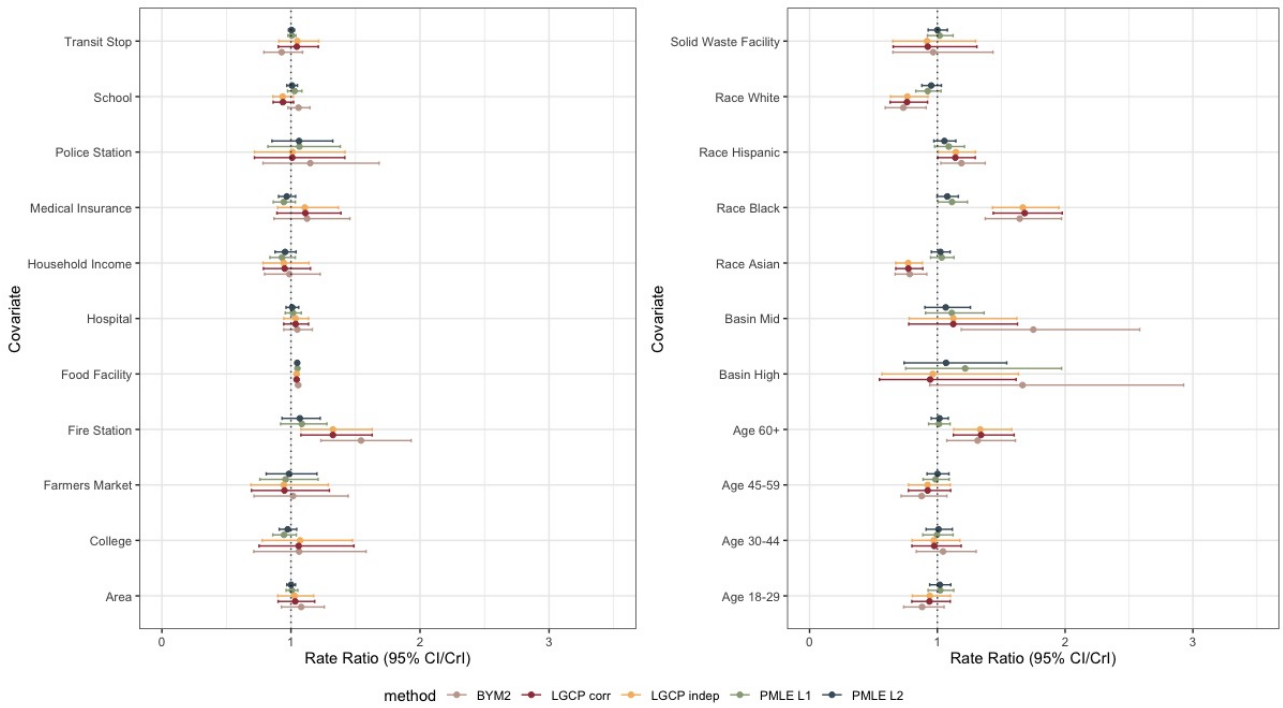


Figure 4.5: Estimated rate ratios with error bars indicating 95% confidence/credible intervals

(Gelfand and Dey, 1994), or CPO, are also helpful, but more suitable for Bayesian models. Figure 4.4 presents the residuals from each model along with their prediction MSEs. The residual plots capture how close each model fits to the data, while the prediction MSEs capture the overall predictive accuracy. The residual plots show that the Bayesian models have smaller bias comparing to penalized PMLE. However, the small bias comes at a cost of large variability, as reflected by the large MSE values and indicates over-fitting. Though LGCP with dependent errors conducts an implicit form of regularization for smoothness as achieved by a fusion penalty, such regularization is not explicit and it may thus be less straightforward to find a near-optimal bias and variance trade-off, compared to methods with explicit penalization.

Figure 4.5 presents point estimates along with 95% confidence/credible intervals (CI/CrI) shown as error bars. The models all identify race and the number of food facilities to be associated with crime incidents within a region. The former matches other studies (Krysan, 2008; Lodge et al., 2021; Uehara, 1994) reporting challenges in the search of housing and/or housing inequalities associated

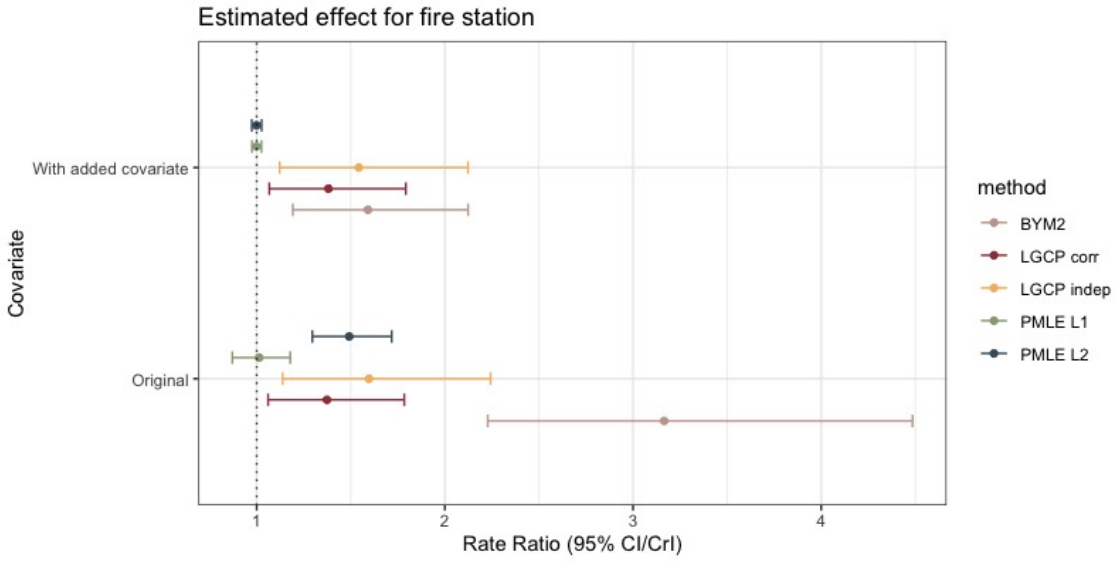


Figure 4.6: Comparison of estimated coefficients and 95% CI/CrI before and after adding a spatially structured covariate to the single covariate model of fire station

with race, as well as residents of underrepresented race being exposed to higher crime rates. Lodge et al. (2021) also found such disparity by race and ethnicity to decrease when the granularity of data is not very high (>1600m buffer size, which is within the range of most census tracts in our case). This aligns with the reduced effect sizes of race and ethnicity estimated via PMLE compared to Bayesian models. PMLE leads to narrower CIs than Bayesian methods for this dataset in general. Also, when there is discrepancy in estimated effects reported by other models, PMLE tends to produce intermediate estimates. This can be seen, for example, for the effect of transit stops and schools. In addition, BYM2 finds medium basin, and both BYM2 and LGCP find the proportion of senior population to be positively associated with crime rates, which is somewhat hard to explain based on common knowledge. These findings align with our observation from Section 4.4 that PMLE could have a better control of type I error without significantly affecting its power.

A common concern in the analysis of spatial data is the effect of *spatial confounding* (Reich et al., 2006). The presence of spatial coufounding, which occurs when covariates contributing to the variability in the response are spatially structured, may introduce biases to the estimated effect sizes. To investigate how much the results of PMLE and the Bayesian models could be

potentially impacted by this issue, we fit each model with fire station as the only covariate, and fit an additional set of models with a synthetic, spatially structured covariate added. Figure 4.6 compares the estimates along with 95% CI/CrI. We see that all models are not completely immune to spatial confounding, as indicated by the change in estimated effects after adding the spatially structured covariate; however, PMLE with ℓ_1 fusion penalty and the two LGCP models are more robust against the inclusion of this spatially structured variable. As a sensitivity analysis for the choice of priors and computational approach, we also present and discuss an alternative version of Figures 4.5 and 4.6 in Appendix C.2 where the Bayesian models are implemented via `RStan` with a different set of priors. We found that the model estimates and CrIs remain highly similar to Figure 4.5, while results from `RStan` are more sensitive to spatial confounding.

4.6 Discussion

We proposed a computationally simple approach to modeling semi-parametric, doubly-stochastic point processes with theoretical guarantees, focusing on the estimation and inference of fixed covariate effects. The key insight in the proposed method, which is based on a penalized regression framework, is that ignoring the stochasticity in the intensity and jointly modeling its realization along with the deterministic component still leads to a valid estimating function for the regression parameters. The nonparametric baseline is captured by a high-dimensional discretized intercept parameter. We solve this over-parametrized model with a fusion penalty for the region-specific intercepts, along with a sparsity penalty for the regression parameters. However, the soft constraint on smoothness does not need to hold exactly to ensure the validity of this penalization approach. We address the extra stochasticity in our statistical inference procedure and introduce robust covariance estimates under scenarios with and without stationarity of the error random field.

In our discussion of spatial confounding (Section 4.5), we focused on the notion of spatial confounding considered by Reich et al. (2006), while alternative notions (e.g., in Keller and Szpiro, 2020; Paciorek, 2010) have been proposed and may guide the choice of models differently. Literature on graph denoising may motivate further simplifications to the computational approach in the proposed Poisson maximum likelihood, or other graph-based spatial models. A sparse approximation to the edge incidence matrix B_n, as in Padilla et al. (2017), or an approximation to the graph Laplacian L_n, as in Sadhanala et al. (2016), could reduce the computational burden for

large-scale settings. Establishing consistency and asymptotic normality in the the presence of such approximations would be an interesting topic of future research. Additional considerations may be helpful for selecting the threshold η in the de-biasing procedure (see Equation 4.8) in practice, which controls the trade-off between type I error rate and power, especially with limited samples. Finaly, as noted in Section 4.2.3, prediction and parameter tuning may not be straightforward for graphical or spatial models, and naïve cross-validation is somewhat ad-hoc for correlated observations. Establishing theoretical guarantees for such an approach by leveraging recent developments in this area (Rabinowicz and Rosset, 2022) and/or incorporating alternative parameter tuning strategies that do not require sample splitting could be of interest.

Chapter 5

ESTIMATION OF DIRECT AND INDIRECT TREATMENT EFFECTS UNDER APPROXIMATE NEIGHBORHOOD INTERFERENCE ON RANDOM GRAPHS

5.1 Introduction

When assessing the casual effect of an exposure or treatment on an outcome of interest, a common assumption is the absence of cross-unit interference, namely, treatment assigned to one unit does not affect the outcome of other units. For example, the stable-unit treatment value assumption (SUTVA) (Angrist et al., 1996; Rubin, 1980) under Rubin's classical potential outcome framework includes this no-interference assumption.

However, there are many application settings where this assumption is violated; for instance, in infectious disease settings, one's vaccination against an infectious disease reduces their susceptibility to such disease and consequently the chance that this disease is transmitted to other individuals within the same neighborhood (Halloran and Struchiner, 1995; Hudgens and Halloran, 2008). Other examples include education, social science and economics studies (Cai et al., 2015; Carter et al., 2021; Hong and Raudenbush, 2006; Paluck et al., 2016; Sobel, 2006) where the causal effect of interventions could spread by "social links" between peers or acquaintances. The social links by which interference occur are often characterized by a graph or social network, where the vertices represent individuals from the population, and two individuals are connected by an edge if some notion of social interaction exists between them.

Existing causal inference methodologies tailored to the cross-unit interference setting typically rely on relatively strong assumptions on the causal mechanism, randomization of intervention, sampling procedure, and/or structure of social network. Baird et al. (2018); Cai et al. (2021); Hudgens and Halloran (2008); Kim et al. (2015); Manski (2013); Miguel and Kremer (2004); VanderWeele and Tchetgen (2011) estimated the causal effect of interventions under interference, explicitly or implicitly exploiting the *partial interference* assumption, which specified that the population consists of a large number of clusters between which interference is not present. This assumption, essentially

requiring SUTVA between clusters, facilitates the extension of classical methods and theories to the network interference setting, but limits more general applications with other network structures or interference mechanisms.

Aronow and Samii (2017); Leung (2020); van der Laan (2014) investigated the single large network scenario without assuming independent clusters. van der Laan (2014) considered a longitudinal setting where the time-dependent outcome depends on previous periods, and assumed the network degrees of all units to be uniformly bounded. Aronow and Samii (2017) imposed high-level week dependence assumptions on the potential outcome, which are less straightforward to justify or interpret in terms of network structure or causal mechanism. Leung (2020) required the degree of the network to be bounded as the sample size grows, and as in most of the aforementioned literature, that the potential outcomes as well as the network are both deterministic. Sävje et al. (2021) also relied on this sparsity condition, and could only accommodate the estimation of indirect effects if such effect can be expressed by a binary indicator.

Li and Wager (2022) discussed the inference of direct and indirect treatment effects on random graphs having growing network degrees as sample size increases. This approach relied on a graphon sampling model underlying the random graph, and restricted interference to occur only between direct neighbors on the network. However, real-world social networks tend to have a more heterogeneous degree distribution (Watts and Strogatz, 1998; Zhou et al., 2020), and interference could also exist for more distant pairs of individuals (Bond et al., 2012; Cai et al., 2015; Guilbeault et al., 2018; Jackson et al., 2008). Choi (2017) and Leung (2022) relaxed the direct neighbor interference assumption, but Choi (2017) assumed monotonicity for the treatment effect, and Leung (2022) imposed an exposure mapping model where the causal estimand was specified as a contrast between summary statistics of direct or indirect effects – this may be harder to interpret outside of the exposure mapping setting.

We are interested in the estimation and inference of direct and indirect treatment effects under *approximate neighborhood interference* (Leung, 2022), that is, interference could exist between distant pairs of individuals, where the magnitude of interference decays with network distance. Furthermore, we explore the random graph setting and only impose high-level assumptions on the network structure, instead of requiring a specific random graph model. The causal mechanism is specified via a semiparametric model, accommodating both continuous and categorical response

variables, where the mean of potential outcome is specified as a combination of nonparametric nuisance effects and parametric direct and indirect treatment effects. By doing this, our estimation and inference procedures are also valid under imperfect randomization of treatment, that is, when there are observed confounders, which is an additional flexibility comparing to most existing approaches.

In Section 5.2, we introduce necessary notations and describe our model setup including the causal mechanism, and the characterization of direct and indirect treatment effects following the framework of Hu et al. (2022) but generalize to non-continuous outcomes. Section 5.3 describes our proposed estimation and inference procedure, for which consistency and asymptotic normality guarantees are established in Section 5.4. Section 5.5 illustrates the performance of our proposed approach through simulations under various settings. Section 5.6 summarizes our key contributions and findings along with concluding remarks.

5.2 Model Setup

5.2.1 Social Network and Causal Model

Suppose the data is observed on a network $G_n = (V_n, E_n)$ of individuals, where the vertices represent individuals $i = 1, \ldots, n$, and the edges $\{E_{ij} \in \{0, 1\} : i, j \in V\}$, which are undirected, are generated from a random graph distribution distribution P_G. Examples include Erdős–Rényi graphs (Erdős and Rényi, 1959) where the E_{ij}'s are independent and identically distributed Bernoulli random variables, or power law graphs (Aiello et al., 2001) specifying $|\{i : \deg(i) = d\}| = n/d^{\alpha}$ for some parameter α, where $\deg(i) = \sum_{j \neq i} E_{ij}$ is the network degree of vertex i.

We let each individual i be associated with covaraites $x_i \in \mathbb{R}^K$ representing personal characteristics such as demographics, socioeconomic or health status. The treatment of interest W is assigned independently to each individual as $W_i \sim \text{Bernoulli}(\pi_i)$, where the treatment assignment probabilities π_i can depend on x_i via an unknown relationship $\pi_i = \phi(x_i)$.

Let $\ell_{ij} := \ell(i, j)$ be the length of the shortest path between individuals i and j, which is zero when $i = j$. We call i an *s-neighbor* of j (and vice versa) if $\ell_{ij} = s$. Further, we define $N_{is} := \sum_j I\{\ell_{ij} = s\}$ to be the number of s-neighbors for individual i, and $M_{is} := \sum_{j:\ell_{ij}=s} w_j$ be the number of treated s-neighbors of individual i. If i has no s-neighbor, we simply take

$M_{is} = N_{is} = 0$.

We denote the potential outcome for individual i as $Y_i(W)$, indexed by the n-dimensional treatment assignment vector instead of W_i alone. In particular, $Y_i(w_j = w; W_{-j})$ denotes the potential outcome of individual i if treatment $w \in \{0, 1\}$ were assigned to individual j, and the treatment assigned to other individuals were maintained $W_{-j} \in \{0, 1\}^{n-1}$.

We assume that the potential outcome is given by the mean model

$$g\left(\mu_i(w_i; W_{-i})\right) = \sum_{k=1}^{K} f_k^*(x_{ik}) + \beta_0^* w_i + \sum_{s=1}^{n-1} \beta_s^* \frac{M_{is}}{N_{is}}, \tag{5.1}$$

where $g(\cdot)$ is a link function as in a generalized linear model (GLM), such as logit function for binary outcomes, and $f_k^*, \beta_0^*, \beta_s^*$ denote the truth functions or parameters. We let $0/0 = 0$ by convention, in the case that the length of the longest path is less than $n - 1$. We define $\mu_i(W) = \mathbb{E}[Y_i(W) \mid W]$ as the expected potential outcome given treatment vector W. Each f_k is an unknown function capturing the effect of the k-th nuisance variable, and recalling that the treatment assignment probabilities π_i may also depend on $\{x_{ik}\}_{k=1}^K$, this formulation allows for imperfect randomization as long as the confounders are observed.

We do not impose parametric assumptions on these f_k's, while we specify linear forms for the direct and indirect treatment effects, given by the second and third terms in (5.1). In particular, (5.1) states that one's potential outcome depends on the treatments received by other individuals only through the proportion of treated s-neighbors for the sequence $s = 1, \ldots, n - 1$, regardless of the individual characteristics of these s-neighbors. This is a form of the *anonymous interference assumption* (Hudgens and Halloran, 2008) that is common in literature; but our requirement is weaker in that s is allowed to grow with n as opposed to remaining fixed (Li and Wager, 2022; Manski, 2013), despite the decaying magnitude of $\beta^*(s)$ as s increases.

5.2.2 Characterization of Direct and Indirect Effects

We follow the framework of Hu et al. (2022) to character direct and indirect effects under the potential outcome paradigm when cross-unit interference is present. In particular, we define the *average direct effect* (ADE) of the treatment as

$$\bar{\tau}_{\text{ADE}} = \frac{1}{n} \sum_i g\left(\mu_i(w_i = 1, W_{-i})\right) - g\left(\mu_i(w_i = 0, W_{-i})\right), \tag{5.2}$$

and the *average indirect effect* (AIE) as

$$\bar{\tau}_{\text{AIE}} = \frac{1}{n}\sum_{i}\sum_{j\neq i} g\left(\mu_j(w_i = 1, W_{-i})\right) - g\left(\mu_j(w_i = 0, W_{-i})\right). \tag{5.3}$$

The form of ADE is more closely connected to the classical no-interference setting. Sävje et al. (2021) showed that for some experimental designs and causal mechanisms, commonly used causal estimators assuming no interference are consistent even when interference is present, justifying (5.2) as a natural extension for the common notion of treatment effect, that is, the average effect of the treatment W_i on the individual that is intervened on, while keeping everything else fixed.

In definition of AIE, the term $g\left(\mu_j(w_i = 1, W_{-i})\right) - g\left(\mu_j(w_i = 0, W_{-i})\right)$ represents the contrast in linear predictors examining the pair (i, j), and in particular, reflects the effect of altering individual i's treatment on the outcome of individual j. The form of $\bar{\tau}_{\text{AIE}}$ calculates the total indirect effects of i's treatment on all other individuals, and averages across all i.

Under the assumed mean model (5.1), it can be seen that our causal estimands of interest satisfy $\bar{\tau}_{\text{ADE}} = \beta_0$ and $\bar{\tau}_{\text{AIE}} = \sum_{s=1}^{n} \beta_s$. We note that they are both defined on the scale of linear predictors, and consequently, their interpretations are also on the transformed scale given by $g(\cdot)$; for example, odds ratios for binary Y. In practical applications of causal inference, however, the natural scale estimands may also be of interest due to considerations such as collapsibility, which determines the validity when generalizing statistical results from sub-populations to the general population (Colnet et al., 2023). To this end, we define the *natural scale treatment effects* analogously as

$$\bar{\tau}^{\text{N}}_{\text{ADE}} = \frac{1}{n}\sum_{i} \mu_i(w_i = 1, W_{-i}) - \mu_i(w_i = 0, W_{-i}) \tag{5.4}$$

and

$$\bar{\tau}^{\text{N}}_{\text{AIE}} = \frac{1}{n}\sum_{i}\sum_{j\neq i} \mu_j(w_i = 1, W_{-i}) - \mu_j(w_i = 0, W_{-i}). \tag{5.5}$$

5.3 Method

We discuss our main methodology in this section, starting with estimation for $\bar{\tau}_{\text{ADE}}$ and $\bar{\tau}_{\text{AIE}}$ under the penalized regression framework. We propose an optimization problem leading consistent estimators for $\bar{\tau}_{\text{ADE}}$ and $\bar{\tau}_{\text{AIE}}$ in Section 5.3.1, and discuss strategies for parameter tuning in dependent data settings in Section 5.3.2. We develop a statistical inference procedure in Section 5.3.3 following

the decorrelated score approach in Ning and Liu (2017). Section 5.3.4 extends to the natural scale estimands $\bar{\tau}^{\mathrm{N}}_{\mathrm{ADE}}$ and $\bar{\tau}^{\mathrm{N}}_{\mathrm{AIE}}$ and establishes an inference procedure via parametric bootstrap.

5.3.1 A Penalized Regression Approach

We are interested in learning about the causal effects of treatment with the presence of nuisance effects. Parametric estimation and inference, e.g., via GLMs, may suffer when the nuisance effects represented by $\{f_k\}$ are non-linear and/or mis-specified (Buja et al., 2019; Vansteelandt and Dukes, 2022; Whitney et al., 2019). Non-parametric estimation, e.g., under the framework of generalized additive models (Hastie and Tibshirani, 1990), is more flexible but may not lead to easily interpretable causal estimands. We follow a similar idea as in semiparametric partially linear models (Carroll et al., 1997; Härdle et al., 2000). In particular, we specify linear terms representing the causal effects of interest, capture the nuisance effects as unknown functions $\{f_k\}$ within certain univariate function class $\mathcal{F}$, and impose structure-inducing penalty enforcing identifiability and regularity of the estimated functions (Haris et al., 2019a; Ravikumar et al., 2009).

We adopt the general framework of Haris et al. (2022) and also incorporate the parametric treatment effects into the optimization problem, leading to

$$\min_{f_1,\ldots,f_k\in\mathcal{F},\beta} \mathbb{P}_n \ell\left(y,\ \sum_{k=1}^{K} f_k(x_k) + w\beta_0 + \sum_{s=1}^{n-1} \frac{M_s}{N_s}\beta_s\right) + \lambda_\beta R_\beta(\beta) + \sum_{k=1}^{K} \lambda_k R_{\mathcal{F}}(f_k). \tag{5.6}$$

Here, $\beta := (\beta_0, \beta_1, \ldots, \beta_{n-1})$ is the vector-form notation combining the direct effect parameter β_0 and indirect effect parameters $\{\beta_s\}$. Likewise, we define $N_s := (N_{1s}, \ldots, N_{ns})$ and $M_s := (M_{1s}, \ldots, M_{ns})$ as the aggregated numbers of s-neighbors and treated s-neighbors, respectively. $\ell(\cdot)$ represents a convex loss function measuring the goodness of fit, e.g., log likelihood for GLMs. We use empirical process notation such that $\mathbb{P}_n \ell(y, \eta) = \sum_{i=1}^{n} \ell(y_i, \eta_i)$ for a linear predictor η summarizing the nuisance and treatment effects.

The second term in (5.6) controls the degree-of-freedom for the key parameters of interest β, whose dimensionality could grow with the sample size n based on the graph structure (i.e. length of the longest path). Besides the dimensionality, this penalty is important also because it handles the potentially strong collinearity between the proportions M_s/N_s induced by network connectivity. Common choices of R_β include the L1 (Tibshirani, 1996) or L2 (Hoerl and Kennard, 1970) penalties.

The last term in (5.6) consists of structure-inducing penalties on each of the nuisance functions f_k. The penalty $R_{\mathcal{F}}$ is required to be a semi-norm as in Haris et al. (2022), which satisfies all definitions of norms except that $R_{\mathcal{F}}(f)$ may be zero for non-zero f. We note that The form of $R_{\mathcal{F}}$ may depend on the function class $\mathcal{F}$ that these f_k's fall into; for example, if $\mathcal{F}$ is a reproducing kernel Hilbert space (RKHS) (Berlinet and Thomas-Agnan, 2011), e.g., the space of square integrable functions, then $R_{\mathcal{F}}$ is the corresponding RKHS norm, e.g., the function L_2 norm.

Under the causal mechanism specified in (5.1), we have that $\bar{\tau}_{\text{ADE}} = \beta_0$ and $\bar{\tau}_{\text{AIE}} = \sum_{s=1}^{n} \beta_s$, leading to the estimators

$$\hat{\tau}_{\text{ADE}} = \hat{\beta}_0, \quad \hat{\tau}_{\text{AIE}} = \sum_{s=1}^{n-1} \hat{\beta}_s.$$

5.3.2 Generalized Cross-Validation Score

A practical question in the implementation of (5.6) is how to select the tuning parameters λ_β and $\{\lambda_k\}$. The challenge is two-fold: first, common strategies such as cross-validation (CV) often rely on sample splitting, which is somewhat ad-hoc for dependent data, especially in our case where the correlation structure is implicit as induced by the graphical structure, for which we make no parametric assumption. We have found in numerical examples that parameter tuning via CV led to subpar model performance in our settings. The second difficulty is that we have $(K+1)$ parameters to determine, and the amount of computation grows exponentially in K, the number of nuisance parameters.

To establish a more justified and computationally simpler approach, we extend the proposal of Li (1985) and develop generalized cross-validation (GCV) scores for non-linear models. Our GCV score is based on an estimate for the degree-of-freedom considering the complexities in both the nuisance and parametric terms.

For simplicity in notation, we consider the matrix form of (5.6), where we suppose each f_k is estimated via $\hat{f}_k = \Psi_k \hat{\alpha}_k$ for matrix Ψ_k (e.g., the kernel matrix if $\mathcal{F}$ is a RKHS, or consisting of basis functions) and estimated coefficients α_k. We stack the K matrices together as $\Psi := [\Psi_1, \ldots, \Psi_K]$ and concatenate the coefficients as $\alpha := [\alpha_1^\top, \ldots, \alpha_K^\top]^\top$. Likewise, we express the treatment-related

variables as $Z := \left[w, \frac{M_1}{N_1}, \ldots, \frac{M_{n-1}}{N_{n-1}}\right]$ so that (5.6) is now written as

$$\min_{\alpha,\beta} \mathbb{P}_n \ell\left(y, \Psi\alpha + Z\beta\right) + \lambda_\beta R_\beta(\beta) + \sum_{k=1}^{K} \lambda_k R_\alpha(\alpha_k). \tag{5.7}$$

The previous regularization term on the functions f_k now translates to a penalty term on each of the α's. For instance, the RKHS penalty can be expressed as a quadratic term $\alpha_k^\top \Psi_k \alpha_k$.

We adopt the definition of generalized degrees of freedom (GDF) in Ye (1998) for a general modeling procedure $\mathcal{M}$ that

$$\mathrm{GDF}(\mathcal{M}) = \sum_{i=1}^{n} \frac{\partial \mathbb{E}_\mu[\hat{\mu}_i(Y)]}{\partial \mu_i} = \sum_{i=1}^{n} \frac{1}{\mathrm{Var}(Y_i)^2} \mathrm{cov}\left(\hat{\mu}_i(Y), Y_i - \mu_i\right), \tag{5.8}$$

where $\hat{\mu}_i = \Psi\hat{\alpha} + Z\hat{\beta}$, i.e., the predicted/fitted values. The last expression aligns with the definition of degrees of freedom in Efron (2004). Note that, denoting e_i as the i-th column of an n-dimensional identity matrix, we have

$$\frac{\partial \mathbb{E}_\mu[\hat{\mu}_i(Y)]}{\partial \mu_i} = \lim_{\delta \to 0} \mathbb{E}_\mu \left[\frac{\hat{\mu}_i(Y + \delta e_i) - \hat{\mu}_i(Y)}{\delta}\right] = \mathbb{E}_\mu \frac{\partial \hat{\mu}_i(Y)}{\partial Y} \tag{5.9}$$

if regularity conditions allow interchanging differentiation and integration (see Assumption 4.1-3). This allows us to estimate the GDF (implicitly indexed by the tuning parameter λ) as

$$\widehat{\mathrm{GDF}} := \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \hat{\mu}_i(Y)}{\partial Y},$$

which is a valid moment-based estimate despite the correlation between the Y_i's. Furthermore, (5.9) also establishes the above definition of GDF as a generalization of hat values in linear models, and specifically, (5.8) reflects the average sensitivity of the fitted values $\hat{\mu}_i$ with respect to the data points Y.

Finally, we combine the aforementioned definition of GDF with the GCV criterion (Hastie et al., 2009)

$$\mathrm{GCV}(\lambda) = \frac{\sum_{i=1}^{n} (y_i - \hat{\mu}_i(Y))^2}{\left(1 - \frac{\widehat{\mathrm{GDF}}(\lambda)}{n}\right)^2} \tag{5.10}$$

which includes the form of GCV score in linear settings (Li, 1985) as a special case. The criterion (5.10) enables the comparison of model performance balancing goodness-of-fit and model parsimony, without requiring sample splitting and out-of-sample prediction, and is a more justified approach for parameter tuning with dependent data as in our case.

5.3.3 Inference of $\bar{\tau}_{ADE}$ and $\bar{\tau}_{AIE}$

It is known that penalized regression estimates are in general biased (Voorman et al., 2014) and their distributions are hard to characterize analytically (Zhao et al., 2021b). To assess the uncertainty of our penalized estimates $\hat{\tau}_{\text{ADE}}$ and $\hat{\tau}_{\text{AIE}}$ and conduct statistical inference, we use a one-step correction to the estimated regression coefficients based on the decorrelated score function developed by Ning and Liu (2017). The high-level idea of the decorrelated score function is that it serves as an approximately unbiased estimating function for the parameters of interest (in our case, β) and is uncorrelated with the high-dimensional nuisance parameters (in our case, α or the f_k's), and therefore leads to valid and semi-parametrically efficient inference.

Denoting the solution to (5.7) as $\hat{\theta} := (\hat{\beta}^\top, \hat{\alpha}^\top)^\top$, we define the decorrelated score function

$$S(\beta, \alpha) := \nabla_\beta \ell(y, \Psi\alpha + Z\beta) - \xi^\top \nabla_\alpha \ell(y, \Psi\alpha + Z\beta) \tag{5.11}$$

as in Ning and Liu (2017), where $\xi^\top = I_{\beta\alpha} I_{\alpha\alpha}^{-1}$, with $I_{\beta\alpha}$ and $I_{\alpha\alpha}$ being the corresponding partitions of the information matrix $I = \mathbb{E}_\theta[\nabla^2 \ell(y, \Psi\alpha + Z\beta)]$. $S(\beta, \alpha)$ is uncorrelated with the nuisance score function $\nabla_\alpha \ell(\cdot)$ in that $\mathbb{E}_\theta[S(\beta, \alpha) \nabla_\alpha \ell(y, \Psi\alpha + Z\beta)] = 0$.

Instead of plugging in $\hat{\beta}, \hat{\alpha}$ to calculate ξ in (5.11), Ning and Liu (2017) proposed imposing a sparsity assumption on ξ (since its dimensionality scales with n) to control the estimation errors. In particular, for inference on the j-th entry of β, the proposed solving for

$$\hat{\xi}_j = \underset{\xi}{\operatorname{argmin}} \frac{1}{2} \mathbb{P}_n \left[\xi^\top \nabla_{\alpha\alpha} \ell(y, \Psi\hat{\alpha} + Z\hat{\beta}) \xi - 2\xi^\top \nabla_{\alpha\beta_j} \ell(y, \Psi\hat{\alpha} + Z\hat{\beta}) \right] + \lambda_\xi \|\xi\|_1,$$

which is equivalent to a weighted Lasso estimator if $\ell(\cdot)$ is the negative log-likelihood of a GLM. This procedure applies to cases with finite-dimensional β; while in our case, the dimensionality of β may increase with n depending on the graphical structure. However, if the interference from distant pairs of individuals decays fast enough with the network distance (formalized in Assumption 5.2), then our true parameter β^* can be approximated with a lower-dimensional parameter via shrinkage or truncation, and the finite-dimensional theory in Ning and Liu (2017) can consequently be adapted to our case. To this end, we modify the estimate for ξ via a group Lasso problem (Yuan and Lin,

2006), where $\xi := [\xi_0, \xi_1, \ldots, \xi_{n-1}]$[1] in (5.11) is solved as

$$\hat{\xi} = \underset{\xi_0,\ldots,\xi_{n-1}}{\operatorname{argmin}} \frac{1}{2}\mathbb{P}_n \left\{ \sum_{j=0}^{n-1} \left[\xi_j^\top \nabla_{\alpha\alpha} \ell(y, \Psi\hat{\alpha} + Z\hat{\beta})\xi_j - 2\xi_j^\top \nabla_{\alpha\beta_j} \ell(y, \Psi\hat{\alpha} + Z\hat{\beta}) \right] \right\} + \lambda_\xi \sum_{j=0}^{n-1} \|\xi_j\|_2. \quad (5.12)$$

The one-step estimator is then defined as

$$\hat{b} := \hat{\beta} - \hat{I}_{\beta|\alpha}^{-1} \hat{S}(\hat{\beta}, \hat{\alpha}), \quad (5.13)$$

where $\hat{I}_{\beta|\alpha} = \nabla^2_{\beta\beta} \ell(y, \Psi\hat{\alpha} + Z\hat{\beta}) - \hat{\xi}^\top \nabla^2_{\alpha\beta} \ell(y, \Psi\hat{\alpha} + Z\hat{\beta})$.

We will establish in Section 5.4.2 that $\sqrt{n}\hat{I}_{\beta|\alpha}^{1/2}(\hat{b} - \beta^*)$ converges to a standard Normal distribution. Thus, denoting $\hat{b} = (\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_{n-1})$ such that the indices are the same as the entries of β, we can construct (1-α) confidence intervals for causal estimands as linear transformations of $\hat{b}$, and particularly, for $\bar{\tau}_{\text{ADE}}$ as

$$\hat{b}_0 \pm q_{1-\frac{\alpha}{2}} n^{-\frac{1}{2}} \left[\hat{I}_{\beta|\alpha}^{-\frac{1}{2}} \right]_{11},$$

and for $\bar{\tau}_{\text{AIE}}$ as

$$\sum_{s=1}^{n-1} \hat{b}_s \pm q_{1-\frac{\alpha}{2}} n^{-\frac{1}{2}} \gamma^\top \hat{I}_{\beta|\alpha}^{-\frac{1}{2}} \gamma,$$

where $\gamma = (0, 1, \ldots, 1)$.

5.3.4 Inference on the Natural Scale

In the linear case where the link function g is identity, $\bar{\tau}^{\text{N}}_{\text{ADE}}$ and $\bar{\tau}^{\text{N}}_{\text{AIE}}$ simply reduces to $\bar{\tau}_{\text{ADE}}$ and $\bar{\tau}_{\text{AIE}}$. In other cases, to conduct statistical inference on the natural scale effects $\bar{\tau}^{\text{N}}_{\text{ADE}}$ and $\bar{\tau}^{\text{N}}_{\text{AIE}}$, a natural idea is to use the delta method to transform the estimated parameters $(\hat{\alpha}, \hat{\beta})$ and obtain natural scale effects. However, this requires assessing the variability of $\hat{\alpha}$ which is not covered in Section 5.3.3 and may lead to efficiency loss due to the high dimensionality of α.

Instead, we propose a parametric bootstrap procedure exploiting the conditional independence between the Y_i's given the graphical structure along with the treatment assignment probabilities. In particular, since Y only depends on the the graph G_n via the numbers of treated s-neighbors and network degrees of each individual, we could obtain valid bootstrap confidence intervals by permuting the edges in G_n and re-assigning the treatment w_i's, as long as the re-sampled network

[1] this indexing (starting from 0) is adopted because it aligns with the indexing of β

degrees, nuisance variables and treatment assignments follow the same distribution as in the true data generating mechanism. This procedure is described in Algorithm 3.

5.4 Theoretical Guarantees

This section presents theoretical properties of our causal estimates. We start by introducing necessary assumptions on the causal mechanism and random graph model, providing interpretations and explanations on each of them. Then, we show the consistency of $\hat{\tau}_{\text{ADE}}$ and $\hat{\tau}_{\text{AIE}}$, which guarantees the consistency of $\hat{\tau}^{\text{N}}_{\text{ADE}}$ and $\hat{\tau}^{\text{N}}_{\text{AIE}}$ under the continuity of link function g. Finally, we establish the asymptotic normality of our one-step estimator based on the decorrelated score function, enabling valid statistical inference for the direct and indirect causal effects. Proofs for theoretical results are provided in Appendix D.1.

5.4.1 Consistency

We first summarize and formalize the assumptions we introduced in Section 5.2 on the causal mechanism.

Assumption 5.1 (Causal mechanism)**.** *Given any treatment assignment vector $W \in \{0,1\}^n$, the observed outcomes are $Y_i = Y_i(W)$ where the expectations $\mathbb{E}[Y_i(W) \mid W]$ satisfy the mean model (5.1).*

Assumption 5.2 (Approximate neighborhood interference (ANI))**.** *For all n, there exists $m = o(\sqrt{n/\log n})$ such that $|\sum_{s=m+1}^{n-1} \beta^*_s| = o_P(1)$ and $R_\beta(\beta^*_{m+1,\ldots,n-1}) = o_P(1)$, where $\beta^*_{m+1,\ldots,n-1}$ represents the vector consisting of entries $(m+1)$ through $(n-1)$ of β^*.*

Assumption 5.2 is different from the ANI assumption in Leung (2022), but similarly states that the interference from "distant" alters, i.e., those more than distance m away, has negligible impact on the ego's potential outcome. This allows us to restrict our "target parameter" for the treatment effects β, i.e., the solution to the population version of (5.6), close to a lower-dimensional space which facilitates our estimation and inference. More fundamentally, this assumption also guarantees that the causal estimand $\bar{\tau}_{\text{AIE}}$ is well-defined for increasing n.

In addition to the causal mechanism, we also require assumptions on the optimization problem (5.6), including the experimental design giving rise to the treatment assignment vector $W \in \{0,1\}^n$,

the function class $\mathcal{F}$ as well as the loss function $\ell(\cdot)$. We start with assumptions on the covariates $X_1, \ldots, X_K$ and the design matrix $Z := \left[w, \frac{M_1}{N_1}, \ldots, \frac{M_{n-1}}{N_{n-1}}\right]$, which depends on the experimental design and graph structure:

Assumption 5.3 (Design matrix and graph structure)**.** *The nuisance covariates $X_1, \ldots, X_K$ are deterministic. Entries of the treatment assignment vector $W \in \{0,1\}^n$ are independent Bernoulli random variables with probabilities $\pi_i = \phi(x_{i1}, \ldots, x_{iK})$, respectively, for a function $\phi(\cdot)$. Furthermore, for each $s = 1, \ldots, n-1$ and $i = 1, \ldots, n$, we have that the Y_i's are uniformly sub-Gaussian conditioning on the treatment assignment W and graph G_n, i.e.*

$$\max_{i=1,\ldots,n} K_0^2 \cdot \mathbb{E}_{Y|W,G_n}\left[\exp\left(\frac{(Y_i W_i - \mathbb{E}[Y_i W_i])^2}{K_0^2} - 1\right) \;\middle|\; W, G_n\right] \leq \sigma_0^2$$

for some constant $K_0, \sigma_0^2 > 0$.

The last condition in Assumption 5.3 can be satisfied by a wide range of models for the potential outcome, and specifically, it only requires sub-Gaussian tails of each Y_i conditioning on the treatment assignment W and network G_n, and hence the regression coefficients. We note that the dependence between the Y_i's only comes from W and G_n, and exploit the conditional independence in the development of our technical results, so that we could leverage existing work under the penalized estimation framework and adapt to this dependent setting.

Next, we introduce some basic regularity conditions that are needed for our estimation procedure, where the last condition is needed only for the calculation of GCV score, and can be omitted if parameter tuning is done with alternative approaches.

Assumption 5.4 (Regularity conditions)**.**

1. *The loss function $\ell(\cdot)$ is convex and can be expressed as*

$$\ell(y, \eta) = \zeta \cdot y\eta + h(\eta)$$

   where $\eta := \sum_{k=1}^K f_k(x_k) + w\beta_0 + \sum_{s=1}^{n-1} \frac{M_s}{N_s}\beta_s$ is the linear predictor, ζ is an unknown model parameter, and $h(\cdot)$ is twice differentiable with bounded first-order derivative, i.e. $|h'(\eta)| < M$.

2. *The true causal parameters satisfy $\|\beta^*\|_\infty < R$ for $R > 0$ not depending on n.*

3. *(For GCV) For all i: the expectation $\mathbb{E}_\mu \left| \left[\frac{\partial \hat{\mu}_i(Y)}{\partial \mu_i} \right] \right| < \infty$. Also, there exists at least one combination of parameters $(\alpha^\dagger, \beta^\dagger)$ such that letting $\mu_i^\dagger = \Psi\alpha^\dagger + Z\beta^\dagger$, the expectation $\mathbb{E}_{\mu^\dagger}|\hat{\mu}_i(Y)| < \infty$. Finally, $\frac{\partial \mathbb{E}_\mu[\hat{\mu}_i(Y)]}{\partial \mu_i}$ is continuous in μ_i.*

Assumptions 5.4-1 and 5.4-2 or their similar variants are standard in penalized regression literature (e.g., Haris et al., 2022; Van de Geer, 2008). Assumption 5.4-3 is needed for the exchangeability of differentiation and integration (Giga et al., 2010) in (5.9), justifying our estimate for GDF and therefore the GCV score.

Assumption 5.4-2 restricts the space that the causal parameters β fall into. A similar notion on the "size", or complexity, of the function space $\mathcal{F}$ is needed to ensure accurate joint estimation of the nuisance f_k and parameters β. To formalize this, for a function space $\mathcal{F}$ equipped with a pseudometric[3] $d(\cdot, \cdot)$, we define

- ϵ-*cover*: a set $\mathcal{F}_1 \subset \mathcal{F}$ such that for any $f \in \mathcal{F}$, there exists $f_1 \in \mathcal{F}_1$ such that $d(f, f_1) \leq \epsilon$;
- ϵ-*covering number*: $N(\epsilon, \mathcal{F}, d) := \min\{|\mathcal{F}_1| : \mathcal{F}_1 \text{ is an } \epsilon\text{-cover of } \mathcal{F}\}$;
- ϵ-*entropy*: $H(\epsilon, \mathcal{F}, d) := \log N(\epsilon, \mathcal{F}, d)$

following Van der Vaart (2000). Intuitively, a small ϵ-covering number or ϵ-entropy of $\mathcal{F}$ means that all functions within $\mathcal{F}$ can be approximated reasonably well with a small set of functions, and $\mathcal{F}$ is thus of smaller "size" or complexity.

Assumption 5.5 (Logarithmic entropy of $\mathcal{F}$)**.** *For each $k = 1, \ldots, K$ and any $\lambda_\mathcal{F} > 0$, we have that*

$$H\left(\epsilon, \{f_k \in \mathcal{F} : \lambda_\mathcal{F} R_\mathcal{F}(f_j) \leq 1, d_{j,r}\}\right) \leq A_0 T_n \log\left(\epsilon^{-1} + 1\right)$$

for A_0 and T_n, where the pseudometric $d_{j,r}$ is defined such that $d_{j,r}(f, g) := \left(\frac{1}{n}\sum_{i=1}^n (f(x_{ij}) - g(x_{ij}))^r\right)^{1/r}$.

The logarithmic entropy bound is a general bound that holds for most finite dimensional function classes (of dimension T_n), e.g. the L_2 function space (Haris et al., 2022). Though we impose

[3]i.e., satisfying all requirements of a metric except that $d(x, y) = 0$ does not necessarily imply $x = y$.

restrictions on the function class $\mathcal{F}$, we do not require the true f_k^*'s, or the minimizer of the population version of (5.6), to fall within $\mathcal{F}$.

We introduce the shorthand notation for the loss function, $L(f, \beta) := \ell\left(y,\ \sum_{k=1}^{K} f_k(x_k) + w\beta_0 + \sum_{s=1}^{n-1} \frac{M_s}{N_s}\beta_s\right)$, and define the *empirical process term* as

$$\nu_n(f, \beta) := (\mathbb{P}_n - \mathbb{P}_0)L(f, \beta).$$

In addition, we define the *excess risk* as

$$\mathcal{E}(f, \beta) := \mathbb{P}_0[L(f, \beta) - L(\tilde{f}, \tilde{\beta})],$$

where $\tilde{f}_1, \ldots, \tilde{f}_K \in \mathcal{F}$ satisfy $\tilde{f}_k(x_{ik}) - f_k^*(x_{ik}) = 0$ for $i = 1, \ldots, n$ and $k = 1, \ldots, K$, and the *target causal parameter* $\tilde{\beta}$ is such that

$$\tilde{\beta} := \arg\min_{\beta} \mathbb{P}_0 \ell\left(y,\ \sum_{k=1}^{K} \tilde{f}_k(x_k) + w\beta_0 + \sum_{s=1}^{n-1} \frac{M_s}{N_s}\beta_s\right).$$

In other words, $\tilde{\beta}$ is the minimizer of the population risk with respect to β, with each nuisance function f_k fixed at $\tilde{f}_k$. These two quantities will be used in our technical proofs. The next assumption on $\mathcal{E}(f, \beta)$ ensures that a small excess risk translates to small estimation error in f and β, and is common in the high-dimensional estimation literature on general convex loss functions (Negahban et al., 2012; Van de Geer, 2008).

Assumption 5.6 (Margin condition)**.** *There exists a strongly convex function G such that $G(0) = 0$ and for any function $f^0 \in \mathcal{F}, \beta^0 \in \mathbb{R}^n$ as well as any (f, β) that is in a neighborhood of (f^0, β^0), we have*

$$\mathcal{E}(f, \beta) \geq G(\|\beta - \beta^0\| + \|f - f^0\|_{\mathcal{F}})$$

for some vector norm $\|\cdot\|$ and some function norm $\|\cdot\|_{\mathcal{F}}$ on $\mathcal{F}$.

An additional assumption we require is the compatibility condition, which specifies the compatibility between the error norm (in our case, the L1 norm for β) and the regularizer used in estimation (in our case, the penalty $R_\beta(\cdot)$) over a certain parameter space.

Assumption 5.7 (Compatibility condition)**.**

$$R_\beta((\beta - \tilde{\beta})_{0,\ldots,m}) \leq \frac{\|\beta - \tilde{\beta}\|\sqrt{m}}{\phi(m)}$$

for all β such that $R((\beta - \tilde{\beta})_{m+1,\ldots,n-1}) \leq 3R((\beta - \tilde{\beta})_{0,\ldots,m})$, where $\|\cdot\|$ is the vector norm in Assumption 5.6 $\phi(m) > 0$ indicates the compatibility between the L1 norm and the regularizer R_β.

A common assumption for high-dimensional regression is the rates of tuning parameters, which we state in Assumption 5.8 below:

Assumption 5.8 (Rates of penalties)**.** *The penalties in (5.6) satisfy $\lambda_\beta = O_P(\sqrt{\log n/n})$ and $\lambda_\beta \phi(m) = O_P(1)$. Furthermore, $\lambda_k = O_P(\sqrt{\log n/n})$ for each k.*

Recalling that the dimensionality of the causal parameters β scale with n, the rates above align with the standard $\sqrt{\log p/n}$ rate for penalized regression estimates (where p is the dimensionality).

We now present our main theory on the consistency of $\hat{\beta}$, the solution to (5.6), and consequently $\hat{\tau}_{\text{ADE}}$, $\hat{\tau}_{\text{AIE}}$, $\hat{\tau}^{\text{N}}_{\text{ADE}}$ and $\hat{\tau}^{\text{N}}_{\text{AIE}}$.

Theorem 5.1 (Consistency of $\hat{\tau}_{\text{ADE}}$, $\hat{\tau}_{\text{AIE}}$, $\hat{\tau}^{\text{N}}_{\text{ADE}}$ and $\hat{\tau}^{\text{N}}_{\text{AIE}}$)**.** *Under Assumptions 5.1-5.8, the solution $(\hat{f}_1, \ldots, \hat{f}_K, \hat{\beta})$ to (5.6) satisfies*

$$\|\hat{\beta} - \beta^*\|_1 + \sum_k R_{\mathcal{F}}(\hat{f}_k - \tilde{f}_k) = O_P\left(\sqrt{m\frac{\log n}{n}}\right)$$

with probability at least $1 - C_1 \exp(-C_2 n\rho^2)$ for $\rho = O_P(\lambda_\beta) = O_P(\sqrt{\log n/n})$, and constants $C_1, C_2 > 0$ not depending on n, as $n \to \infty$.

Consequently, we have that $\hat{\tau}_{ADE} \xrightarrow{p} \bar{\tau}_{ADE}$, $\hat{\tau}_{AIE} \xrightarrow{p} \bar{\tau}_{AIE}$, $\hat{\tau}^N_{ADE} \xrightarrow{p} \bar{\tau}^N_{ADE}$, and $\hat{\tau}^N_{AIE} \xrightarrow{p} \hat{\tau}^N_{AIE}$ as $n \to \infty$.

5.4.2 Asymptotic Normality

In this section, we establish the asymptotic normality of the one-step estimator defined in (5.13), justifying our confidence intervals and hypothesis testing procedures. This requires a few additional assumptions, which are analogues of those in Ning and Liu (2017) and are introduced below for convenience. For two sequences a_n and b_n, we use $a_n \preceq b_n$ to denote that $a_n \leq C b_n$ for some constant C for large enough n.

Assumption 5.9 (Error bound of ξ)**.** *Define $\xi^* := {I^*_{\alpha\alpha}}^{-1} I^*_{\beta\alpha}$, where $I^*_{\alpha\alpha}$ and $I^*_{\beta\alpha}$ are partitions of the the information matrix evaluated at the truth β^* and population risk minimizer α^*. The*

solution $\hat{\xi}$ *to (5.12) satisfies*

$$\lim_{n\to\infty} \Pr\left(\|\hat{\xi} - \xi^*\|_1 \preceq \varepsilon_1(n)\right) = 1$$

for a sequence $\varepsilon_1(n) \to 0$ *as* $n \to \infty$.

Assumption 5.10 (Stability of Hessian)**.** *For* $\alpha_v := v\alpha^* + (1-v)\hat{\alpha}$ *with* $v \in [0,1]$,

$$\lim_{n\to\infty} \Pr\left(\sup_{v\in[0,1]} \left\|\nabla^2_{\beta\alpha}\ell(y, \Psi\alpha_v + Z\beta^*) - \hat{\xi}^\top \nabla^2_{\alpha\alpha}\ell(y, \Psi\alpha_v + Z\beta^*)\right\|_1 \preceq \varepsilon_2(n)\right) = 1$$

for a sequence $\varepsilon_2(n) \to 0$ *as* $n \to \infty$.

Assumption 5.11 (Convergence of Hessian)**.** *For* $\beta_v := v\beta^* + (1-v)\hat{\beta}$ *with* $v \in [0,1]$,

$$\lim_{n\to\infty} \Pr\left(\sup_{v\in[0,1]} \|\nabla^2_{\beta\beta}\ell(y, \Psi\hat{\alpha} + Z\beta_v) - I^*\|_1 \preceq \varepsilon_3(n)\right) = 1$$

for a sequence $\varepsilon_3(n) \to 0$ *as* $n \to \infty$.

Assumption 5.9 states that the optimization procedure of (5.12) leads to a solution ξ that is close to the partitions of information matrix evaluated at the truth. Assumption 5.10 translates to the stability of the Hessian matrix against small perturbations in the nuisance parameters α near the population risk minimizers. And in addition, Assumption 5.11 guarantees that the evaluation of Hessian matrix near the solution $(\hat{\alpha}, \hat{\beta})$ is close to the true information matrix despite perturbations in the causal parameters β. Following the framework of Ning and Liu (2017), we have that

Theorem 5.2 (Asymptotic normality of $\hat{b}$)**.** *Under Assumptions 5.1-5.11, and further assuming* $\|I^*_{\beta\alpha}\|_1\varepsilon_1(n) = o(1)$, $\sqrt{n}(\varepsilon_1(n) + \varepsilon_2(n)) = O_P(1)$ *and* $\|\xi^*\|_1\varepsilon_3(n) = o(1/\sqrt{\log n})$, *the one-step estimator* $\hat{b}$ *defined in (5.13) satisfies, for any weight vector* $v \in \mathbb{R}^n$ *with non-negative entries such that* $v^\top v = 1$,

$$\sqrt{n}v^\top \hat{I}^{1/2}_{\beta|\alpha}(\hat{b} - \beta^*) \xrightarrow{d} N(0,1).$$

Directly applying Theorem 5.2 to cases with $v = (1, 0, \ldots, 0)$ and $v = (0, 1/\sqrt{n-1}, \ldots, 1/\sqrt{n-1})$ establishes the asymptotic normality of $\hat{\tau}_{\text{ADE}}$ and $\hat{\tau}_{\text{AIE}}$, respectively, which justifies our construction of confidence intervals in Section 5.3.3.

5.5 Simulations

We illustration the validity of our estimation and inference approaches with numerical examples in this section. We focus on a linear setting with continuous Y and identity link $g(\cdot)$, along with a binary setting with logit $g(\cdot)$. For the latter case, we present inference results on both the original (i.e., $\hat{\tau}_{\text{AIE}}$, $\hat{\tau}_{\text{ADE}}$) and natural scale (i.e., $\hat{\tau}^{\text{N}}_{\text{AIE}}$, $\hat{\tau}^{\text{N}}_{\text{ADE}}$).

We investigated three random graph models with different levels of heterogeneity in network degree distribution, including the Erdős–Rényi (Erdős and Rényi, 1959), power law (Aiello et al., 2001) and small world (Watts and Strogatz, 1998) random graphs, implemented by the `igraph` `R` package. For each type of random graph, we considered three sparsity levels $p_n = O(n^{-r})$ with $r = \frac{1}{4}, \frac{1}{2}$ and $\frac{2}{3}$, respectively, where $p_n := \sum_{ij} E_{ij}/n^2$ is the average probabilities of edge connectivity. Note that by setting $r < 1$, we were allowing each individual to have an increasing number of neighbors as the network grows, which is not handled by some existing methods (e.g. Leung, 2020; van der Laan, 2014).

For both settings and each of the scenarios, we simulated $N = 100$ replicates of data with increasing sample size $n = 10, 100, 500, 1000$. $K = 2$ nuisance variables X_1, X_2 were included, both of which were sampled from Uniform$[0, 1]$, and we set the nuisance functions to be $f_1(x_1) = x_1^2$ and $f_2(x_2) = -x_2$. The binary treatments W were assigned such that $W_i \sim \text{Bernoulli}(\pi_i)$, with $\pi_i = 0.5 + 0.5x_1$, so that the randomization was imperfect with X_1 being the confounder. The causal parameter β^* was defined such that $\beta^*_1 = 1$, and $\beta^*_s = \exp(-s)$ for $s = 1, \ldots, n$. In other words, the interference decays exponentially in network distance. We used polynomial basis to estimate the nuisance functions f_1, f_2. Selection of tuning parameters was done by minimizing the GCV score developed in Section 5.3.2.

5.5.1 Linear Case

For the linear case, we generated the outcome $Y_i = Y_i(W)$ from Normal distribution with mean given by (5.1), and variance parameter $\sigma^2 = 0.1$. Recall that in this case, $\bar{\tau}_{\text{ADE}}$, $\bar{\tau}_{\text{AIE}}$ are equivalent to $\bar{\tau}^{\text{N}}_{\text{ADE}}$ and $\bar{\tau}^{\text{N}}_{\text{AIE}}$.

Figure 5.1 presents the estimation errors of $\hat{\tau}_{\text{ADE}}$ (top) and $\hat{\tau}_{\text{AIE}}$ (bottom) from our estimation procedure (5.6). The colors correspond to the average edge connectivity probabilities, where darker

color means overall sparser graphs. We observe that the magnitudes of estimation errors shrink towards zero as n increases, while convergence is faster for the direct effect estimate $\hat{\tau}_{\text{ADE}}$. This is expected since the estimation of indirect effects involves a larger number of parameters (potentially scaling with n) and therefore has relatively lower efficiency, compared to $\hat{\tau}_{\text{ADE}}$.

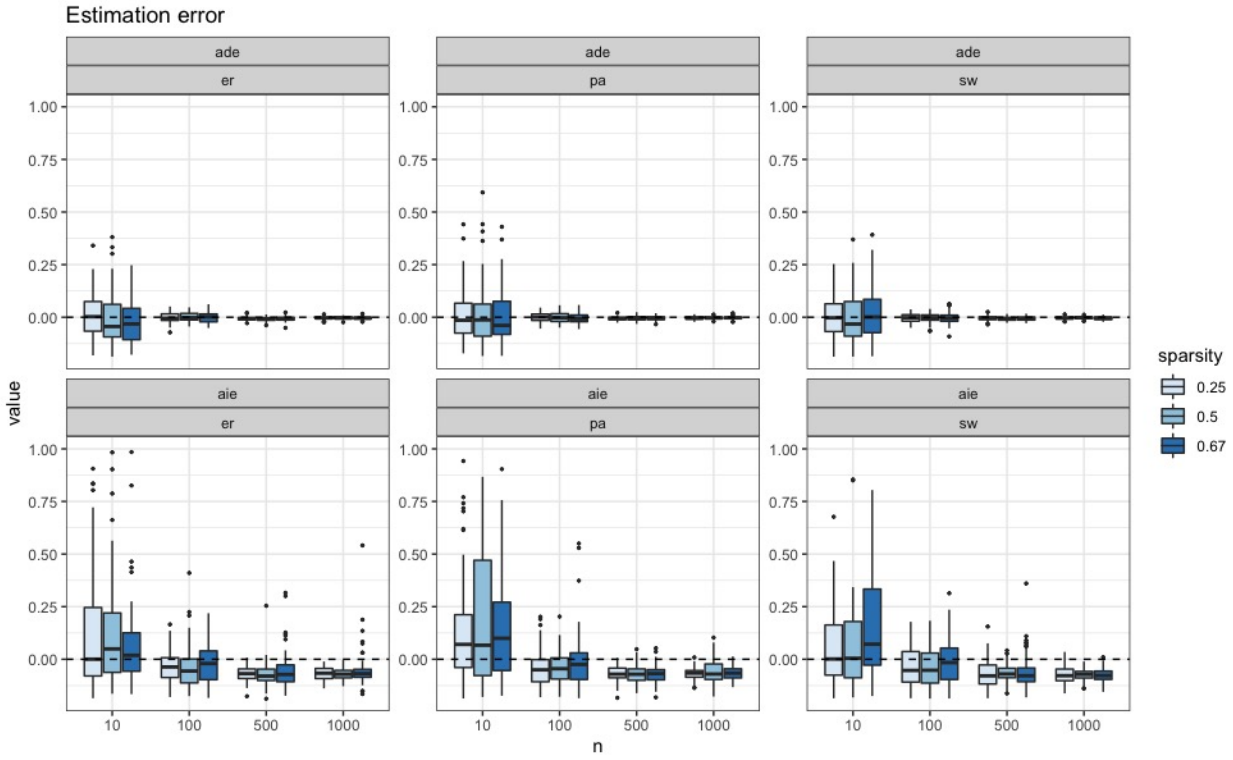


Figure 5.1: Distribution of estimation errors of $\hat{\tau}_{\text{ADE}}$ (top) and $\hat{\tau}_{\text{AIE}}$ (bottom) in the linear case, for Erdős–Rényi (left), power law (middle) and small world (right) random graphs. The color corresponds to the exponent r in the average edge probabilities $p_n = O(n^{-r})$.

To verify the validity of our inference procedure proposed in Section 5.3.3, we calculate the Z scores by scaling the one-step estimates with the derived asymptotic variance in (5.13), and compare with the theoretical quantiles of the standard Normal distribution. Figure 5.2 visualizes this comparison. We observe close alignment with the $y = x$ dashed, reference line, indicating accurate characterization for the asymptotic distribution of $\hat{b}$, even with moderate sample sizes ($n = 100$). Also, for sparser graphs (the darkest color), the longest network length is most likely larger, translating to a higher degree-of-freedom for the inference of $\bar{\tau}_{\text{AIE}}$. Consequently, we observe

signs of slightly conservative inference at this sparsity level for $\bar{\tau}_{\text{AIE}}$, reflected by the larger quantiles falling below the reference line in the rightmost column. In general, our estimation and inference approaches are valid with all three random graph models, demonstrating their flexibility when there is heterogeneity in the degree distribution of the networks.

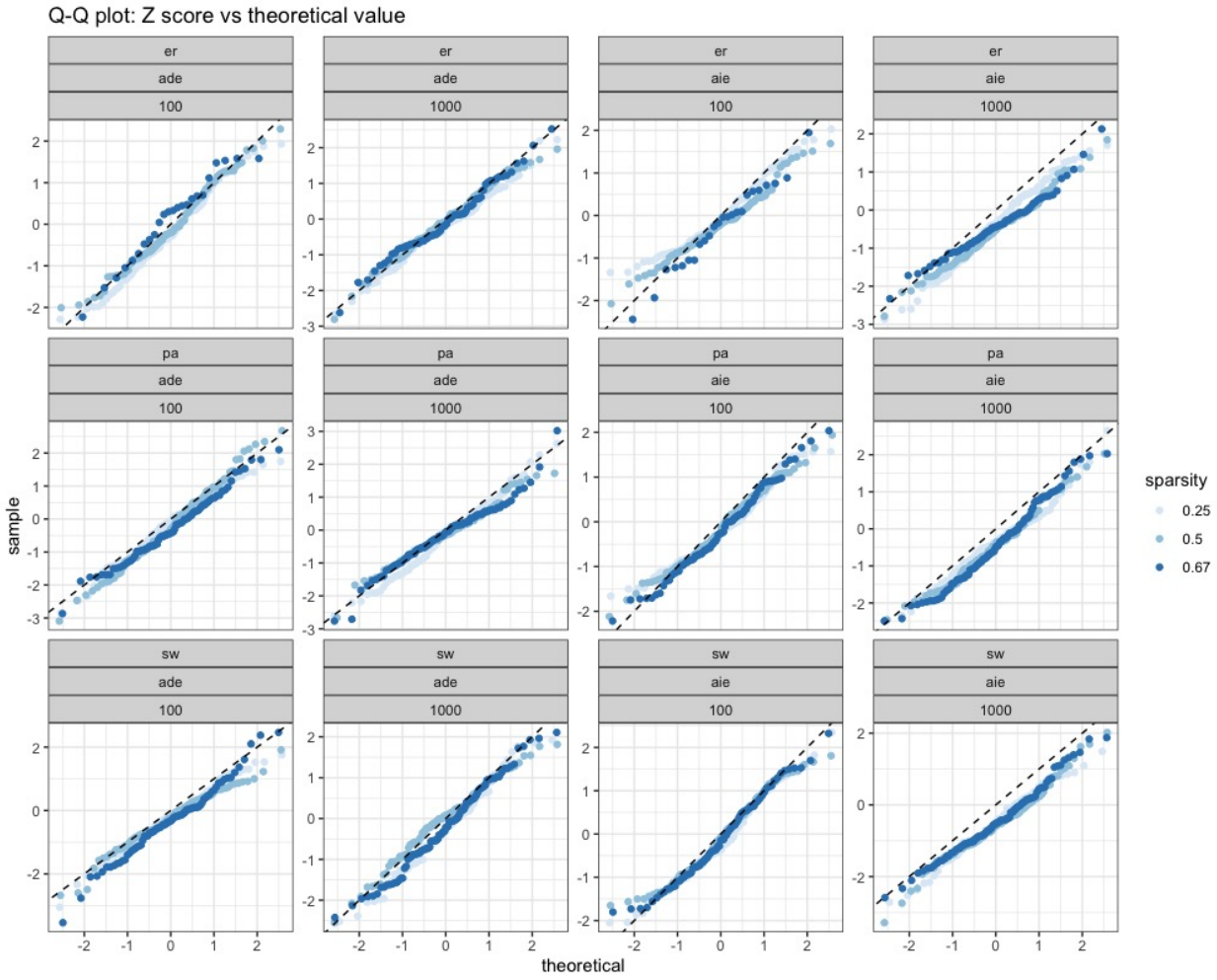


Figure 5.2: Q-Q plot of Z scores versus theoretical quantiles of a Normal distribution in the linear case, for the one-step estimators of $\bar{\tau}_{\text{ADE}}$ (the 1st and 2nd columns) and $\bar{\tau}_{\text{AIE}}$ (the 3rd and 4th columns). We compare small ($n = 100$, the 1st and 3rd columns) and large ($n = 1000$, the 2nd and 4th columns) sample sizes.

5.5.2 Binary Case

In the second setting, we simulated binary outcomes $Y_i = Y_i(W)$ from Bernoulli($\mu_i(W)$), where the means $\mu_i(W)$ are given by (5.1) with link function $g(\mu) := \text{logit}(\mu) = \log(\mu/(1-\mu))$. Parametric bootstrap was conducted according to Algorithm 3 with $B = 100$ bootstrap samples to estimate the standard errors of $\hat{\tau}^{\text{N}}_{\text{ADE}}$ and $\hat{\tau}^{\text{N}}_{\text{AIE}}$.

Figure 5.3 shows the estimation errors of $\hat{\tau}_{\text{ADE}}$ and $\hat{\tau}_{\text{AIE}}$ on the original (logit) scale, where the estimation errors approach zero with growing sample sizes, and $\hat{\tau}_{\text{ADE}}$ showing faster convergence as in the linear case.

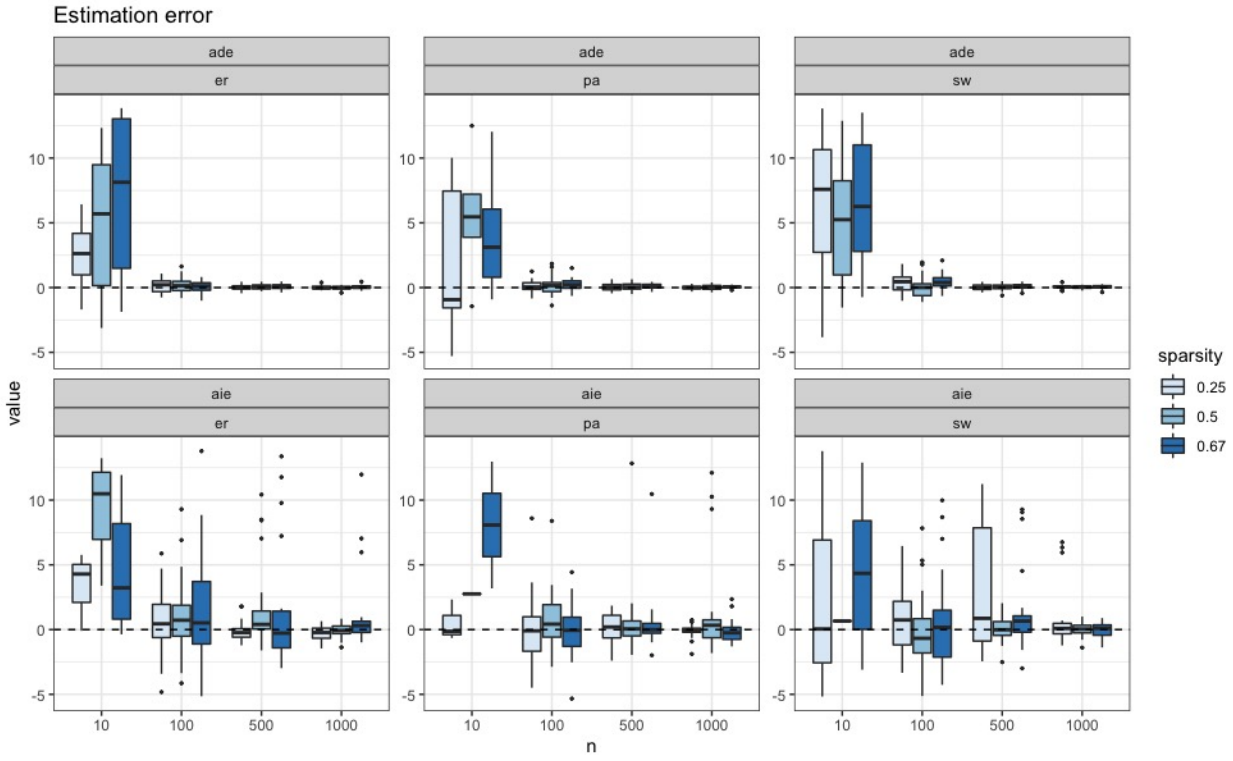


Figure 5.3: Distribution of estimation errors of $\hat{\tau}_{\text{ADE}}$ (top) and $\hat{\tau}_{\text{AIE}}$ (bottom) in the binary case, for Erdős–Rényi (left), power law (middle) and small world (right) random graphs. The color corresponds to the exponent r in the average edge probabilities $p_n = O(n^{-r})$.

Comparison between the Z scores to Normal quantiles in Figure 5.4 illustrates the validity of our inference procedure based on the decorrelated score function. In particular, we observe the

empirical quantiles aligning more closely to the reference line when the sample size grows from $n = 100$ to 1000.

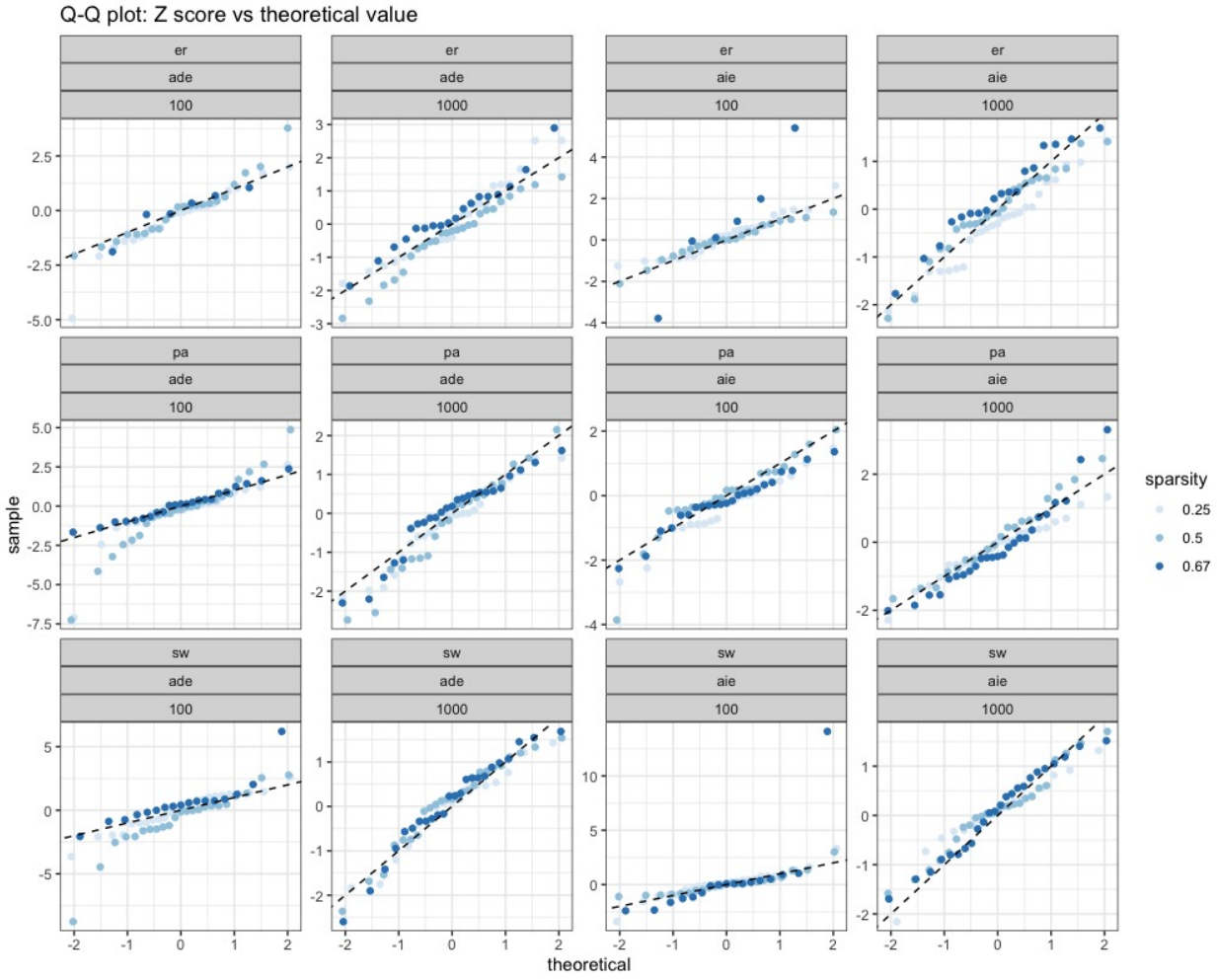


Figure 5.4: Q-Q plot of Z scores versus theoretical quantiles of a Normal distribution in the binary case, for the one-step estimators of $\bar{\tau}_{\text{ADE}}$ (the 1st and 2nd columns) and $\bar{\tau}_{\text{AIE}}$ (the 3rd and 4th columns). We compare small ($n = 100$, the 1st and 3rd columns) and large ($n = 1000$, the 2nd and 4th columns) sample sizes.

We then examine the natural scale estimates $\hat{\tau}^{\text{N}}_{\text{ADE}}$ and $\hat{\tau}^{\text{N}}_{\text{AIE}}$. Figure 5.5 is an analog of the error plot in Figure 5.3, showing the convergence of $\hat{\tau}^{\text{N}}_{\text{ADE}}$ and $\hat{\tau}^{\text{N}}_{\text{AIE}}$ to the truth with increasing network size. This is guaranteed by the continuity of $g^{-1}(\cdot)$, despite that $\hat{\tau}^{\text{N}}_{\text{ADE}}$ and $\hat{\tau}^{\text{N}}_{\text{AIE}}$ converge more slowly compared to the logit scale estimates.

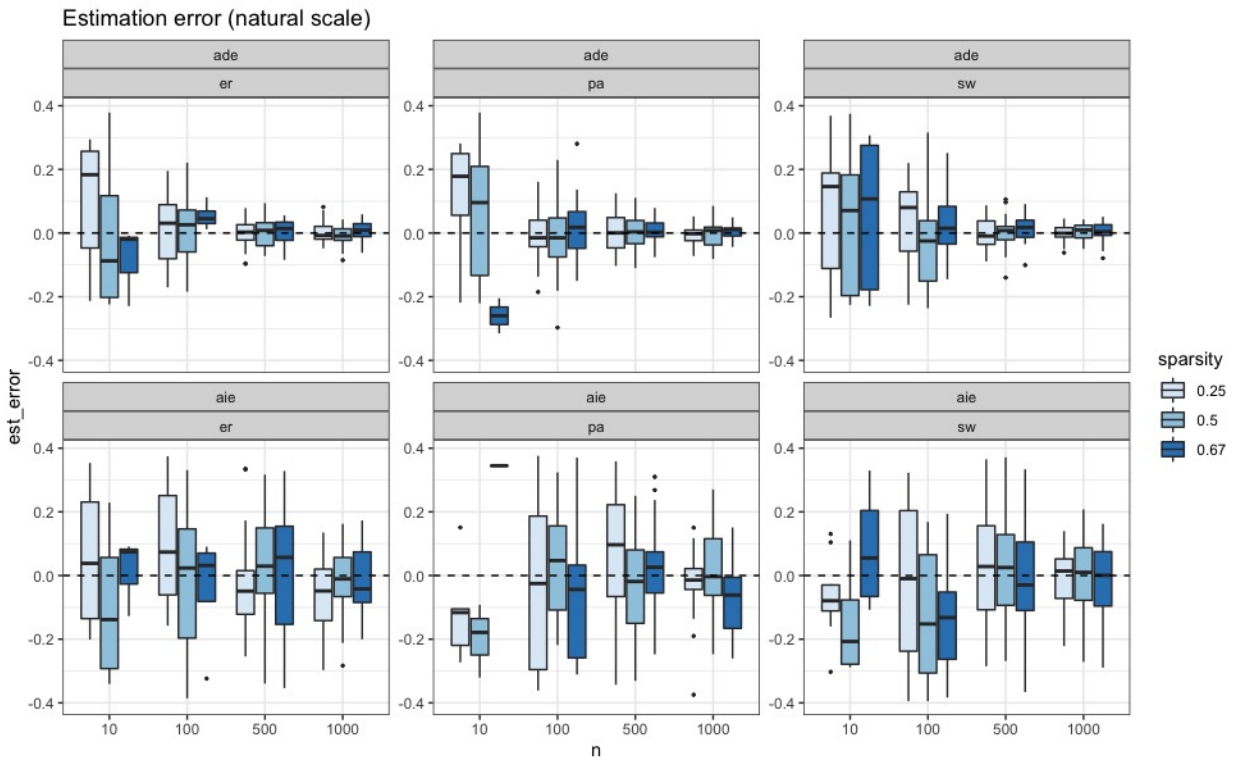


Figure 5.5: Distribution of estimation errors of the natural scale effects $\hat{\tau}^{\mathrm{N}}_{\mathrm{ADE}}$ (top) and $\hat{\tau}^{\mathrm{N}}_{\mathrm{AIE}}$ (bottom) in the binary case, for Erdős–Rényi (left), power law (middle) and small world (right) random graphs. The color corresponds to the exponent r in the average edge probabilities $p_n = O(n^{-r})$.

Figure 5.6 illustrates the parametric bootstrap procedure described in Algorithm 3 estimating the standard errors of the natural scale causal effects $\bar{\tau}^{\mathrm{N}}_{\mathrm{ADE}}$ and $\bar{\tau}^{\mathrm{N}}_{\mathrm{AIE}}$. The bootstrap and empirical standard errors are plotted with dashed and solid lines, respectively, against the sample size n on the log scale. We observe that the estimated standard errors approach the empirical ones with sufficient sample size, despite that results for the power law graph are noisier and likely require a larger sample size to stabilize.

5.6 Discussion

We developed a semi-parametric framework for the estimation and inference of direct and indirect causal effects in network settings, with the presence of cross-unit interference. This framework applies to a broad class of causal mechanisms with the expected potential outcome specified by

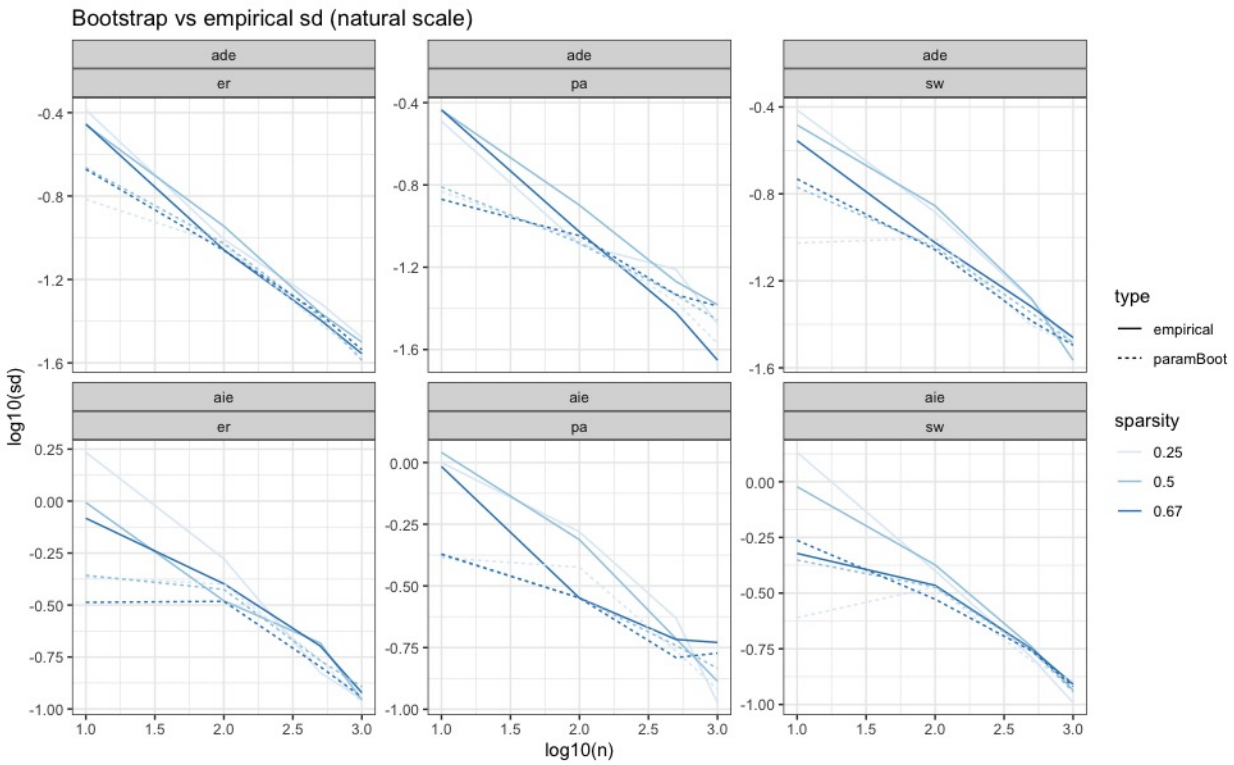


Figure 5.6: Comparison of bootstrap (dashed lines) versus empirical (solid lines) standard errors for the natural scale estimates in the binary case, where the SDs and sample size n are both plotted on the log scale.

a combination of non-parametric nuisance effects and parametric treatment effects (Eq. 5.1). In particular, the indirect treatment effect could spread through all connected paths over the network as opposed to direct neighbors only, which is more general than many existing methodologies. We also considered the network as a realization of a random graph model, imposing minimal assumptions on its generating mechanism.

Our estimation procedure belongs to the class of penalized regression methods (e.g., Haris et al., 2022), where we impose smoothing penalties on the nuisance functions and a (near-)sparsity penalty on the potentially high-dimensional causal parameter. We proposed a GCV criterion that applies to non-linear and dependent data, facilitating the simultaneous tuning for multiple parameters with the presence of network correlation. We leveraged the decorrelated score function proposed by Ning and Liu (2017) to conduct efficient inference on the causal parameters of interest with

the presence of high-dimensional nuisance. For non-linear mean models with non-identity (more precisely, non-collapsible) link functions, a parametric bootstrap algorithm was developed to handle the inference of natural scale causal effects. Extensive simulations in both linear and non-linear settings with a variety of random graph models and varying levels of sparsity illustrate the validity of our approach.

In the parametric bootstrap procedure described by Algorithm 3, we assumed that the distribution of network degrees is readily available for us to sample from. However, this is not always guaranteed in practical settings. For example, in our simulation studies for the binary case (Section 5.5.2), we conducted re-sampling by permuting the edges in each graph G_n while preserving the original degree distribution, as implemented by the `simple` algorithm of `igraph`. This approach could lead to self edges or multiple edges, which are not allowed in our causal mechanism. We therefore trimmed these edges, causing the actual degree distribution to shift slightly to the lower end compared to the original mechanism (Figure 5.7). The difference, nevertheless, is very small and Figure 5.6 still suggests reasonable performance despite the slightly distorted degree distribution. Sophisticated graph re-sampling algorithms could provide more theoretically justified inference, likely at the cost of higher computational complexity.

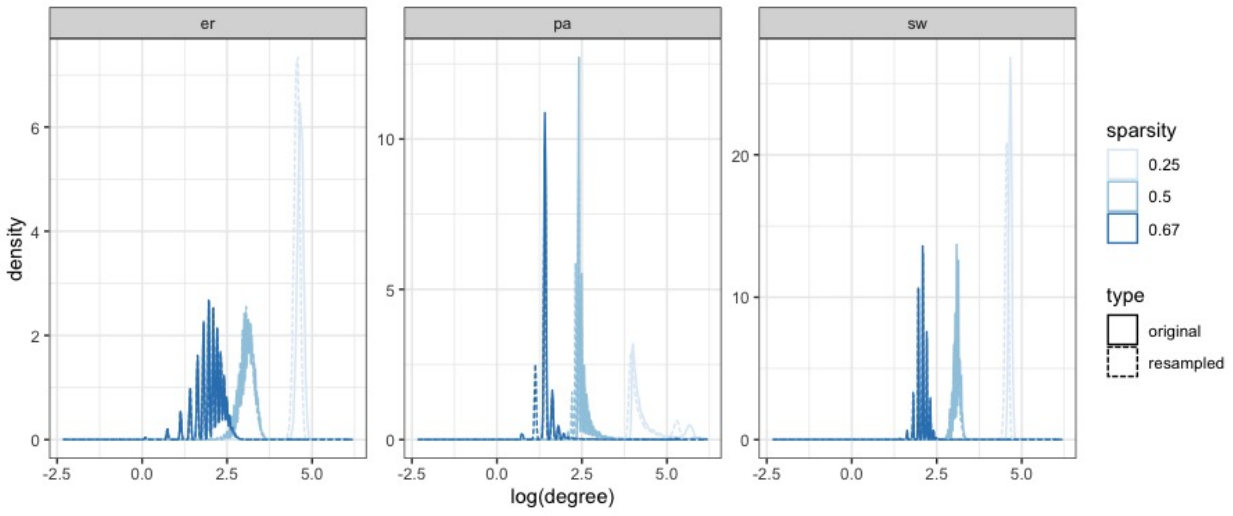


Figure 5.7: Densities of the original (solid) and re-sampled (dashed) degree distributions across all bootstrap samples (slightly jittered), plotted on the log scale, for Erdős–Rényi (left), power law (middle) and small world (right) random graphs.

Algorithm 3: Parametric bootstrap for $\bar{\tau}^{\mathrm{N}}_{\mathrm{ADE}}$ and $\bar{\tau}^{\mathrm{N}}_{\mathrm{AIE}}$

Input: estimated treatment assignment rule $\hat{\pi} = \hat{\pi}(x_1, \ldots, x_k)$; estimated causal parameters $\hat{\beta}$; estimated nuisance functions $\hat{f}_k$; desired number of bootstrap samples B

for $t = 0, 1, \ldots B$ **do**

for $i = 1, \ldots, n$ **do**

Generate nuisance variables from their empirical distributions $\{x_{ik}^{(t)}\}_{k=1}^{K} \sim \mathbb{P}_n^X$

Assign treatment indicators as $w_i^{(t)} \leftarrow \hat{\pi}\left(x_{i1}^{(t)}, \ldots, x_{iK}^{(t)}\right)$

Sample a new graph $G_n^{(t)}$ with the same degree distribution[2] as G_n: $N_{i1}^{(t)} \sim \mathbb{P}_n^G$

Calculate the number of s-neighbors as $N_{is}^{(t)} \leftarrow \sum_j I\{\ell_{ij}^{(t)} = s\}$

Calculate the nuisance effects $\{\hat{f}_k(x_{ik}^{(t)})\}_{k=1}^{K}$

end

for $i' = 1, \ldots, n$ **do**

Calculate the number of treated s-neighbors as $M_{i's}^{(t)} \leftarrow \sum_{j:\ell_{i'j}^{(t)}=s} w_j^{(t)}$

Calculate the treatment effects $w_{i'}^{(t)}\hat{\beta}_0 + \sum_{s=1}^{n-1} \frac{M_{i's}^{(t)}}{N_{i's}^{(t)}}$

Generate the potential outcome $Y_{i'}^{(t)}(w_{i'}, W_{-i'})$

end

Solve (5.6) with the newly generated data plugged in:

$$(\hat{f}_1^{(t)}, \ldots, \hat{f}_k^{(t)}, \hat{\beta}^{(t)}) \leftarrow \underset{f_1,\ldots,f_k\in\mathcal{F},\beta}{\operatorname{argmin}} \mathbb{P}_n \ell\left(y^{(t)}, \sum_{k=1}^{K} f_k(x_k^{(t)}) + w^{(t)}\beta_0 + \sum_{s=1}^{n-1} \frac{M_s^{(t)}}{N_s^{(t)}}\beta_s\right) + \lambda_\beta R_\beta(\beta) + \sum_{k=1}^{K} \lambda_k R_{\mathcal{F}}(f_k)$$

Calculate the natural scale direct effect

$$\hat{\tau}^{\mathrm{N}}_{\mathrm{ADE}}{}^{(t)} \leftarrow \frac{1}{n}\sum_i \hat{\mu}_i(w_i^{(t)} = 1; W_{-i}^{(t)}) - \hat{\mu}_i(w_i^{(t)} = 0; W_{-i}^{(t)})$$

Calculate the natural scale indirect effect

$$\hat{\tau}^{\mathrm{N}}_{\mathrm{AIE}}{}^{(t)} \leftarrow \frac{1}{n}\sum_i \sum_{j\neq i} \hat{\mu}_j(w_i^{(t)} = 1; W_{-i}^{(t)}) - \hat{\mu}_j(w_i^{(t)} = 0; W_{-i}^{(t)})$$

end

Result: Empirical distributions $\hat{F}_{\mathrm{ADE}}$, $\hat{F}_{\mathrm{AIE}}$ of the bootstrap samples $\{\hat{\tau}^{\mathrm{N}}_{\mathrm{ADE}}{}^{(t)}\}_{t=1}^{B}$ and $\{\hat{\tau}^{\mathrm{N}}_{\mathrm{AIE}}{}^{(t)}\}_{t=1}^{B}$

BIBLIOGRAPHY

Souzana Achilleos, Marianthi-Anna Kioumourtzoglou, Chih-Da Wu, Joel D Schwartz, Petros Koutrakis, and Stefania I Papatheodorou. Acute effects of fine particulate matter constituents on mortality: A systematic review and meta-regression analysis. *Environment International*, 109: 89–100, 2017.

Yash P Aggarwal, Lynn R Sykes, David W Simpson, and Paul G Richards. Spatial and temporal variations in ts/tp and in P wave residuals at Blue Mountain Lake, New York: Application to earthquake prediction. *Journal of Geophysical Research*, 80(5):718–732, 1975.

William Aiello, Fan Chung, and Linyuan Lu. A random graph model for power law graphs. *Experimental Mathematics*, 10(1):53–66, 2001.

John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

Alma Andersson, Ludvig Larsson, Linnea Stenbeck, Fredrik Salmén, Anna Ehinger, Sunny Z Wu, Ghamdan Al-Eryani, Daniel Roden, Alex Swarbrick, Åke Borg, et al. Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nature Communications*, 12(1):6012, 2021.

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.

Luc Anselin. *Spatial econometrics: methods and models*, volume 4. Springer Science & Business Media, 1988.

Larry Armijo. Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16(1):1–3, 1966.

Patrick Arnaud, Christophe Bouvier, Leonardo Cisneros, and Ramon Dominguez. Influence of rainfall spatial variability on flood prediction. *Journal of Hydrology*, 260(1-4):216–230, 2002.

Peter M Aronow and Cyrus Samii. Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4):1912–1947, 2017.

Adrian Baddeley and Gopalan Nair. Fast approximation of the intensity of Gibbs point processes. *Electronic Journal of Statistics*, 6:1155–1169, 2012.

Sarah Baird, J Aislinn Bohren, Craig McIntosh, and Berk Özler. Optimal design of experiments in the presence of interference. *Review of Economics and Statistics*, 100(5):844–860, 2018.

Sudipto Banerjee. High-dimensional Bayesian geostatistics. *Bayesian Analysis*, 12(2):583, 2017.

Maurice Stevenson Bartlett. The spectral analysis of point processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 25(2):264–281, 1963.

Amir Behnamian, Koreen Millard, Sarah N Banks, Lori White, Murray Richardson, and Jon Pasher. A systematic approach for variable selection with random forests: achieving stable variable importance values. *IEEE Geoscience and Remote Sensing Letters*, 14(11):1988–1992, 2017.

Katja M Bendtsen, Elizabeth Bengtsen, Anne T Saber, and Ulla Vogel. A review of health effects associated with exposure to jet engine emissions in and around airports. *Environmental Health*, 20(1):1–21, 2021.

Silas Bergen, Lianne Sheppard, Paul D Sampson, Sun-Young Kim, Mark Richards, Sverre Vedal, Joel D Kaufman, and Adam A Szpiro. A national prediction model for PM2.5 component exposures and measurement error–corrected health effect inference. *Environmental Health Perspectives*, 121(9):1017–1025, 2013.

Silas Bergen, Lianne Sheppard, Joel D Kaufman, and Adam A Szpiro. Multipollutant measurement error in air pollution epidemiology studies arising from predicting exposures with penalized regression splines. *Journal of the Royal Statistical Society. Series C, Applied statistics*, 65(5): 731, 2016.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

Stefania Bertazzon, Markey Johnson, Kristin Eccles, and Gilaad G Kaplan. Accounting for spatial effects in land use regression for urban air pollution modeling. *Spatial and Spatio-Temporal Epidemiology*, 14:9–21, 2015.

Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225, 1974.

Julian Besag, Jeremy York, and Annie Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–20, 1991.

Nicky Best, Sylvia Richardson, and Andrew Thomson. A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, 14(1):35–59, 2005.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Magali N Blanco. *Traffic-related Air Pollution and Dementia Incidence in a Seattle-based, Prospective Cohort Study*. PhD thesis, University of Washington, 2021.

Magali N Blanco, Annie Doubleday, Elena Austin, Julian D Marshall, Edmund Seto, Timothy V Larson, and Lianne Sheppard. Design and evaluation of short-term monitoring campaigns for long-term air pollution exposure assessment. *Journal of Exposure Science & Environmental Epidemiology*, pages 1–9, 2022a.

Magali N Blanco, Amanda Gassett, Timothy Gould, Annie Doubleday, David L Slager, Elena Austin, Edmund Seto, Timothy V Larson, Julian D Marshall, and Lianne Sheppard. Characterization of annual average traffic-related air pollution concentrations in the greater Seattle area from a year-long mobile monitoring campaign. *Environmental Science & Technology*, 2022b.

Robert M Bond, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298, 2012.

H Boogaard, AP Patton, RW Atkinson, JR Brook, HH Chang, DL Crouse, JC Fussell, G Hoek, B Hoffmann, R Kappeler, et al. Long-term exposure to traffic-related air pollution and se-

lected health outcomes: A systematic review and meta-analysis. *Environment International*, page 107262, 2022.

Maitreyee Bose, Timothy Larson, and Adam A Szpiro. Adaptive predictive principal components for modeling multivariate air pollution. *Environmetrics*, 29(8):e2525, 2018.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

HL Brantley, GSW Hagler, ES Kimbrough, RW Williams, S Mukerjee, and LM Neas. Mobile air monitoring data-processing strategies and effects on spatial air pollution trends. *Atmospheric Measurement Techniques*, 7(7):2169–2183, 2014.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Pierre Brémaud. *Point processes and queues: martingale dynamics*, volume 50. Springer, 1981.

Timothy R Brick, Rachel E Koffer, Denis Gerstorf, and Nilam Ram. Feature selection methods for optimal design of studies for developmental inquiry. *The Journals of Gerontology: Series B*, 73 (1):113–123, 2018.

David R Brillinger. The identification of point process systems. *The Annals of Probability*, pages 909–924, 1975.

Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Dieu Tien Bui, Mahdi Panahi, Himan Shahabi, Vijay P Singh, Ataollah Shirzadi, Kamran Chapi, Khabat Khosravi, Wei Chen, Somayeh Panahi, Shaojun Li, et al. Novel hybrid evolutionary algorithms for spatial prediction of floods. *Scientific reports*, 8(1):1–14, 2018.

Andreas Buja, Lawrence Brown, Richard Berk, Edward George, Emil Pitkin, Mikhail Traskin, Kai Zhang, and Linda Zhao. Models as approximations I: Consequences illustrated with linear regression. *Statistical Science*, 34(4):523–544, 2019.

Jing Cai, Alain De Janvry, and Elisabeth Sadoulet. Social networks and the decision to insure. *American Economic Journal: Applied Economics*, 7(2):81–108, 2015.

Liqian Cai and Tapabrata Maiti. Variable selection and estimation for high-dimensional spatial autoregressive models. *Scandinavian Journal of Statistics*, 47(2):587–607, 2020.

Xiaoxuan Cai, Wen Wei Loh, and Forrest W Crawford. Identification of causal intervention effects under contagion. *Journal of Causal Inference*, 9(1):9–38, 2021.

Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:'model-X'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

Raymond J Carroll, Jianqing Fan, Irene Gijbels, and Matt P Wand. Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92(438):477–489, 1997.

Michael Carter, Rachid Laajaj, and Dean Yang. Subsidies and the African Green Revolution: direct effects and social network spillovers of randomized input subsidies in Mozambique. *American Economic Journal: Applied Economics*, 13(2):206–229, 2021.

Carlos M Carvalho, Nicholas G Polson, and James G Scott. Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*, pages 73–80. PMLR, 2009.

Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

Matthew Cefalu and Francesca Dominici. Does exposure prediction bias health effect estimation? The relationship between confounding adjustment and exposure prediction. *Epidemiology*, 25(4): 583, 2014.

Spencer Chainey, Lisa Tompson, and Sebastian Uhlig. The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal*, 21(1):4–28, 2008.

Stephanie H Chan, Victor C Van Hee, Silas Bergen, Adam A Szpiro, Lisa A DeRoo, Stephanie J London, Julian D Marshall, Joel D Kaufman, and Dale P Sandler. Long-term air pollution exposure and blood pressure in the sister study. *Environmental Health Perspectives*, 123(10): 951–958, 2015.

Hung Chen. Convergence rates for parametric components in a partly linear model. *The Annals of Statistics*, pages 136–146, 1988.

Xi Chen, Qihang Lin, Seyoung Kim, Jaime G. Carbonell, and Eric P. Xing. Smoothing proximal gradient method for general structured sparse regression. *The Annals of Applied Statistics*, 6(2): 719 – 752, 2012. doi: 10.1214/11-AOAS514.

Zhao-Yue Chen, Rong Zhang, Tian-Hao Zhang, Chun-Quan Ou, and Yuming Guo. A kriging-calibrated machine learning method for estimating daily ground-level NO2 in mainland China. *Science of the Total Environment*, 690:556–564, 2019.

Sung Nok Chiu, Dietrich Stoyan, Wilfrid S Kendall, and Joseph Mecke. *Stochastic Geometry and Its Applications*. John Wiley & Sons, 2013.

David Choi. Estimation of monotone treatment effects in network experiments. *Journal of the American Statistical Association*, 112(519):1147–1155, 2017.

Fan RK Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

Bénédicte Colnet, Julie Josse, Gaël Varoquaux, and Erwan Scornet. Risk ratio, odds ratio, risk difference... Which causal measure is easier to generalize? *arXiv preprint arXiv:2303.16008*, 2023.

David Roxbee Cox. Some statistical models related with series of events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 17:129–164, 1955.

Matthew J Cracknell and Anya M Reading. Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. *Computers & Geosciences*, 63:22–33, 2014.

Noel Cressie. *Statistics for Spatial Data*. John Wiley & Sons, 2015.

Lingzhen Dai, Antonella Zanobetti, Petros Koutrakis, and Joel D Schwartz. Associations of fine particulate matter species with mortality in the United States: a multicity time-series analysis. *Environmental Health Perspectives*, 122(8):837–842, 2014.

Abhirup Datta, Sudipto Banerjee, Andrew O Finley, and Alan E Gelfand. Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812, 2016.

David Dereudre and Frédéric Lavancier. Consistency of likelihood estimation for gibbs point processes. *The Annals of Statistics*, 45(2):744–770, 2017.

Qian Di, Itai Kloog, Petros Koutrakis, Alexei Lyapustin, Yujie Wang, and Joel Schwartz. Assessing PM2.5 exposures with high spatiotemporal resolution across the continental United States. *Environmental Science & Technology*, 50(9):4712–4721, 2016.

Daniela Dias and Oxana Tchepel. Spatial and temporal dynamics in air pollution exposure assessment. *International Journal of Environmental Research and Public Health*, 15(3):558, 2018.

Ivan Diaz, Alan Hubbard, Anna Decker, and Mitchell Cohen. Variable importance and prediction methods for longitudinal problems with missing variables. *PloS one*, 10(3):e0120031, 2015.

Peter J. Diggle. *Statistical Analysis of Spatial Point Patterns*. Edward Arnold, 2003. ISBN 0340740701. 2nd edition.

Peter J Diggle, Yongtao Guan, Anthony C Hart, Fauzia Paize, and Michelle Stanton. Estimating individual-level risk in spatial epidemiology using spatially aggregated information on the population at risk. *Journal of the American Statistical Association*, 105(492):1394–1402, 2010.

Peter J Diggle, Paula Moraga, Barry Rowlingson, and Benjamin M Taylor. Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. *Statistical Science*, 28(4):542–563, 2013.

Francesca Dominici, Lianne Sheppard, and Merlise Clyde. Health effects of air pollution: a statistical review. *International Statistical Review*, 71(2):243–276, 2003.

Kangning Dong and Shihua Zhang. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature Communications*, 13(1):1739, 2022.

Peijun Du, Xuyu Bai, Kun Tan, Zhaohui Xue, Alim Samat, Junshi Xia, Erzhu Li, Hongjun Su, and Wei Liu. Advances of four machine learning methods for spatial data handling: A review. *Journal of Geovisualization and Spatial Analysis*, 4(1):1–25, 2020.

Jiří Dvořák, Jesper Møller, Tomáš Mrkvička, and Samuel Soubeyrand. Quick inference for log Gaussian Cox processes with non-stationary underlying random fields. *Spatial Statistics*, 33: 100388, 2019.

Bradley Efron. The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99(467):619–632, 2004.

P Erdős and A Rényi. On random graphs. *Publicationes Mathematicae*, 6(290-297):18, 1959.

Theodoros Evgeniou, Massimiliano Pontil, and Tomaso Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13:1–50, 2000.

Yin Feng, Jinhua Cheng, Jun Shen, and Han Sun. Spatial effects of air pollution on public health in China. *Environmental and Resource Economics*, 73:229–250, 2019.

Jorge Ferreira, Paulo João, and José Martins. GIS for crime analysis: Geography for predictive models. *Electronic Journal of Information Systems Evaluation*, 15(1):pp36–49, 2012.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

Ivan Franch-Pardo, Brian M Napoletano, Fernando Rosete-Verges, and Lawal Billa. Spatial analysis and GIS in the study of COVID-19. A review. *Science of The Total Environment*, 739:140033, 2020.

Jerome H Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67, 1991.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

Matthias Galipaud, Mark AF Gillingham, and François-Xavier Dechaume-Moncharmont. A farewell to the sum of Akaike weights: The benefits of alternative metrics for variable importance estimations in model selection. *Methods in Ecology and Evolution*, 8(12):1668–1678, 2017.

Alan E Gelfand and Dipak K Dey. Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(3):501–514, 1994.

Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot. Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225–2236, 2010.

Stefanos Georganos, Tais Grippa, Assane Niang Gadiaga, Catherine Linard, Moritz Lennert, Sabine Vanhuysse, Nicholus Mboga, Eléonore Wolff, and Stamatis Kalogirou. Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto International*, 36(2):121–136, 2021.

Mi-Ho Giga, Yoshikazu Giga, Jürgen Saal, Mi-Ho Giga, Yoshikazu Giga, and Jürgen Saal. *Nonlinear Partial Differential Equations: Asymptotic Behavior of Solutions and Self-Similar Solutions*, chapter Convergence Theorems in the Theory of Integration, pages 239–247. Springer, 2010.

Mark S Goldberg. On the interpretation of epidemiological studies of ambient air pollution. *Journal of Exposure Science & Environmental Epidemiology*, 17(2):S66–S70, 2007.

Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics*, 24(1):44–65, 2015.

Romina Gonella, Mathias Bourel, and Liliane Bel. Facing spatial massive data in science and society: Variable selection for spatial models. *Spatial Statistics*, page 100627, 2022.

Michael F Goodchild and Robert P Haining. GIS and spatial data analysis: Converging perspectives. *Papers in Regional Science*, 83(1):363–385, 2004.

Jochen Gorski, Frank Pfeuffer, and Kathrin Klamroth. Biconvex sets and optimization with biconvex functions: a survey and extensions. *Mathematical Methods of Operations Research*, 66(3): 373–407, 2007.

Brandon M Greenwell, Bradley C Boehmke, and Andrew J McCarthy. A simple and effective model-based variable importance measure. *arXiv preprint arXiv:1805.04755*, 2018.

Baptiste Gregorutti, Bertrand Michel, and Philippe Saint-Pierre. Correlation and variable importance in random forests. *Statistics and Computing*, 27(3):659–678, 2017.

Yongtao Guan. A composite likelihood approach in fitting spatial point process models. *Journal of the American Statistical Association*, 101(476):1502–1512, 2006.

Yongtao Guan. On consistent nonparametric intensity estimation for inhomogeneous spatial point processes. *Journal of the American Statistical Association*, 103(483):1238–1247, 2008.

Douglas Guilbeault, Joshua Becker, and Damon Centola. Complex contagions: A decade in review. *Complex spreading phenomena in social systems: Influence and contagion in real-world social networks*, pages 3–25, 2018.

Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, 20(1):296, 2019.

M Elizabeth Halloran and Claudio J Struchiner. Causal inference in infectious diseases. *Epidemiology*, pages 142–151, 1995.

Wolfgang Härdle, Hua Liang, and Jiti Gao. *Partially linear models*. Springer Science & Business Media, 2000.

Asad Haris, Ali Shojaie, and Noah Simon. Nonparametric regression with adaptive truncation via a convex hierarchical penalty. *Biometrika*, 106(1):87–107, 2019a.

Asad Haris, Noah Simon, and Ali Shojaie. Generalized sparse additive models. *arXiv preprint arXiv:1903.04641*, 2019b.

Asad Haris, Noah Simon, and Ali Shojaie. Generalized sparse additive models. *The Journal of Machine Learning Research*, 23(1):3035–3090, 2022.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: the Lasso and Generalizations*. Chapman and Hall/CRC, 2019.

Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*, volume 43. CRC press, 1990.

Yichun He, Xin Tang, Jiahao Huang, Jingyi Ren, Haowen Zhou, Kevin Chen, Albert Liu, Hailing Shi, Zuwan Lin, Qiang Li, et al. ClusterMap for multi-scale clustering analysis of spatial gene expression. *Nature Communications*, 12(1):5909, 2021.

Tomislav Hengl, Gerard BM Heuvelink, Bas Kempen, Johan GB Leenaars, Markus G Walsh, Keith D Shepherd, Andrew Sila, Robert A MacMillan, Jorge Mendes de Jesus, Lulseged Tamene, et al. Mapping soil properties of Africa at 250m resolution: Random forests significantly improve current predictions. *PloS one*, 10(6):e0125814, 2015.

Tomislav Hengl, Madlene Nussbaum, Marvin N Wright, Gerard BM Heuvelink, and Benedikt Gräler. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, 6:e5518, 2018.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Guanglei Hong and Stephen W Raudenbush. Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, 101(475):901–910, 2006.

Giles Hooker, Lucas Mentch, and Siyu Zhou. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31(6):1–16, 2021.

Jian Hu, Xiangjie Li, Kyle Coleman, Amelia Schroeder, Nan Ma, David J Irwin, Edward B Lee, Russell T Shinohara, and Mingyao Li. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature Methods*, 18(11):1342–1351, 2021.

Shishan Hu, Suzanne E Paulson, Scott Fruin, Kathleen Kozawa, Steve Mara, and Arthur M Winer. Observation of elevated air pollutant concentrations in a residential neighborhood of los angeles california using a mobile platform. *Atmospheric Environment*, 51:311–319, 2012.

Yuchen Hu, Shuangning Li, and Stefan Wager. Average direct and indirect causal effects under interference. *Biometrika*, 109(4):1165–1172, 2022.

Zhilian Huang, Huiling Guo, Yee-Mun Lee, Eu Chin Ho, Hou Ang, Angela Chow, et al. Performance of digital contact tracing tools for COVID-19 response in Singapore: cross-sectional study. *JMIR mHealth and uHealth*, 8(10):e23148, 2020.

N Hudda, MC Simon, W Zamore, and JL Durant. Aviation-related impacts on ultrafine particle number concentrations outside and inside residences near an airport. *Environmental Science & Technology*, 52(4):1765–1772, 2018.

Michael G Hudgens and M Elizabeth Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.

Janine Illian, Antti Penttinen, Helga Stoyan, and Dietrich Stoyan. *Statistical Analysis and Modelling of Spatial Point Patterns*, volume 70. John Wiley & Sons, 2008.

Matthew O Jackson et al. *Social and economic networks*, volume 3. Princeton university press Princeton, 2008.

Roman A Jandarov, Lianne A Sheppard, Paul D Sampson, and Adam A Szpiro. A novel principal component analysis for spatially misaligned multivariate air pollution data. *Journal of the Royal Statistical Society. Series C, Applied statistics*, 66(1):3, 2017.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

Jens Ledet Jensen. Asymptotic normality of estimates in spatial point processes. *Scandinavian Journal of Statistics*, pages 97–109, 1993.

Ian T Jolliffe. *Principal component analysis for special types of data*. Springer, 2002.

David Kahle and Hadley Wickham. ggmap: Spatial Visualization with ggplot2. *The R Journal*, 5(1): 144–161, 2013. URL `https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf`.

Mikhail Kanevski. *Machine learning for spatial environmental data: theory, applications, and software*. EPFL press, 2009.

Sadra Karimzadeh, Masashi Matsuoka, Jianming Kuang, and Linlin Ge. Spatial prediction of aftershocks triggered by a major earthquake: A binary machine learning perspective. *ISPRS International Journal of Geo-Information*, 8(10):462, 2019.

Joel D Kaufman, Sara D Adar, R Graham Barr, Matthew Budoff, Gregory L Burke, Cynthia L Curl, Martha L Daviglus, Ana V Diez Roux, Amanda J Gassett, David R Jacobs Jr, et al. Association between air pollution and coronary artery calcification within six metropolitan areas in the usa (the Multi-Ethnic Study of Atherosclerosis and Air pollution): a longitudinal cohort study. *The Lancet*, 388(10045):696–704, 2016.

Joshua P Keller and Adam A Szpiro. Selecting a scale for spatial confounding adjustment. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 183(3):1121–1143, 2020.

Joshua P Keller, Casey Olives, Sun-Young Kim, Lianne Sheppard, Paul D Sampson, Adam A Szpiro, Assaf P Oron, Johan Lindström, Sverre Vedal, and Joel D Kaufman. A unified spatiotemporal modeling approach for predicting concentrations of multiple air pollutants in the multi-ethnic study of atherosclerosis and air pollution. *Environmental Health Perspectives*, 123(4):301–309, 2015.

Joshua P Keller, Mathias Drton, Timothy Larson, Joel D Kaufman, Dale P Sandler, and Adam A Szpiro. Covariate-adaptive clustering of exposures for air pollution epidemiology cohorts. *The Annals of Applied Statistics*, 11(1):93, 2017.

BM Golam Kibria, Li Sun, James V Zidek, and Nhu D Le. Bayesian spatial prediction of random space-time fields with application to mapping PM2.5 exposure. *Journal of the American Statistical Association*, 97(457):112–124, 2002.

David A Kim, Alison R Hwong, Derek Stafford, D Alex Hughes, A James O'Malley, James H

Fowler, and Nicholas A Christakis. Social network targeting to maximise population behaviour change: a cluster randomised controlled trial. *The Lancet*, 386(9989):145–153, 2015.

Sun-Young Kim, Lianne Sheppard, and Ho Kim. Health effects of long-term air pollution: influence of exposure prediction methods. *Epidemiology*, pages 442–450, 2009.

Kipruto Kirwa, Adam A Szpiro, Lianne Sheppard, Paul D Sampson, Meng Wang, Joshua P Keller, Michael T Young, Sun-Young Kim, Timothy V Larson, and Joel D Kaufman. Fine-scale air pollution models for epidemiologic research: insights from approaches developed in the multi-ethnic study of atherosclerosis and air pollution (MESA Air). *Current Environmental Health Reports*, 8(2):113–126, 2021.

Vladimir Yu Kiselev, Kristina Kirschner, Michael T Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N Natarajan, Wolf Reik, Mauricio Barahona, and Anthony R Green. SC3: consensus clustering of single-cell RNA-seq data. *Nature Methods*, 14(5):483–486, 2017.

Miloš Kovacevic, Branislav Bajat, Branislav Trivic, and Radmila Pavlovic. Geological units classification of multispectral images by using support vector machines. In *2009 International Conference on Intelligent Networking and Collaborative Systems*, pages 267–272. IEEE, 2009.

Maria Krysan. Does race matter in the search for housing? An exploratory study of search strategies, experiences, and locations. *Social Science Research*, 37(2):581–603, 2008.

Frédéric Lavancier, Jesper Møller, and Ege Rubak. Determinantal point process models and statistical inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77 (4):853–877, 2015.

Richard Law, Janine Illian, David FRP Burslem, Georg Gratzer, CVS Gunatilleke, and IAUN Gunatilleke. Ecological information from spatial patterns of plants: insights from point process theory. *Journal of Ecology*, 97(4):616–628, 2009.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

Kelvin Leong and Anna Sung. A review of spatio-temporal pattern analysis approaches on crime analysis. *International E-journal of Criminal Sciences*, 9:1–33, 2015.

Michael P Leung. Treatment and spillover effects under network interference. *Review of Economics and Statistics*, 102(2):368–380, 2020.

Michael P Leung. Causal inference under approximate neighborhood interference. *Econometrica*, 90(1):267–293, 2022.

Jin Li, Andrew D Heap, Anna Potter, and James J Daniell. Application of machine learning methods to spatial interpolation of environmental variables. *Environmental Modelling & Software*, 26(12): 1647–1659, 2011.

Ker-Chau Li. From Stein's unbiased risk estimates to the method of generalized cross validation. *The Annals of Statistics*, pages 1352–1377, 1985.

Shuangning Li and Stefan Wager. Random graph asymptotics for treatment effect estimation under network interference. *The Annals of Statistics*, 50(4):2334–2358, 2022.

Tianxi Li, Elizaveta Levina, and Ji Zhu. Prediction models for network-linked data. *The Annals of Applied Statistics*, 13(1):132–164, 2019.

Ye Li, Patrick Brown, Dionne C Gesink, and Håvard Rue. Log Gaussian Cox processes and spatially aggregated disease incidence data. *Statistical Methods in Medical Research*, 21(5):479–507, 2012.

Zhuliu Li, Tianci Song, Jeongsik Yong, and Rui Kuang. Imputation of spatially-resolved transcriptomes by graph-regularized tensor completion. *PLoS Computational Biology*, 17(4):e1008218, 2021.

Bruce G Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 80(1):221–239, 1988.

Morton Lippmann, Lung Chi Chen, Terry Gordon, Kazuhiko Ito, and George D Thurston. National Particle Component Toxicity (NPACT) Initiative: integrated epidemiologic and toxicologic studies of the health effects of particulate matter components. *Research Report (Health Effects Institute)*, 177:5–13, 2013.

Wei Liu, Xu Liao, Yi Yang, Huazhen Lin, Joe Yeong, Xiang Zhou, Xingjie Shi, and Jin Liu. Joint dimension reduction and clustering analysis of single-cell RNA-seq and spatial transcriptomics data. *Nucleic Acids Research*, 50(12):e72–e72, 2022.

Ying Liu, Guofeng Cao, Naizhuo Zhao, Kevin Mulligan, and Xinyue Ye. Improve ground-level PM2.5 concentration mapping using a random forests-based geostatistical approach. *Environmental Pollution*, 235:272–282, 2018.

Evans K Lodge, Cathrine Hoyo, Carmen M Gutierrez, Kristen M Rappazzo, Michael E Emch, and Chantel L Martin. Estimating exposure to neighborhood crime by race and ethnicity for public health research. *BMC Public Health*, 21(1):1–13, 2021.

Yahui Long, Kok Siong Ang, Mengwei Li, Kian Long Kelvin Chong, Raman Sethi, Chengwei Zhong, Hang Xu, Zhiwei Ong, Karishma Sachaphibulkij, Ao Chen, et al. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. *Nature Communications*, 14(1):1155, 2023.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

Odile Macchi. The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7(1):83–122, 1975.

Charles F Manski. Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23, 2013.

Sahar Masmoudi, Haytham Elghazel, Dalila Taieb, Orhan Yazar, and Amjad Kallel. A machine-learning framework for predicting multiple air pollutants' concentrations via multi-target regression and feature selection. *Science of the Total Environment*, 715:136991, 2020.

Colin McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, 141(1): 148–188, 1989.

Lucas Mentch and Giles Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research*, 17(1):841–881, 2016.

MESA Air. Data Organization and Operating Procedures for the Multi-Ethnic Study of Atherosclerosis and Air Pollution and Associated Studies. `https://kaufman-lab.github.io/doop/`, 2019a.

MESA Air. Data organization and operating procedures for the multi-ethnic study of atherosclerosis and air pollution and associated studies. *MESA Air*, 2019b. URL `https://kaufman-lab.github.io/doop/`.

Hanna Meyer, Christoph Reudenbach, Stephan Wöllauer, and Thomas Nauss. Importance of spatial predictor variable selection in machine learning applications – moving from data reproduction to spatial prediction. *Ecological Modelling*, 411:108815, 2019.

Edward Miguel and Michael Kremer. Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, 72(1):159–217, 2004.

Jesper Møller and Rasmus P Waagepetersen. Modern statistics for spatial point processes. *Scandinavian Journal of Statistics*, 34(4):643–684, 2007.

Jesper Møller and Rasmus P Waagepetersen. Some recent developments in statistics for spatial point patterns. *Annual Review of Statistics and Its Application*, 4:317–342, 2017.

Jesper Møller and Rasmus Plenge Waagepetersen. *Statistical Inference and Simulation for Spatial Point Processes*. CRC Press, 2003.

Jesper Møller, Anne Randi Syversveen, and Rasmus Plenge Waagepetersen. Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482, 1998.

Christian Monn. Exposure assessment of air pollutants: a review on spatial heterogeneity and indoor/outdoor/personal exposure to suspended particulate matter, nitrogen dioxide and ozone. *Atmospheric Environment*, 35(1):1–32, 2001.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Jersey Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10(1):1–51, 1923.

Kristin K Nicodemus, James D Malley, Carolin Strobl, and Andreas Ziegler. The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, 11(1):1–13, 2010.

Yang Ning and Han Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Annals of Statistics*, 45(1):158–195, 2017.

Yoichi Nishiyama. Local asymptotic normality of a sequential model for marked point processes and its applications. *Annals of the Institute of Statistical Mathematics*, 47(2):195–209, 1995.

Halûk Özkaynak, Lisa K Baxter, Kathie L Dionisio, and Janet Burke. Air pollution exposure prediction approaches used in air pollution epidemiology studies. *Journal of Exposure Science & Environmental Epidemiology*, 23(6):566–572, 2013.

Christopher J Paciorek. The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Statistical Science*, 25(1):107, 2010.

Oscar Hernan Madrid Padilla, James Sharpnack, James G Scott, and Ryan J Tibshirani. The DFS fused lasso: Linear-time denoising over general graphs. *Journal of Machine Learning Research*, 18:176–1, 2017.

Elizabeth Levy Paluck, Hana Shepherd, and Peter M Aronow. Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences*, 113 (3):566–571, 2016.

Seong Suk Park, Kathleen Kozawa, Scott Fruin, Steve Mara, Ying-Kuang Hsu, Chris Jakober, Arthur Winer, and Jorn Herner. Emission factors for high-emitting vehicles based on on-road measurements of individual vehicle exhaust with a mobile measurement platform. *Journal of the Air & Waste Management Association*, 61(10):1046–1056, 2011.

Robert N Parker, Nicholas J Rosser, and Tristram C Hales. Spatial prediction of earthquake-

induced landslide probability. *Natural Hazards and Earth System Sciences Discussions*, pages 1–29, 2017.

Carolina Perez-Heydrich, Michael G Hudgens, M Elizabeth Halloran, John D Clemens, Mohammad Ali, and Michael E Emch. Assessing effects of cholera vaccination in the presence of interference. *Biometrics*, 70(3):731–741, 2014.

Emma Pierson and Christopher Yau. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*, 16(1):1–10, 2015.

Joshua B Plotkin, Matthew D Potts, Nandi Leslie, N Manokaran, James LaFrankie, and Peter S Ashton. Species-area curves, spatial aggregation, and habitat specialization in tropical forests. *Journal of Theoretical Biology*, 207(1):81–99, 2000.

Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. In *Computer and Information Sciences - ISCIS 2005: 20th International Symposium, Istanbul, Turkey, October 26-28, 2005. Proceedings 20*, pages 284–293. Springer, 2005.

Sandhya Prabhakaran, Elham Azizi, Ambrose Carr, and Dana Pe'er. Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. In *International Conference on Machine Learning*, pages 1070–1079. PMLR, 2016.

Assaf Rabinowicz and Saharon Rosset. Cross-validation for correlated data. *Journal of the American Statistical Association*, 117(538):718–731, 2022.

Anjali Rao, Dalia Barkley, Gustavo S França, and Itai Yanai. Exploring tissue architecture using spatial transcriptomics. *Nature*, 596(7871):211–220, 2021.

Stephen L Rathbun. Asymptotic properties of the maximum likelihood estimator for spatio-temporal point processes. *Journal of Statistical Planning and Inference*, 51(1):55–74, 1996.

Stephen L Rathbun and Noel Cressie. Asymptotic properties of estimators for the parameters of spatial inhomogeneous Poisson point processes. *Advances in Applied Probability*, 26(1):122–154, 1994.

Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(5):1009–1030, 2009.

Brian J Reich, James S Hodges, and Vesna Zadnik. Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, 62(4):1197–1206, 2006.

Ian W Renner, Jane Elith, Adrian Baddeley, William Fithian, Trevor Hastie, Steven J Phillips, Gordana Popovic, and David I Warton. Point process models for presence-only analysis. *Methods in Ecology and Evolution*, 6(4):366–379, 2015.

Andrea Riebler, Sigrunn H Sørbye, Daniel Simpson, and Håvard Rue. An intuitive bayesian spatial model for disease mapping that accounts for scaling. *Statistical Methods in Medical Research*, 25 (4):1145–1165, 2016.

Donald Rubin. Comment on "randomization analysis of experimental data: The Fisher randomization test" by B. BasuV. *Journal of the American Statistical Association*, 91:267, 1980.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Donald B Rubin. Formal mode of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, 25(3):279–292, 1990.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.

Georg Ruß and Alexander Brenning. Spatial variable importance assessment for yield prediction in precision agriculture. In *International Symposium on Intelligent Data Analysis*, pages 184–195. Springer, 2010.

J Sacks, B Buckley, N Alexis, M Angrish, R Beardslee, A Benson, J Brown, B Buckley, M Campen, E Chan, et al. Integrated science assessment ISA for particulate matter (final report, December 2009). *Environmental Protection Agency*, 2009. URL `https://cfpub.epa.gov/ncea/isa/recordisplay.cfm`.

Veeru Sadhanala, Yu-Xiang Wang, and Ryan Tibshirani. Graph sparsification approaches for Laplacian smoothing. In *Artificial Intelligence and Statistics*, pages 1250–1259. PMLR, 2016.

Arkajyoti Saha, Sumanta Basu, and Abhirup Datta. Random forests for spatially dependent data. *Journal of the American Statistical Association*, pages 1–19, 2021.

Paul D Sampson, Adam A Szpiro, Lianne Sheppard, Johan Lindström, and Joel D Kaufman. Pragmatic estimation of a spatio-temporal air quality model with irregular monitoring data. *Atmospheric Environment*, 45(36):6593–6606, 2011.

Fredrik Sävje, Peter Aronow, and Michael Hudgens. Average treatment effects in the presence of unknown interference. *Annals of Statistics*, 49(2):673, 2021.

Frederic Paik Schoenberg. Consistent parametric estimation of the intensity of a spatial–temporal point process. *Journal of Statistical Planning and Inference*, 128(1):79–93, 2005.

Richard F Serfozo. Conditional poisson processes. *Journal of Applied Probability*, 9(2):288–302, 1972.

Lulu Shang and Xiang Zhou. Spatially aware dimension reduction for spatial transcriptomics. *Nature Communications*, 13(1):7203, 2022.

Shavneet Sharma, Gurmeet Singh, Rashmini Sharma, Paul Jones, Sascha Kraus, and Yogesh K Dwivedi. Digital health innovation: exploring adoption of COVID-19 digital contact tracing apps. *IEEE Transactions on Engineering Management*, 2020.

Haipeng Shen and Jianhua Z Huang. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99(6):1015–1034, 2008.

Daniel Simpson, Janine Baerbel Illian, Finn Lindgren, Sigrunn H Sørbye, and Håvard Rue. Going off grid: Computationally efficient inference for log-Gaussian Cox processes. *Biometrika*, 103(1): 49–70, 2016.

Daniel Simpson, Håvard Rue, Andrea Riebler, Thiago G Martins, and Sigrunn H Sørbye. Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, 32(1):1–28, 2017.

Michael E Sobel. What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. *Journal of the American Statistical Association*, 101(476):1398–1407, 2006.

Patrik L Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353 (6294):78–82, 2016.

Michael L Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Science & Business Media, 1999.

Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1): 1–21, 2007.

Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9:1–11, 2008.

Shiquan Sun, Jiaqiang Zhu, Ying Ma, and Xiang Zhou. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biology*, 20(1):1–21, 2019.

Adam A Szpiro and Christopher J Paciorek. Measurement error in two-stage analyses, with application to air pollution epidemiology. *Environmetrics*, 24(8):501–517, 2013.

Adam A Szpiro, Christopher J Paciorek, and Lianne Sheppard. Does more accurate exposure prediction necessarily improve health effect estimates? *Epidemiology*, 22(5):680, 2011.

Benjamin M Taylor, Ricardo Andrade-Pacheco, and Hugh JW Sturrock. Continuous inference for aggregated point process data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(4):1125–1150, 2018.

Ming Teng, Farouk Nathoo, and Timothy D Johnson. Bayesian computation for log-Gaussian

Cox processes: A comparative analysis of methods. *Journal of Statistical Computation and Simulation*, 87(11):2227–2252, 2017.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

Ryan J Tibshirani and Jonathan Taylor. The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335–1371, 2011.

Edwina S Uehara. Race, gender, and housing inequality: An exploration of the correlates of low-quality housing among clients diagnosed with severe and persistent mental illness. *Journal of Health and Social Behavior*, pages 309–321, 1994.

Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

Sara A van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Sara A Van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

Mark J Van der Laan. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1), 2006.

Mark J van der Laan. Causal inference for a population of causally connected units. *Journal of Causal Inference*, 2(1):13–74, 2014.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Tyler J VanderWeele and Eric J Tchetgen Tchetgen. Bounding the infectiousness effect in vaccine trials. *Epidemiology*, 22(5):686, 2011.

Stijn Vansteelandt and Oliver Dukes. Assumption-lean inference for generalised linear model parameters. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):657–685, 2022.

Sverre Vedal, Matthew J Campen, Jacob D McDonald, Timothy V Larson, Paul D Sampson, Lianne Sheppard, Christopher D Simpson, and Adam A Szpiro. National Particle Component Toxicity (NPACT) Initiative report on cardiovascular effects. *Research Report (Health Effects Institute)*, 178:5–8, 2013.

Fabien Viger and Matthieu Latapy. Efficient and simple generation of random simple connected graphs with prescribed degree sequence. In *International Computing and Combinatorics Conference*, pages 440–449. Springer, 2005.

Fabrice Vinatier, Philippe Tixier, Pierre-François Duyck, and Francoise Lescourret. Factors and mechanisms explaining spatial heterogeneity: a review of methods for insect populations. *Methods in Ecology and Evolution*, 2(1):11–22, 2011.

Arend Voorman, Ali Shojaie, and Daniela Witten. Inference in high dimensions with the penalized score test. *arXiv preprint arXiv:1401.2678*, 2014.

Phuong T Vu, Timothy V Larson, and Adam A Szpiro. Probabilistic predictive principal component analysis for spatially misaligned and high-dimensional air pollution data with missing observations. *Environmetrics*, 31(4):e2614, 2020.

Phuong T Vu, Adam A Szpiro, and Noah Simon. Spatial matrix completion for spatially misaligned and high-dimensional air pollution data. *Environmetrics*, 33(4):e2713, 2022.

Rasmus Waagepetersen and Yongtao Guan. Two-step estimation for inhomogeneous spatial point processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3): 685–702, 2009.

Rasmus P Waagepetersen. Convergence of posteriors for discretized log Gaussian Cox processes. *Statistics & Probability Letters*, 66(3):229–235, 2004.

Rasmus Plenge Waagepetersen. An estimating function approach to inference for inhomogeneous Neyman–Scott processes. *Biometrics*, 63(1):252–258, 2007.

Grace Wahba. *Spline models for observational data*. SIAM, 1990.

Travis Hee Wai, Michael T Young, and Adam A Szpiro. Random spatial forests. *arXiv preprint arXiv:2006.00150*, 2020.

Xing Wang, Dane Westerdahl, Ye Wu, Xiaochuan Pan, and K Max Zhang. On-road emission factor distributions of individual diesel vehicles in and around Beijing, China. *Atmospheric Environment*, 45(2):503–513, 2011.

Yixin Wang and David M Blei. Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, 114(527):1147–1161, 2019.

Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

Pengfei Wei, Zhenzhou Lu, and Jingwen Song. Variable importance analysis: a comprehensive review. *Reliability Engineering & System Safety*, 142:399–432, 2015.

David Whitney, Ali Shojaie, and Marco Carone. Comment: Models as (deliberate) approximations. *Statistical Science*, 34(4):591, 2019.

Brian D Williamson, Peter B Gilbert, Marco Carone, and Noah Simon. Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, 77(1):9–22, 2021.

Svante Wold, Arnold Ruhe, Herman Wold, and WJ Dunn, Iii. The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3):735–743, 1984.

Simon N Wood. Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114, 2003.

Junshi Xu, Mingqian Zhang, Arman Ganji, Keni Mallinen, An Wang, Marshall Lloyd, Alessya Venuta, Leora Simon, Junwon Kang, James Gong, et al. Prediction of short-term ultrafine

particle exposures using real-time street-level images paired with air quality measurements. *Environmental Science & Technology*, 56(18):12886–12897, 2022.

Jeff D Yanosky, Christopher J Paciorek, and Helen H Suh. Predicting chronic fine and coarse particulate exposures using spatiotemporal models for the Northeastern and Midwestern United States, 2009.

Jianming Ye. On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93(441):120–131, 1998.

Fei Yi, Zhiwen Yu, Fuzhen Zhuang, Xiao Zhang, and Hui Xiong. An integrated model for crime prediction using temporal and spatial factors. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 1386–1391. IEEE, 2018.

Kenneth Yip and Feng Zhao. Spatial aggregation: theory and applications. *Journal of Artificial Intelligence Research*, 5:1–26, 1996.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, 2006.

Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. *Social media mining: an introduction*. Cambridge University Press, 2014.

Hao Zhang and Dale L Zimmerman. Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika*, 92(4):921–936, 2005.

Xiaole Zhang, Xi Chen, and Jing Wang. A number-based inventory of size-resolved black carbon particle emissions by global civil aviation. *Nature Communications*, 10(1):534, 2019.

Edward Zhao, Matthew R Stone, Xing Ren, Jamie Guenthoer, Kimberly S Smythe, Thomas Pulliam, Stephen R Williams, Cedric R Uytingco, Sarah EB Taylor, Paul Nghiem, et al. Spatial transcriptomics at subspot resolution with BayesSpace. *Nature Biotechnology*, 39(11):1375–1384, 2021a.

Sen Zhao and Ali Shojaie. A significance test for graph-constrained estimation. *Biometrics*, 72(2): 484–493, 2016a.

Sen Zhao and Ali Shojaie. A significance test for graph-constrained estimation. *Biometrics*, 72(2): 484–493, 2016b.

Sen Zhao, Daniela Witten, and Ali Shojaie. In defense of the indefensible: A very naive approach to high-dimensional inference. *Statistical Science*, 36(4):562–577, 2021b.

Xiangyu Zhao and Jiliang Tang. Modeling temporal-spatial correlations for crime prediction. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 497–506, 2017.

Bin Zhou, Xiangyi Meng, and H Eugene Stanley. Power-law distribution of degree–degree distance: A better representation of the scale-free property of complex networks. *Proceedings of the National Academy of Sciences*, 117(26):14812–14818, 2020.

Appendix A

SUPPLEMENTARY MATERIALS FOR CHAPTER 2

A.1 Annual Average Pollutant Concentration and Prediction Results

A.1.1 Data Description

This section presents the distributions of pollutants that are not discussed in detail in the main text. Figure A.1 visualizes the estimated annual average concentration of BC, NO_2, CO_2 and $PM_{2.5}$ in the Seattle TRAP study, and Figure A.2 presents the annual average concentration of OC, S and Si in the national data.

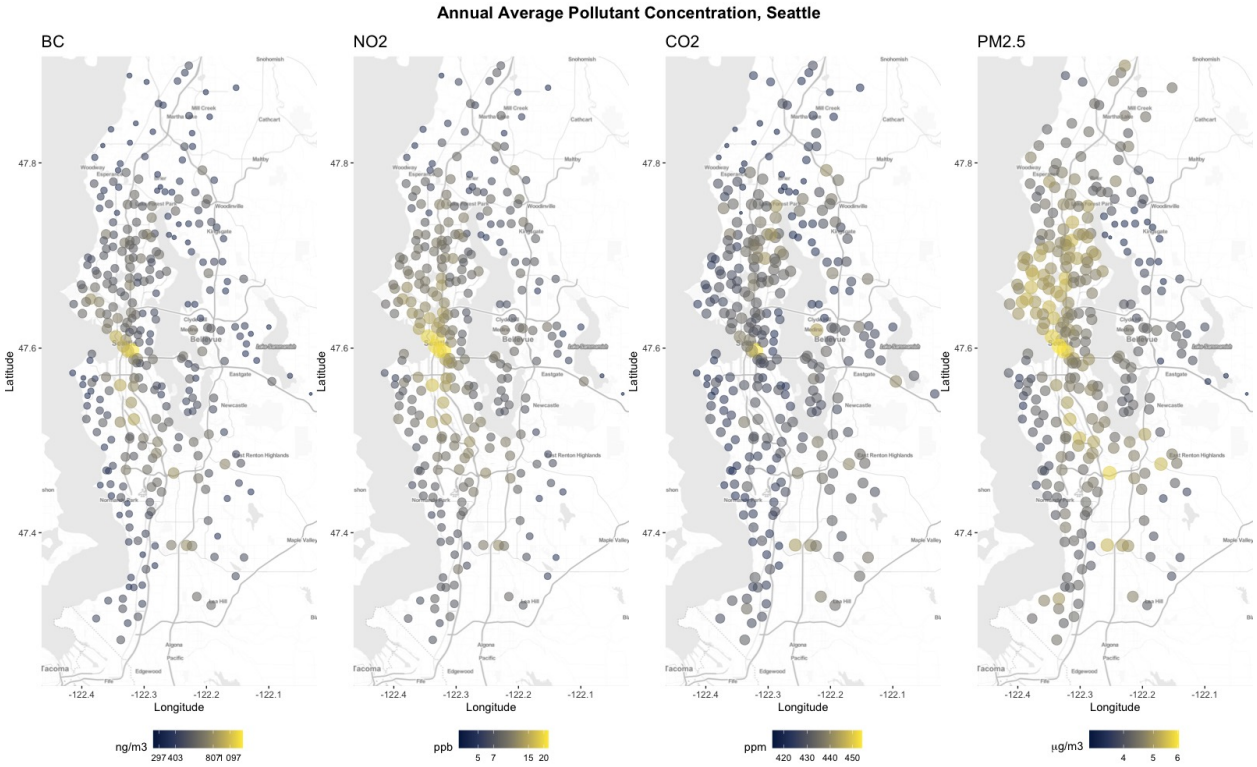


Figure A.1: Annual average concentration of BC, NO_2, CO_2 and $PM_{2.5}$ at mobile monitoring locations in the Seattle dataset

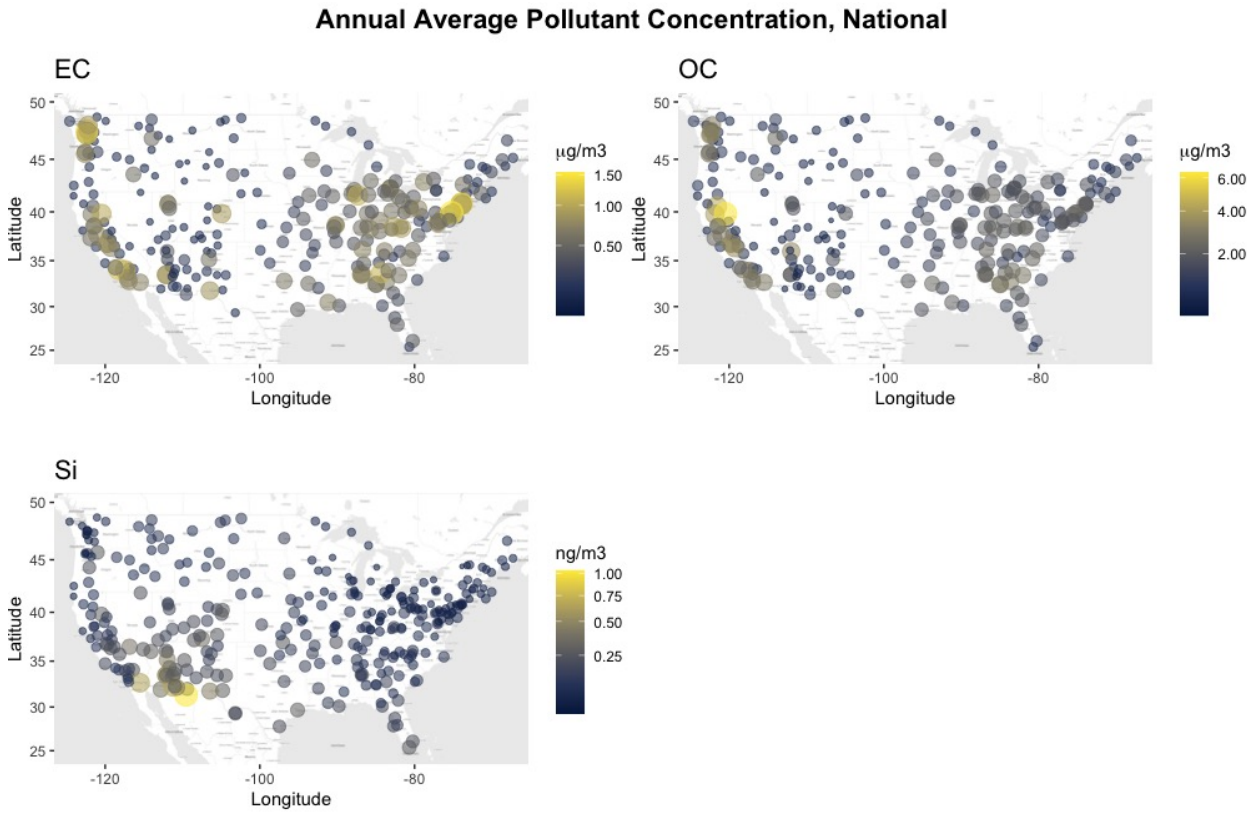


Figure A.2: Annual average concentration of EC, OC and Si at monitoring locations in the national dataset

A.1.2 Synthetic Data Description

Figure A.3 visualizes the overall distribution and decomposition of the synthetic data, i.e. variability coming from the mean, partial sill and nugget, respectively.

A.1.3 Prediction Models

In addition to our primary models, UK-PLS and spatial RF-PL, we also investigated the performance of spatial RF with nonparametric optimization approach (SpatRF-NP, see Wai et al., 2020) along with four benchmark models as a comparison:

- RF: random forest implemented by the `randomForest` R package ignoring spatial correlation;

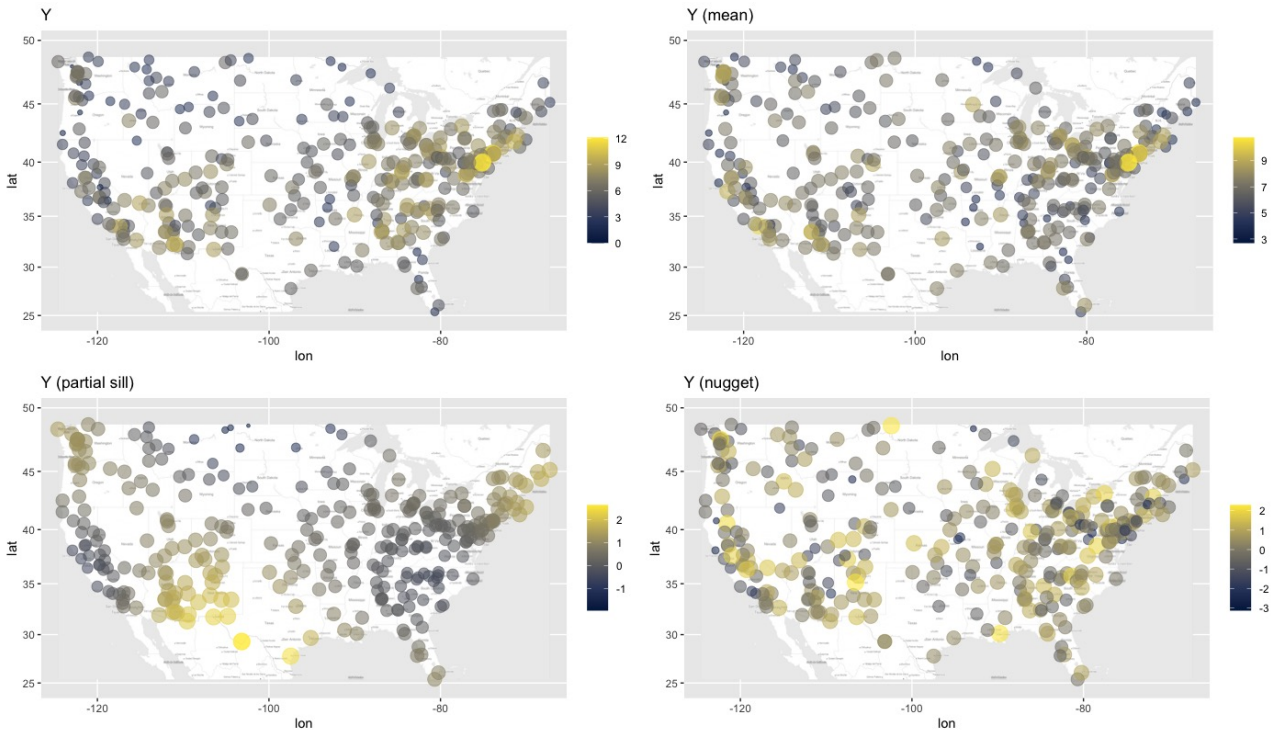


Figure A.3: Decomposition of the synthetic outcome

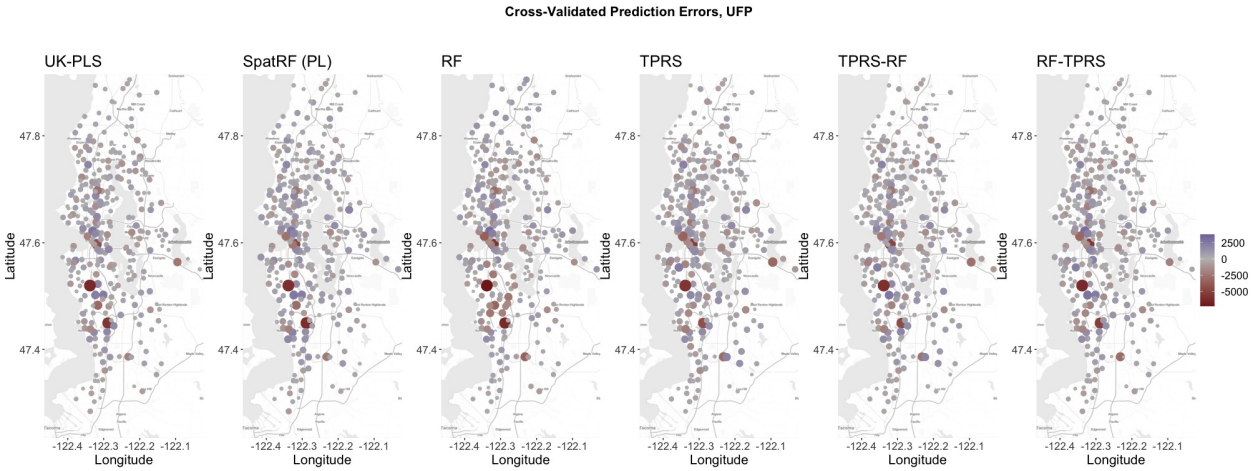
- TPRS: spatial smoothing via thin plate regression splines implemented by the `mgcv` R package;
- RF-TPRS: a two-step procedure that first runs RF, and then conducts TPRS spatial smoothing on the residuals from RF;
- TPRS-RF: a two-step procedure that first runs TPRS, and then applies RF on the residuals from TPRS.

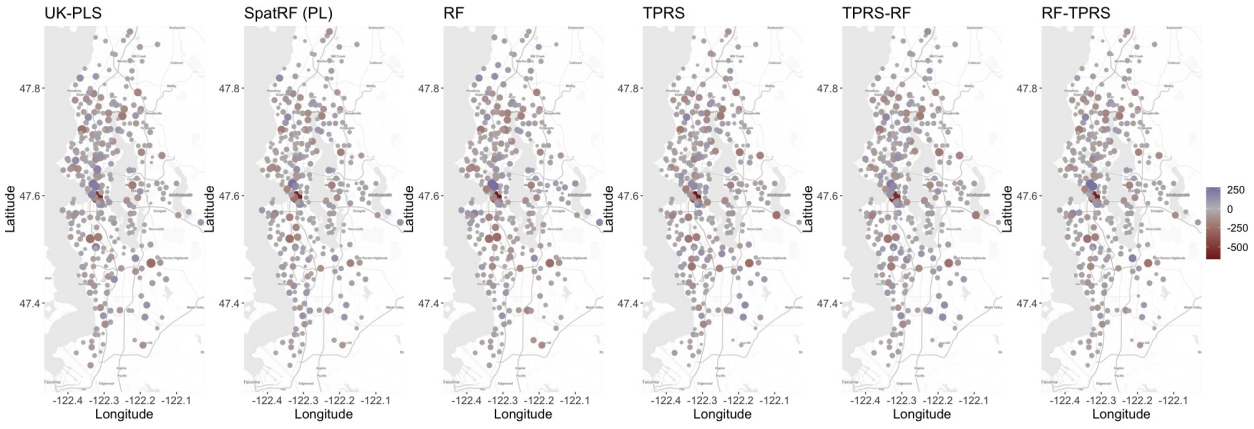
	UK-PLS	RF	TPRS	RF-TPRS	TPRS-RF	SpatRF (PL)	SpatRF (NP)
UFP	0.81	0.75	0.76	0.79	0.80	0.78	0.78
BC	0.65	0.60	0.57	0.67	0.64	0.67	0.67
NO_2	0.77	0.70	0.70	0.76	0.74	0.75	0.74
CO_2	0.56	0.47	0.44	0.57	0.55	0.56	0.54
$PM_{2.5}$	0.76	0.66	0.73	0.72	0.73	0.74	0.71

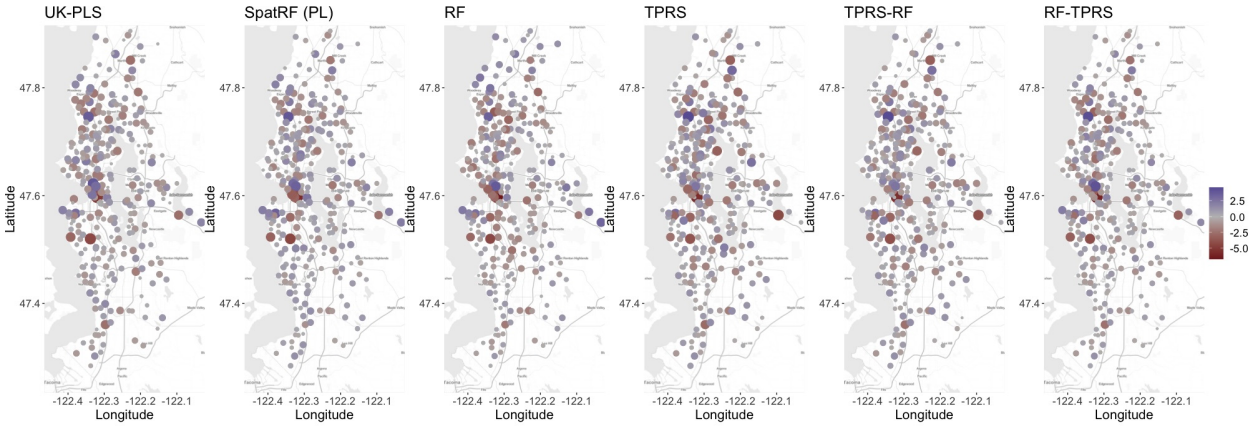
Table A.1: Cross-validated R^2 for each method on the Seattle TRAP data

Table A.1 summarizes the cross-validated R^2 for all models and all pollutants on the Seattle data. The performance of different models relative to each other reveals different sources of heterogeneity in pollutant concentration: for UFP and NO_2, purely covariate and spatial effects both account for part of the spatial heterogeneity, reflected by reasonable performance of RF or TPRS alone; accounting for both of them together either in a joint or two-step manner further leads to increased accuracy. The case is similar for BC and CO_2, where covariate effects appear to be more discernible than spatial effects. $PM_{2.5}$, on the other hand, illustrates a scenario where spatial smoothing alone captures the major source of heterogeneity. UK-PLS and spatial RF have the best overall performance for all pollutants, while neither shows clearly better or worse accuracy than the other. Figure A.4 shows the cross-validated prediction errors for all models and all pollutants in the Seattle dataset.

Table A.2 compares the predictive performance of each model on the national $PM_{2.5}$ sub-species data. Such comparison reflects various scenarios under which different sources of heterogeneity


Cross-Validated Prediction Errors, UFP
UK-PLS
SpatRF (PL)
RF
TPRS
TPRS-RF
RF-TPRS
Latitude
Longitude
47.8
47.6
47.4
-122.4
-122.3
-122.2
-122.1
2500
0
-2500
-5000


Cross-Validated Prediction Errors, BC
UK-PLS
SpatRF (PL)
RF
TPRS
TPRS-RF
RF-TPRS
Latitude
Longitude
47.8
47.6
47.4
-122.4
-122.3
-122.2
-122.1
250
0
-250
-500


Cross-Validated Prediction Errors, NO2
UK-PLS
SpatRF (PL)
RF
TPRS
TPRS-RF
RF-TPRS
Latitude
Longitude
47.8
47.6
47.4
-122.4
-122.3
-122.2
-122.1
2.5
0.0
-2.5
-5.0

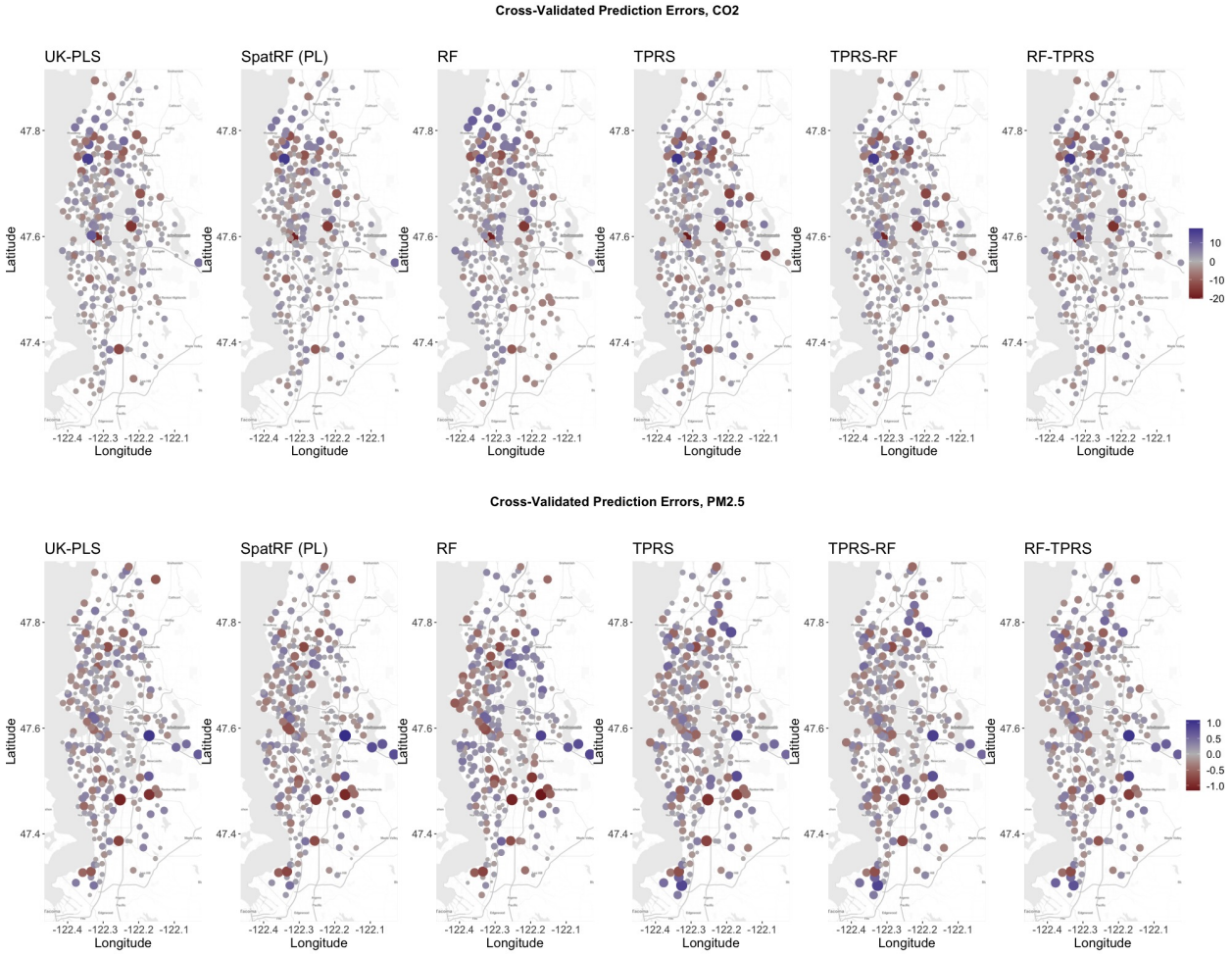


Figure A.4: Prediction errors for all pollutants with all models for the Seattle dataset

best explain the distribution of outcomes. For EC and OC, the majority of variability comes from covariate effects, as indicated by the poor performance of spatial smoothing (TPRS) alone, while non-linear effects (as captured by RF and spatial RF) are more evident for EC. On the contrary, the spatial component captures a considerable amount of variability for Si and S, and RF has the worst performance on them. This agrees with the findings in Bergen et al. (2013) where adding universal kriging on top of PLS leads to clearly improved accuracy. Figure 2.4 shows the cross-validated prediction errors at all study sites for each method.

	UK-PLS	RF	TPRS	RF-TPRS	TPRS-RF	SpatRF (PL)	SpatRF (NP)
EC	0.72	0.79	0.23	0.81	0.73	0.82	0.81
OC	0.59	0.54	0.23	0.64	0.57	0.62	0.59
Si	0.53	0.41	0.56	0.52	0.57	0.55	0.55
S	0.89	0.76	0.93	0.89	0.94	0.90	0.87

Table A.2: Cross-validated R^2 for each method on the national $PM_{2.5}$ sub-species data

Table A.3 presents the prediction R^2 of different models on the synthetic data, where we observe that models capturing the mean (RF) or correlation (TPRS) only have the lowest accuracy, and UK-PLS which only captures linear relationship has worse performance comparing to more flexible models (two-step models or Spatial RF).

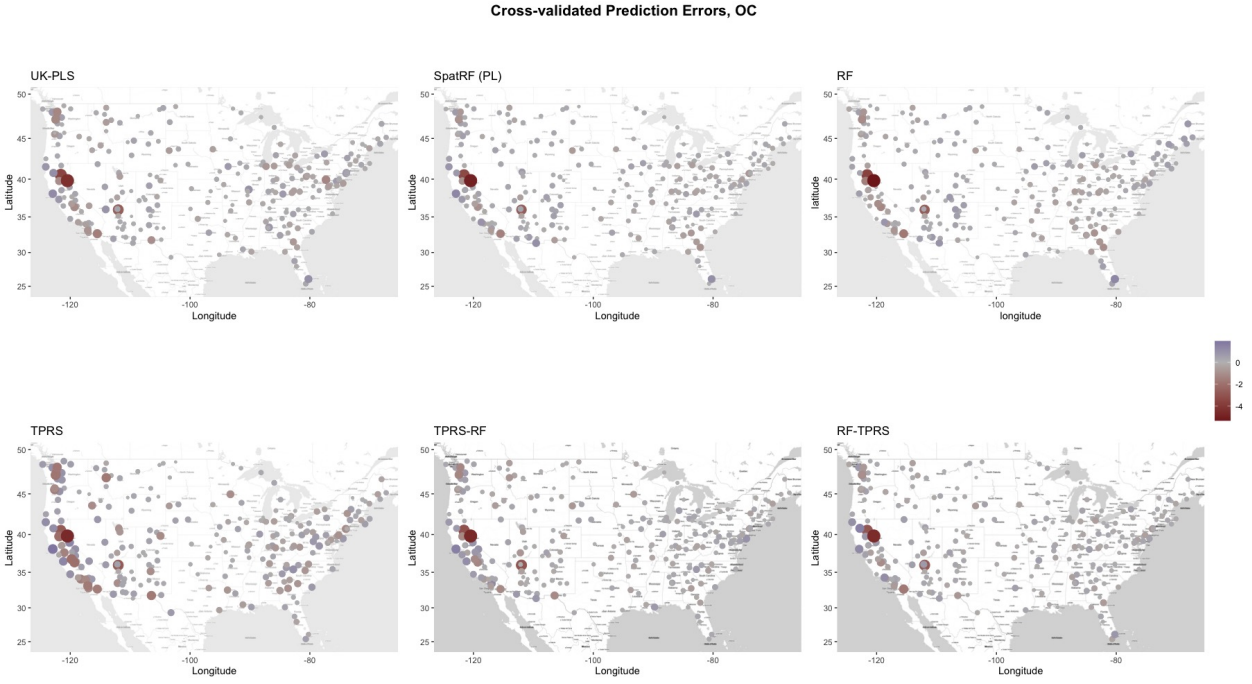
UK-PLS	RF	TPRS	RF-TPRS	TPRS-RF	SpatRF (PL)	SpatRF (NP)
0.62	0.61	0.32	0.69	0.67	0.72	0.72

Table A.3: Cross-validated R^2 for each method on synthetic data

A.2 Variable Importance Analyses

Figures A.6 and A.7 present our full variable importance results, for all pollutants and all models in the Seattle and national studies, respectively.


Cross-validated Prediction Errors, EC
UK-PLS
SpatRF (PL)
RF
TPRS
TPRS-RF
RF-TPRS
Latitude
Longitude
0.5
0.0
-0.5


Cross-validated Prediction Errors, OC
UK-PLS
SpatRF (PL)
RF
TPRS
TPRS-RF
RF-TPRS
Latitude
Longitude
0
-2
-4

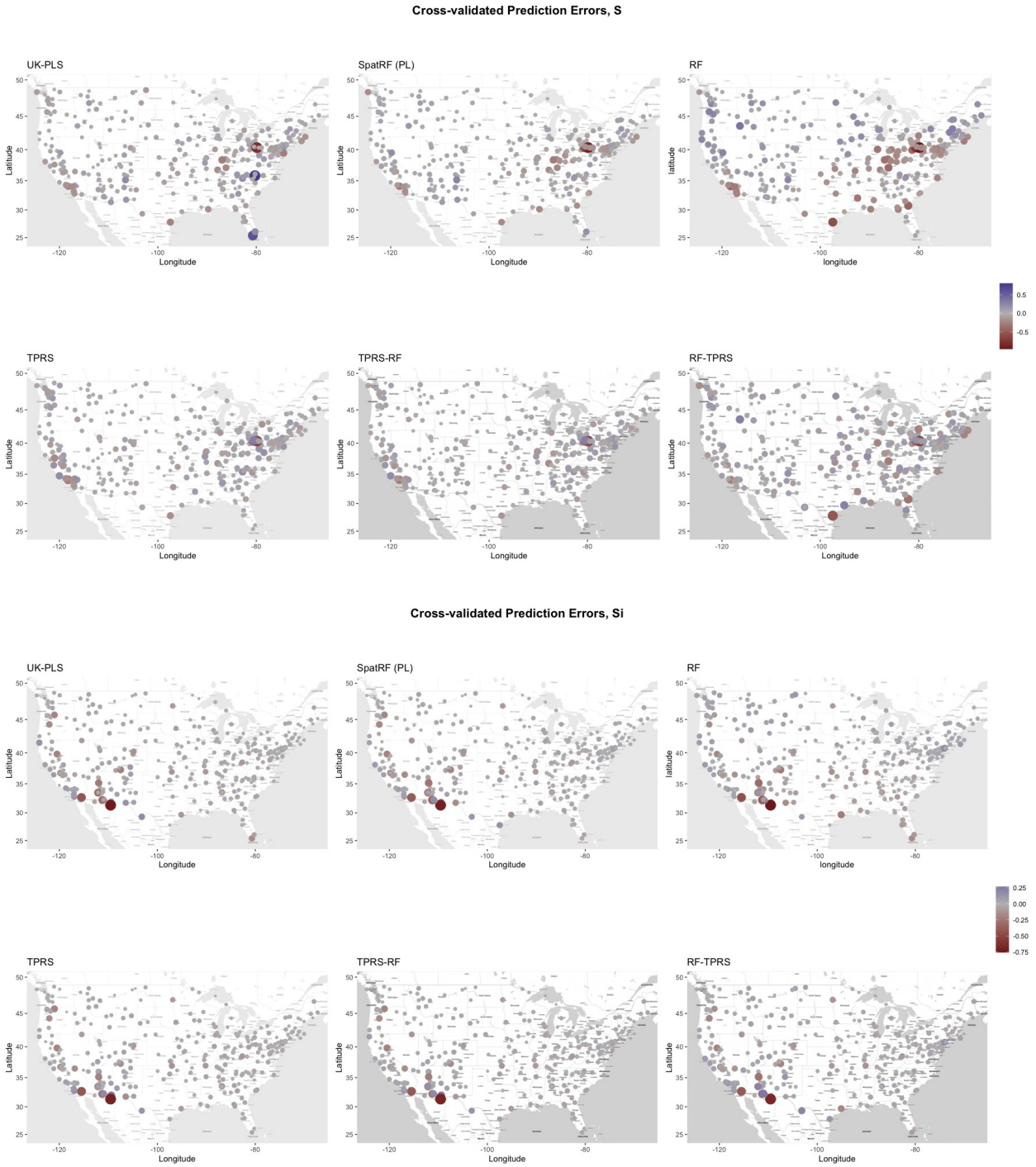
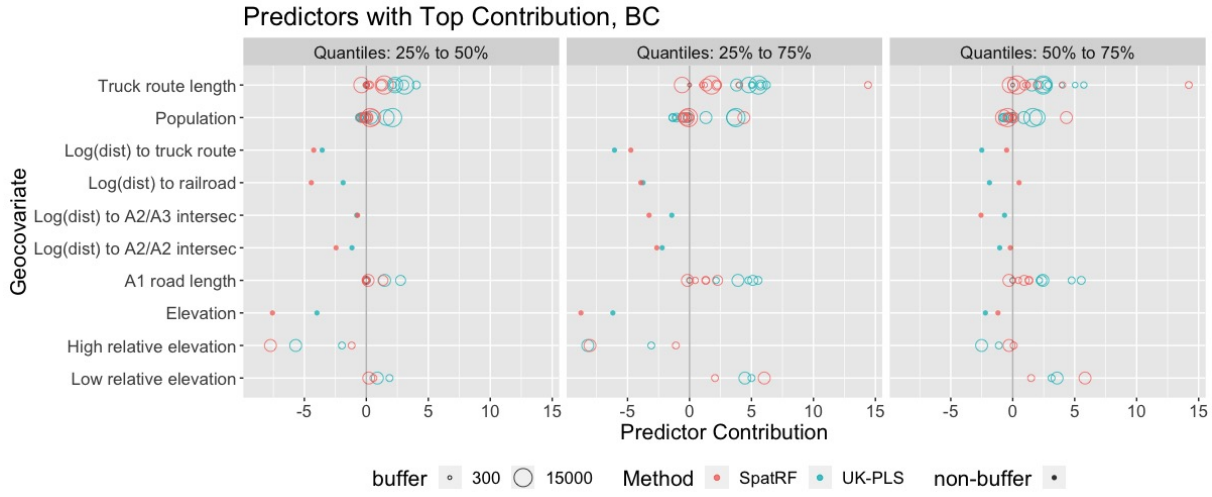


Figure A.5: Prediction errors for all pollutents with all models for the national dataset

Figure A.8 visualizes the proposed variable importance measure together with the maximum absolute correlation between each predictor and each truly active predictor, for the synthetic data.

Due to the autocorrelation between predictors, it is unlikely that any variable importance measure would exactly recover the true predictor contributions. Instead, a reasonably good measure would highlight predictors that are highly correlated the truly active ones, which would explain the mechanism well enough for practical purposes such as prediction. Our method achieves this, as reflected by the observation that predictors found to have high contribution in the first three panels are either the true ones (e.g. annual median NDVI) or highly correlated with at least one truly active predictor (e.g. transportation land use, which has a maximum absolute correlation above 0.75 with the true predictors as seen from the last panel). Through this variable importance measure, both UK-PLS and Spatial RF correctly identified the truly active predictors, despite lower magnitude due to autocorrelation between predictors.



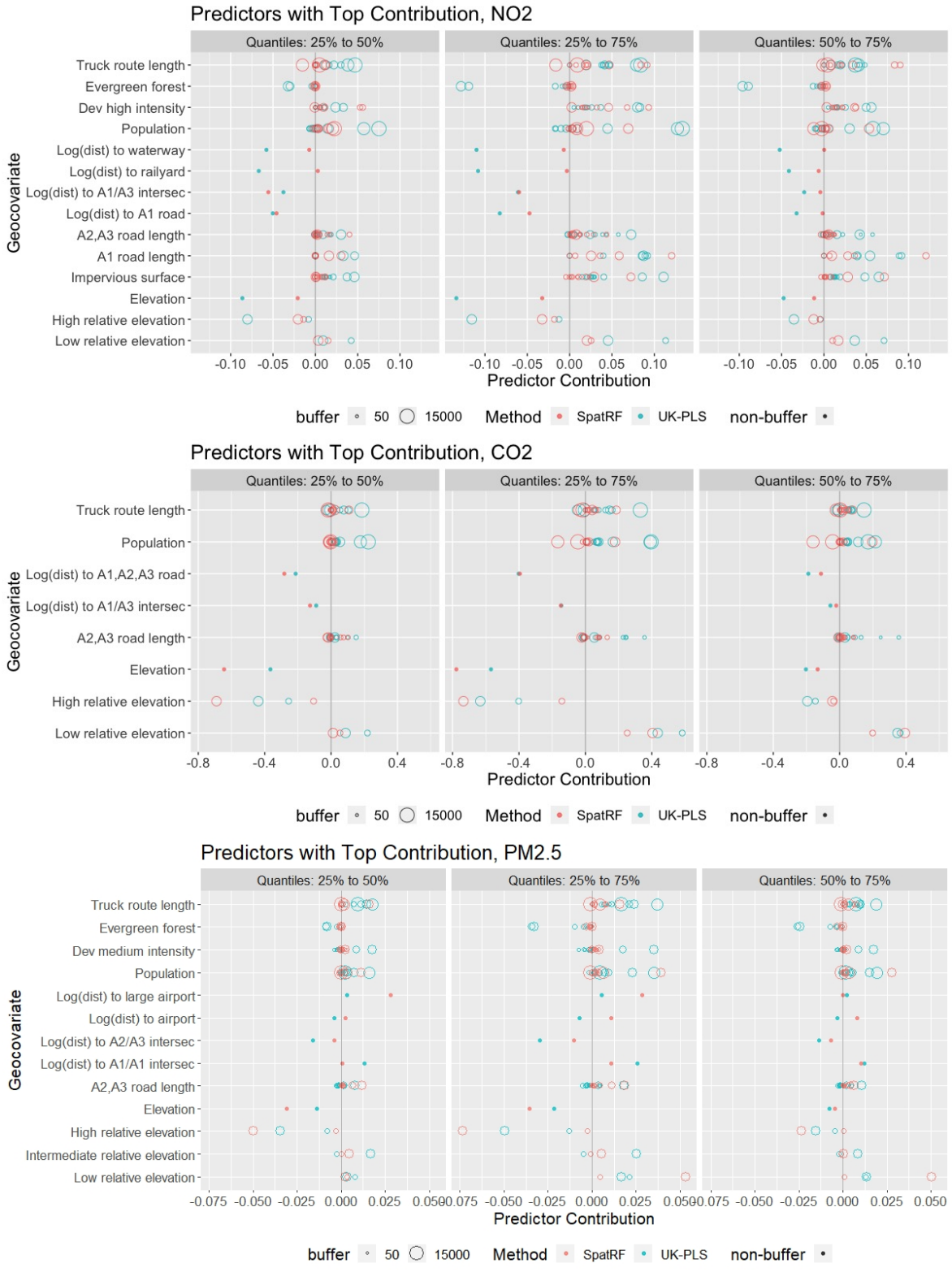


Figure A.6: Variable importance plot for the prediction of BC, NO_2, CO_2 and $PM_{2.5}$ concentrations in the Seattle data, showing predictors with top 5 contribution for either method for at least one contrast.

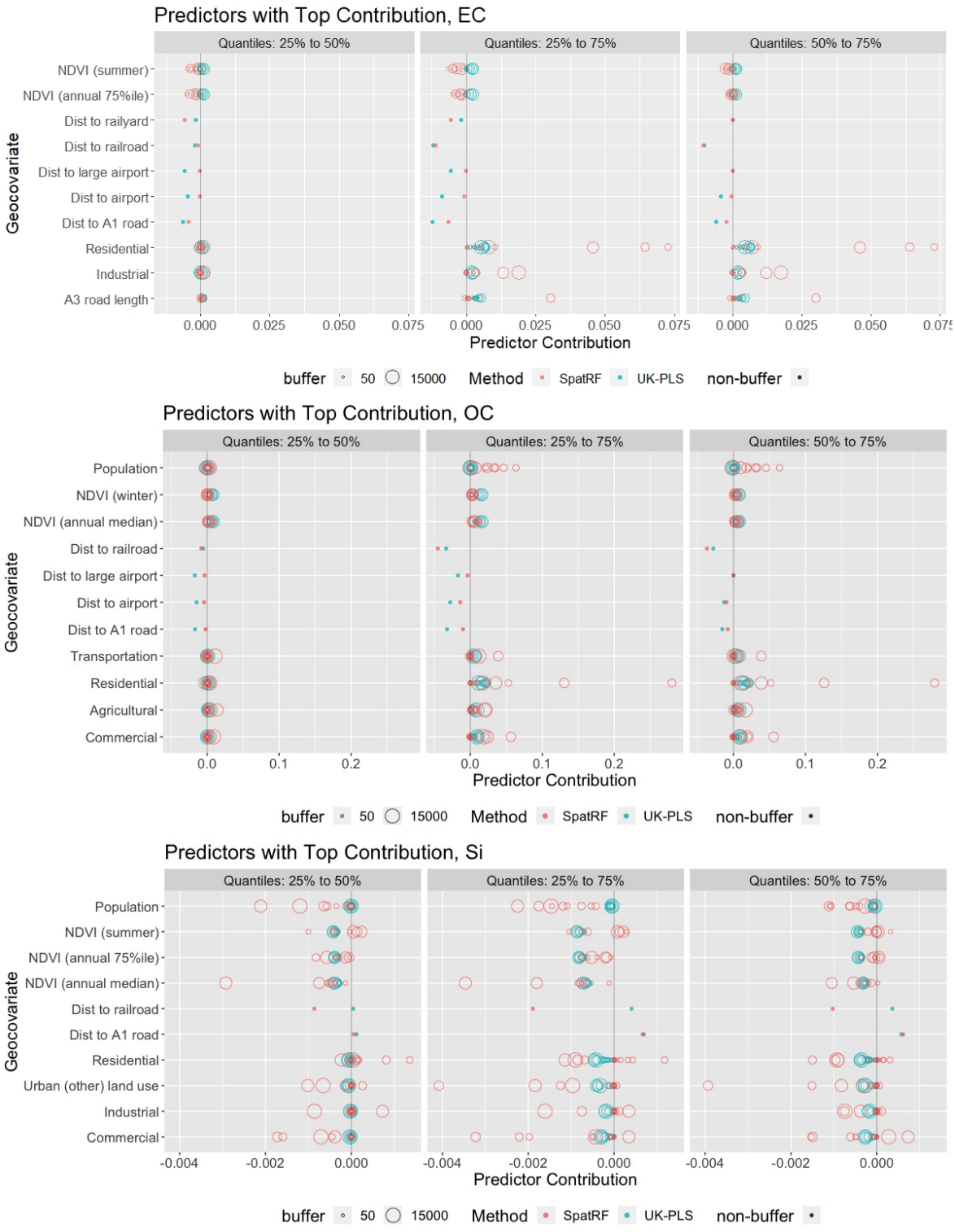


Figure A.7: Variable importance plot for the prediction of EC, OC and Si concentration in the national data, showing predictors with top 5 contribution for either method for at least one contrast.

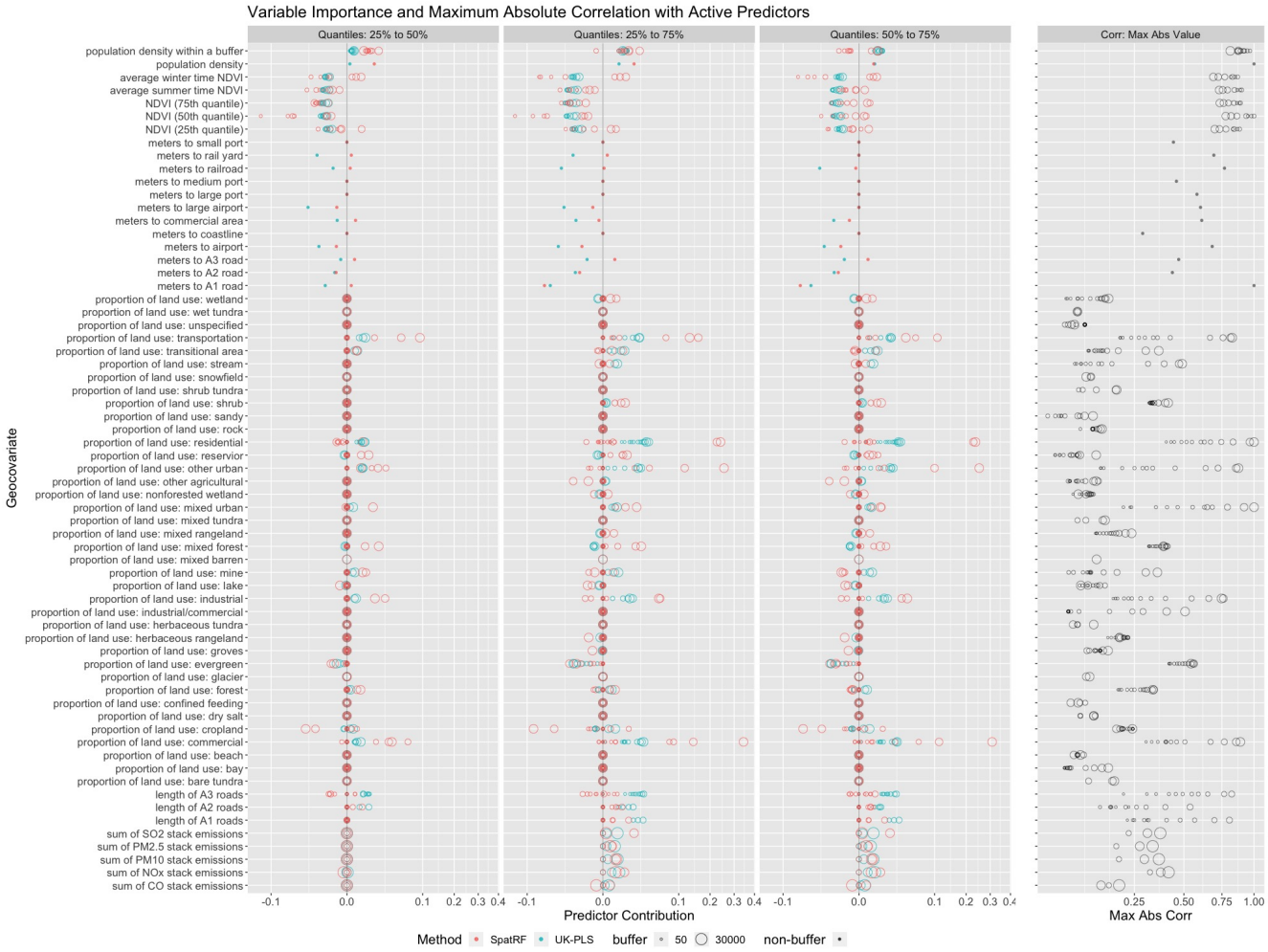


Figure A.8: The full variable importance plot for the synthetic data, along with correlation between each predictor and the active predictors. First three columns: variable importance of spatial RF and UK-PLS; last column: maximum absolute correlation between each predictor and the 5 truly active predictors

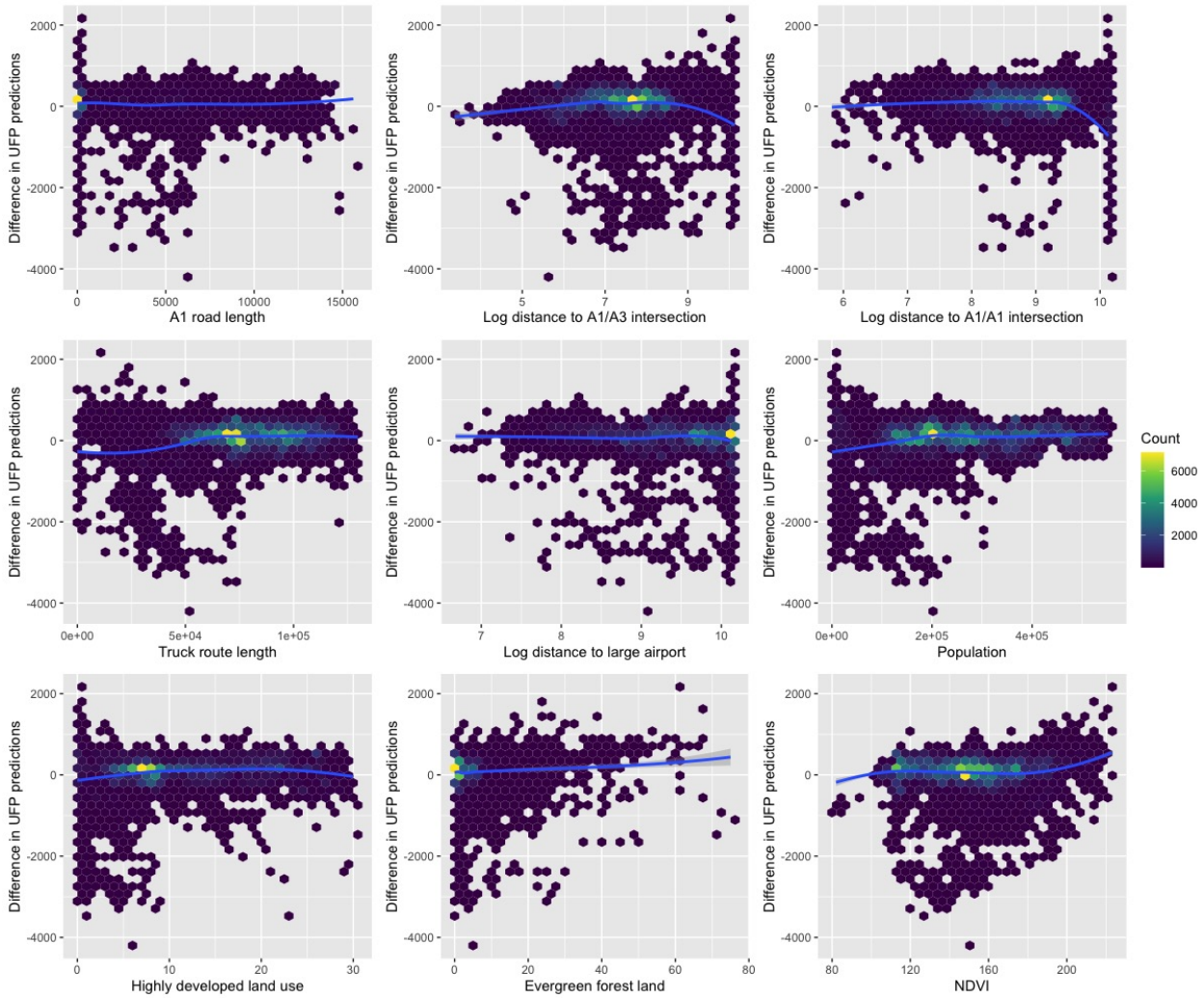


Figure A.9: Hexagonal bin plot showing the difference between spatial RF (PL) and UK-PLS (the subtrahend) predictions of UFP concentration at the residential locations of an epidemiological cohort, versus the distribution of predictors with the greatest difference in variable importance between models. The color reflects the number of points falling to each small region of the plot. Locally weighted scatterplot smoothing (LOESS) curves are added to show the overall trend.

Appendix B

SUPPLEMENTARY MATERIALS FOR CHAPTER 3

B.1 Proof

Proof of Theorem 3.1. Recall the form of the optimization problem given in (3.3):

$$\min_{u,v,\alpha,\beta} \; f_{\gamma,\lambda_1,\lambda_2}(u,v,\alpha,\beta) := \left\|Y^{(l)} - uv^\top\right\|_F^2 + \gamma \left\|u - (K\alpha + B\beta)\right\|_2^2 + \lambda_1 \alpha^\top \tilde{K}\alpha + \lambda_2 \beta^\top \tilde{Q}\beta$$
$$\text{s.t. } u = Y^{(l)}v, \; v^\top v = 1.$$

To find the lth PC via (3.3), we denote $y = Y^{(l)}$ and start by examining the first term in the objective function, which, under the constraint, can be expressed as

$$\begin{aligned}\left\|Y^{(l)} - uv^\top\right\|_F^2 &= \text{tr}\left((y - uv^\top)^\top (y - uv^\top)\right) \\ &= \text{tr}\left(y^\top y - 2y^\top yvv^\top + vv^\top y^\top yvv^\top\right) \\ &\propto -2\text{tr}\left(v^\top VDU^\top UDV^\top v\right) + \text{tr}\left(v^\top vv^\top VDU^\top UDV^\top v\right) \qquad \text{(B.1)}\end{aligned}$$

where $\text{tr}(\cdot)$ denotes the trace of a matrix, $\propto$ means equal up to a constant not depending on u, v, α or β, and we have made use of the cyclic property of trace. Since $v^\top v = 1$ and $U^\top U = I$, denoting $q := V^\top v$, (B.1) continues as

$$\left\|Y^{(l)} - uv^\top\right\|_F^2 = -2\text{tr}\left(q^\top D^2 q\right) + \text{tr}\left(q^\top D^2 q\right) = -q^\top D^2 q.$$

Under the reparametrization in Theorem 3.1, namely, $\eta := \left[\alpha^\top \quad \sqrt{\frac{\lambda_2}{\lambda_1}}\beta^\top\right]^\top$ and $Z := \left[K \quad \sqrt{\frac{\lambda_2}{\lambda_1}}B\right]$, the objective function becomes

$$f_{\gamma,\lambda_1,\lambda_2}(u,v,\eta) = -q^\top D^2 q + \gamma \left\|UDq - Z\eta\right\|_2^2 + \lambda_1 \eta^\top P\eta$$

which is a quadratic function of η for fixed q, and

$$\frac{\partial f}{\partial \eta} = 2\gamma Z^\top (Z\eta - UDq) + 2\lambda_1 P\eta.$$

Since the hessian $\partial^2 f/\partial\eta^2 = 2\gamma Z^\top Z$ is positive semidefinite, we can profile out η by setting it to the minimizer $\tilde{\eta}(q) := (\gamma Z^\top Z + \lambda_1 P)^{-1}(\gamma Z^\top UDq)$. The objective function can thus be rearranged as

$$\begin{aligned} f_{\gamma,\lambda_1,\lambda_2}(u,v,\alpha,\beta) := g_{\gamma,\lambda_1}(q,\tilde{\eta}) &= -q^\top D^2 q + \gamma(UDq - Z\tilde{\eta})^\top(UDq - Z\tilde{\eta}) + \lambda_1\tilde{\eta}^\top P\tilde{\eta} \\ &= -q^\top D^2 q + \gamma(q^\top D^2 q - 2\tilde{\eta}^\top Z^\top UDq + \tilde{\eta}^\top Z^\top Z\tilde{\eta}) + \lambda_1\tilde{\eta}^\top P\tilde{\eta} \\ &= (\gamma-1)q^\top D^2 q - 2\gamma^2 q^\top DU^\top Z(\gamma Z^\top Z + \lambda_1 P)^{-1}Z^\top UDq \\ &\quad + \gamma^2 q^\top DU^\top Z(\gamma Z^\top Z + \lambda_1 P)^{-1}(\gamma Z^\top Z + \lambda_1 P)(\gamma Z^\top Z + \lambda_1 P)^{-1}Z^\top UDq \\ &= -q^\top\left[-(\gamma-1)D^2 + \gamma^2 DU^\top Z(\gamma Z^\top Z + \lambda_1 P)^{-1}Z^\top UD\right]q := -q^\top Aq \end{aligned} \quad \text{(B.2)}$$

where $A := -(\gamma-1)D^2 + \gamma^2 DU^\top Z(\gamma Z^\top Z + \lambda_1 P)^{-1}Z^\top UD$.

It then follows that the original optimization problem (3.3) is equivalent to

$$\min_q \; f_{\gamma,\lambda_1}(q) := -q^\top Aq \quad \text{s.t. } q^\top q = 1,$$

for which the optimal solution is the normalized first eigenvector of A (i.e. the one corresponding to the largest eigenvalue). Recalling the form of $\tilde{\eta}(q)$ and the fact that the untransformed optimal solution satisfies $V^\top\tilde{v} = \tilde{q}$ and $\tilde{u} = Y^{(l)}\tilde{v}$, this completes the proof. □

B.2 Additional Numerical Results

We use the first replicate (out of 100 total) of data in Simulation Scenario 1 (Section 3.4.1) to numerically verify our claim in Theorem 3.1, i.e., the optimality of the derived solution.

Specifically, we solve the optimization problem (3.3) to extract the first PC score, and vary the first two entries of the optimal loadings $\tilde{v}$ while keeping everything else (including the PC scores $\tilde{u}$ and coefficients $\tilde{\eta}$) intact, and ensuring that the constraint $v^\top v$ is still satisfied. This is done under polar coordinates: letting $\theta \in [0, 2\pi)$, we write $v^*(\theta)$ as the modified loading vector, where the entries $v_j^*(\theta) = \tilde{v}_j$ for $j > 2$. We let $v_1^*(\theta) = \rho\sin\theta$ and $v_2^*(\theta) = \rho\cos\theta$ where $\rho = \sqrt{1 - \sum_{j>2}\tilde{v}_j^2}$. Then as θ varies within $[0, 2\pi)$, the two entries $(v_1^*(\theta), v_2^*(\theta))$ take all possible combinations that satisfy the constraint.

We explore different values of tuning parameters $\gamma, \lambda_1, \lambda_2$ and compare the objective function $f_{\gamma,\lambda_1,\lambda_2}(u,v,\eta)$ evaluated at the optimal solution $(\tilde{u},\tilde{v},\tilde{\eta})$ versus the modified solution $(\tilde{u}, v^*(\theta), \tilde{\eta})$

for $\theta \in [0, 2\pi)$. We plot their differences against θ for various values of tuning parameters in Figures B.1 and B.2. The fact that each curve is always above or equal to zero, and reaches zero exactly once, (partially) verifies the uniqueness of the optimal solution, as well as the optimality of the proposed form of $\tilde{v}$, at least for fixed $\tilde{u}$ and $\tilde{\eta}$.

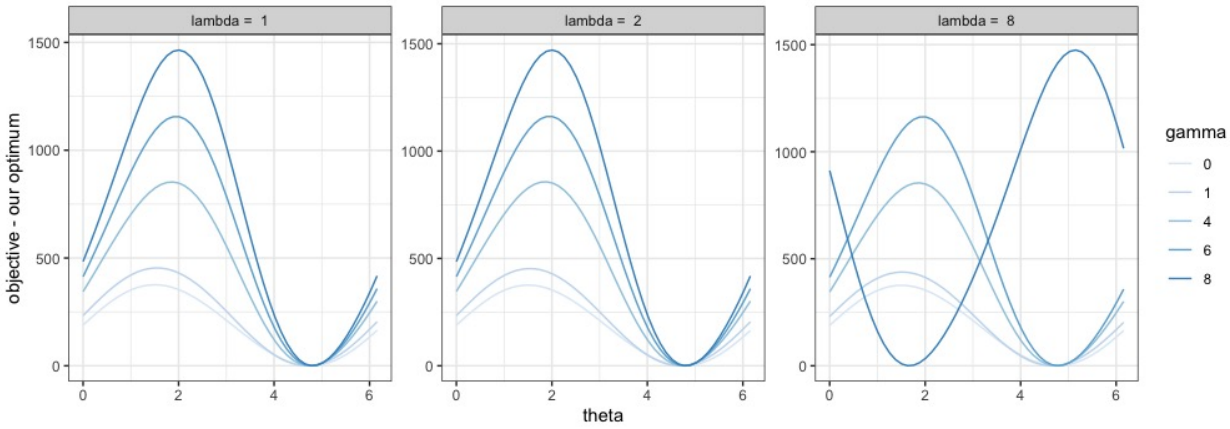


Figure B.1: Differences between the modified objective function versus the optimum achieved by our solution, plotted against θ, where we fix $\lambda_1 = \lambda_2$. Each panel corresponds to a value of λ_1 and λ_2, and the colors of curves reflect the values of γ.

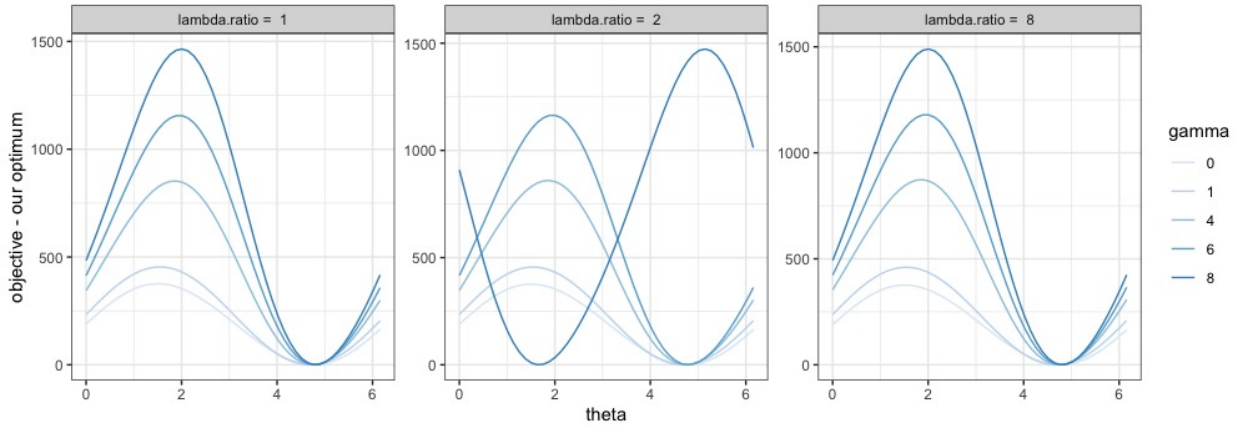


Figure B.2: Differences between the modified objective function versus the optimum achieved by our solution, plotted against θ, where we fix $\lambda_1 = 1$. Each panel corresponds to different ratios λ_2/λ_1, and the colors of curves reflect the values of γ.

Appendix C

SUPPLEMENTARY MATERIALS FOR CHAPTER 4

C.1 Summary of Related Methods

Please refer to Table C.1 for a summary of existing models for spatial point process or areal data.

C.2 Additional Results

Figures C.1 and C.2 present analogs of Figures 4.5 and 4.6, but with the BYM2 and LGCP models implemented in `RStan` based on two MCMC chains with 5000 samples each. All slope parameters are assigned Normal$(0, 10)$ priors, and all variance parameters are assigned truncated Normal$(0, 5)$ priors. We observe that the estimation and inference of the Bayesian models remain similar to the results shown in Figure 4.5, while LGCP models fitted by `RStan` are more sensitive to the inclusion of the spatially structured covariate (Figure C.2), compared with `INLA`. In other words, the estimated the slope parameters are highly simlar between `RStan` and `INLA`, while the predicted spatial random effects from `INLA` appear to be more robust against spatial confounding. Such robustness is likely due to the properties of the PC priors (Simpson et al., 2017), along with the computational advantages of `INLA`.

C.3 Proofs

This section includes proofs for our theoretical claims in Section 4.3. We reintroduce our notation for clarity. The true, continuous baseline intensity is denoted as $\alpha^0(\cdot)$, and the true regression parameters are denoted as $\boldsymbol{\beta}^0$. We denote the discretized baseline vector, i.e. $\alpha^0(\cdot)$ evaluated at locations $\boldsymbol{s} = (s_1, \ldots, s_n)$, as $\tilde{\boldsymbol{\alpha}}(\boldsymbol{s})$ to distinguish it from the baseline intensity function. We also define $\phi(s) = \log \mathbb{E}_0[\exp \varepsilon(s)]$, $\boldsymbol{A} = (|\Omega_1|, \ldots, |\Omega_n|)$, $\boldsymbol{B} = (P_1|\Omega_1|, \ldots, P_n|\Omega_n|)$, and recall that $\ell(\cdot)$ is the Possion log-likelihood as defined in Section 4.2.

Empirical process notations are adopted, where under discretization of the observation window Ω, we denote $\mathbb{P}_0 f(\theta; X, Y) := \mathbb{E}_0[f(\theta; X, Y)]$ with $\mathbb{E}_0$ being the expectation taken under the true

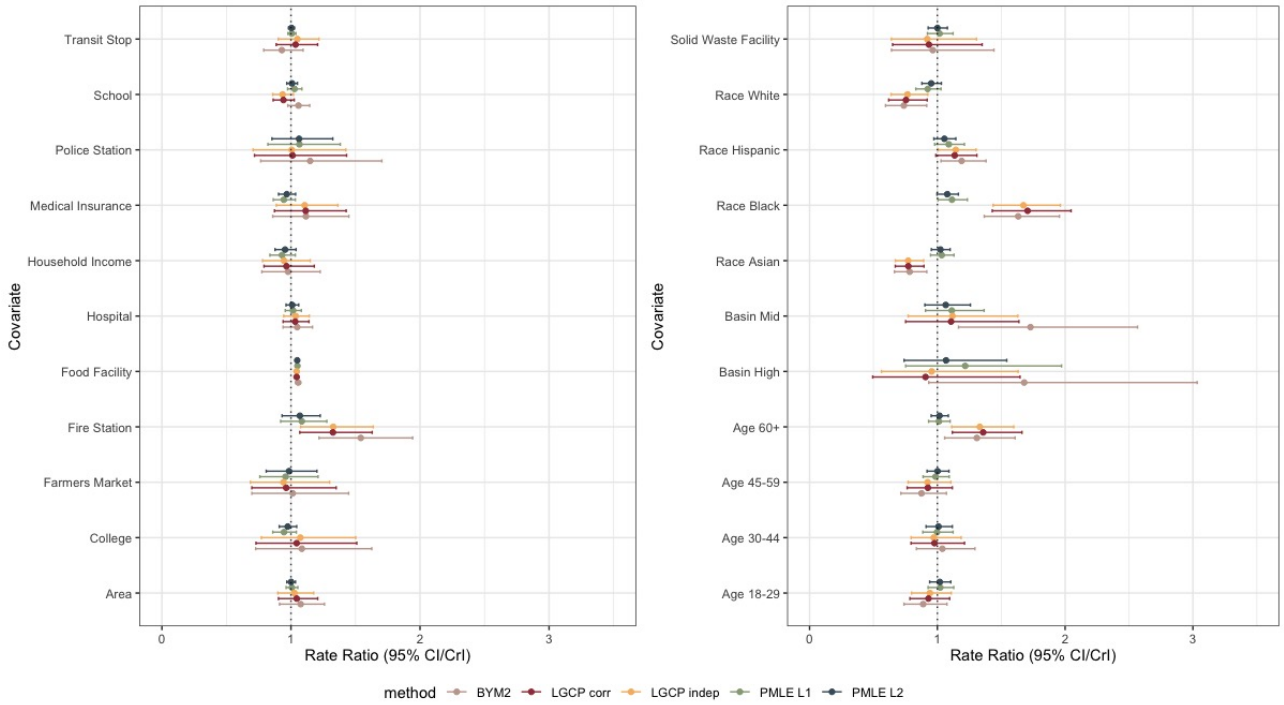


Figure C.1: Estimated rate ratios with error bars indicating 95% confidence/credible intervals from LGCP models fitted by `RStan`.

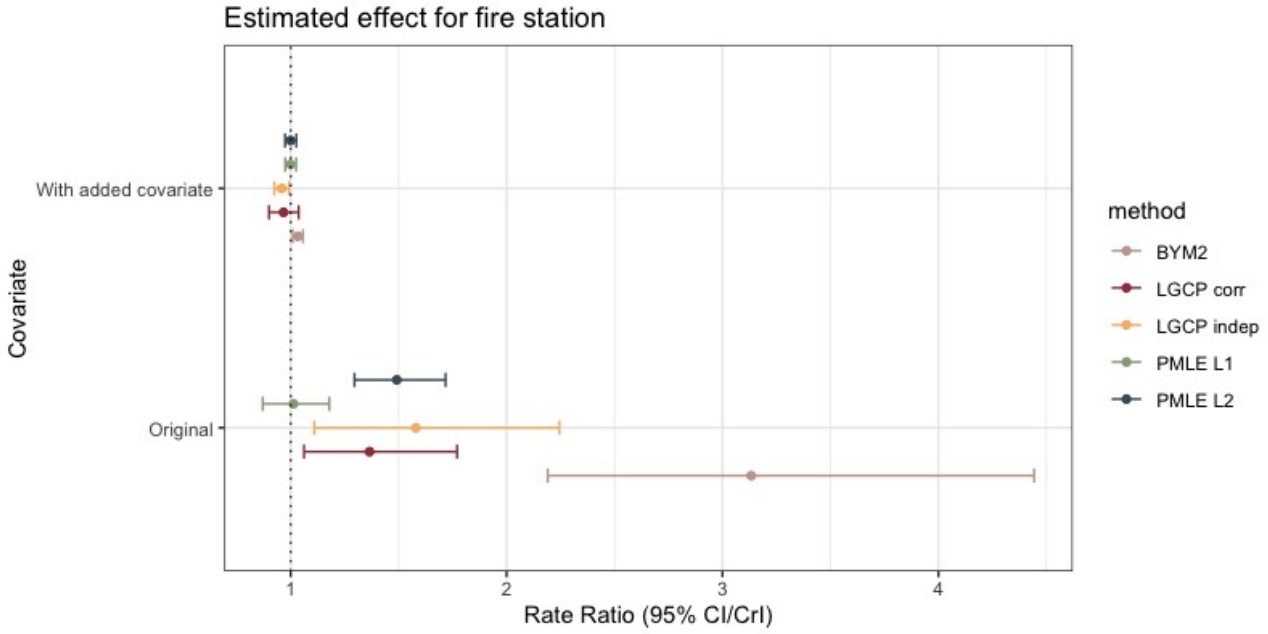


Figure C.2: Comparison of estimated coefficients and 95% CI/CrI before and after adding a spatially structured covariate to the single covariate model of fire station, with LGCP models fitted by `RStan`.

distribution of X, Y, and $\mathbb{P}_n f(\theta; X, Y) := n^{-1} \sum_i f(\theta; X_i, Y_i)$.

We first prove Lemma 4.1 by examining the relationship between the target parameter $\boldsymbol{\theta}^\dagger := (\tilde{\boldsymbol{\alpha}}^\dagger, \boldsymbol{\beta}^\dagger)$, which is the solution to

$$-\nabla_{(\tilde{\boldsymbol{\alpha}}, \boldsymbol{\beta})} \mathbb{P}_0 \ell(\tilde{\boldsymbol{\alpha}}, \boldsymbol{\beta}) = 0,$$

and the true parameter $\boldsymbol{\beta}^0$ along with the function $\alpha^0(\cdot)$ underlying the Cox process. In particular, we show that the Poisson likelihood yields an unbiased estimating equation for $\boldsymbol{\beta}$ despite the ignored error random field as well as misspecification of $\alpha^0(\cdot)$. With the fusion penalty $R(\tilde{\boldsymbol{\alpha}}; \mathcal{G}_n)$ incorporated into the objective function, we further bound the gap between the penalized solution $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^\dagger$ under different conditions on the smoothness of $\alpha^0(\cdot)$.

We then use empirical process arguments to show the convergence of the penalized PMLE to the target parameters, following a similar outline as in Haris et al. (2019b), with an adaptation to the heavy-tailed distribution of the observations in our setting due to double stochasticity.

Finally, we establish the asymptotic linearity of the de-biased estimator $\hat{\boldsymbol{b}}$ and in turn show the validity of our variance estimator along with the inference procedure.

C.3.1 Consistency

Throughout this section, we denote the smooth portion of our objective function as

$$\mathcal{L}(\boldsymbol{\theta}) := -\ell(\tilde{\boldsymbol{\alpha}}, \boldsymbol{\beta}) + \gamma_n R(\tilde{\boldsymbol{\alpha}}; \mathcal{G}_n).$$

Also, without loss of generality, we assume a uniform offset $\boldsymbol{P} = (1, \ldots, 1)^\top$ across all regions.

Proof of Lemma 4.1. Note that for region $\Omega_i, i \in \{1, \ldots, n\}$,

$$\begin{aligned}
-\mathbb{P}_0 \frac{\partial \ell}{\partial \tilde{\alpha}_i} &= -\mathbb{E}_X \left[\mathbb{E}_\varepsilon \mathbb{E}[Y_i \mid \varepsilon(\cdot)] + |\Omega_i| \exp\left(\tilde{\alpha}_i + X_i \boldsymbol{\beta}\right) \right] \\
&= -\mathbb{E}_X \left[\mathbb{E}_\varepsilon \int_{\Omega_i} \Lambda(s) \mathrm{d}s + |\Omega_i| \exp\left(\tilde{\alpha}_i + X_i \boldsymbol{\beta}\right) \right] \\
&= -\mathbb{E}_X \left[\int_{\Omega_i} \mathbb{E}_\varepsilon \exp\left[\alpha^0(s) + X_i \boldsymbol{\beta}^0 + \varepsilon(s)\right] \mathrm{d}s + |\Omega_i| \exp\left(\tilde{\alpha}_i + X_i \boldsymbol{\beta}\right) \right] \qquad \text{(C.1)} \\
&= -\mathbb{E}_X \left[\int_{\Omega_i} \exp\left[\alpha^0(s) + X_i \boldsymbol{\beta}^0 + \phi(s)\right] \mathrm{d}s + |\Omega_i| \exp\left(\tilde{\alpha}_i + X_i \boldsymbol{\beta}\right) \right], \qquad \text{(C.2)}
\end{aligned}$$

where $\mathbb{E}_\varepsilon$ denotes expectation taken with respect to the error random field. "=" in (C.1) holds due to Fubini's Theorem under Assumption 4.1. Furthermore, the mean value theorem for integrals together with Assumption 4.1 imply the existence of some $s_i^* \in \Omega_i$ such that

$$\begin{aligned}
&\int_{\Omega_i} \exp\left[\alpha^0(s) + X_i \boldsymbol{\beta}^0 + \phi(s)\right] \mathrm{d}s + |\Omega_i| \exp\left(\tilde{\alpha}_i + X_i \boldsymbol{\beta}\right) \\
&= -|\Omega_i| \exp\left[\alpha^0(s_i^*) + X_i \boldsymbol{\beta}^0 + \phi(s_i^*)\right] + |\Omega_i| \exp\left(\tilde{\alpha}_i + X_i \boldsymbol{\beta}\right) \qquad \text{(C.3)}
\end{aligned}$$

for any realization of X. We write $\boldsymbol{s}^* = (s_1^*, \ldots, s_n^*)$ such that $s_i^* \in \Omega_i$ for all i. Define the target parameter $\boldsymbol{\theta}^\dagger = (\boldsymbol{\alpha}^\dagger, \boldsymbol{\beta}^\dagger)$ such that $\boldsymbol{\alpha}^\dagger = \alpha^0(\boldsymbol{s}^*) + \phi(\boldsymbol{s}^*)$ and $\boldsymbol{\beta}^\dagger = \boldsymbol{\beta}^0$. Examining the expression in (C.3) leads to

$$-\frac{\partial}{\partial \tilde{\alpha}_i} \mathbb{P}_0 \ell(\tilde{\boldsymbol{\alpha}}, \boldsymbol{\beta}) \big|_{\boldsymbol{\alpha}^\dagger, \boldsymbol{\beta}^\dagger} = 0.$$

Likewise, since

$$-\nabla_{\boldsymbol{\beta}} \mathbb{E}_0[\ell(\tilde{\boldsymbol{\alpha}}, \boldsymbol{\beta}) \mid X] = -\sum_{i=1}^n X_i \frac{\partial}{\partial \tilde{\alpha}_i} \mathbb{E}_0[\ell(\tilde{\boldsymbol{\alpha}}, \boldsymbol{\beta}) \mid X],$$

we also have

$$-\nabla_{\boldsymbol{\beta}} \mathbb{P}_0 \ell(\tilde{\boldsymbol{\alpha}}, \boldsymbol{\beta}) \big|_{\boldsymbol{\alpha}^\dagger, \boldsymbol{\beta}^\dagger} = 0.$$

Together with the convexity of $-\ell(\cdot)$, we established the form of target parameter $\boldsymbol{\theta}^\dagger$ as claimed in Lemma 4.1.

We now examine the minimizer $\boldsymbol{\theta}^*$ of $\mathcal{L}(\cdot)$. First, observe that $\mathcal{L}(\boldsymbol{\theta})$ involves $\boldsymbol{\beta}$ only through $-\ell(\boldsymbol{\theta})$, we immediately have $\nabla_{\boldsymbol{\beta}}\mathcal{L}(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^\dagger) = 0$ which yields $\boldsymbol{\beta}^* = \boldsymbol{\beta}^\dagger = \boldsymbol{\beta}^0$ again by the convexity of $-\ell(\cdot)$.

We first discuss the case with ℓ_2 smoothing penalty. By the optimality of $\boldsymbol{\theta}^*$ and conducting a Taylor expansion of $\mathcal{L}$ around $\boldsymbol{\theta}^\dagger$, we have

$$\begin{aligned}
0 &= -\nabla_{\tilde{\boldsymbol{\alpha}}}[\mathbb{P}_0\ell(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) + \gamma_n R(\boldsymbol{\alpha}^*)] \\
&= -\mathbb{E}_0\left[\boldsymbol{Y} - A \odot \exp(\boldsymbol{\alpha}^* + \boldsymbol{X}\boldsymbol{\beta}^*)\right] + \gamma_n \nabla_{\tilde{\boldsymbol{\alpha}}} R(\boldsymbol{\alpha}^*) \qquad \text{(C.4)} \\
&= -\mathbb{E}_0\left[\boldsymbol{Y} - A \odot \exp(\boldsymbol{\alpha}^\dagger + \boldsymbol{X}\boldsymbol{\beta}^\dagger)\right] + \gamma_n \nabla_{\tilde{\boldsymbol{\alpha}}} R(\boldsymbol{\alpha}^\dagger) + \left[\operatorname{diag}\left(A \odot \exp(\breve{\boldsymbol{\alpha}} + \boldsymbol{X}\boldsymbol{\beta}^\dagger)\right) + \gamma_n \nabla^2_{\tilde{\boldsymbol{\alpha}}} R(\breve{\boldsymbol{\alpha}})\right]\left(\boldsymbol{\alpha}^* - \boldsymbol{\alpha}^\dagger\right)
\end{aligned}$$

for some $\breve{\boldsymbol{\alpha}}$ between $\boldsymbol{\alpha}^*$ and $\boldsymbol{\alpha}^\dagger$, where $\odot$ denotes element-wise multiplication. Since $-\mathbb{E}_0\left[\boldsymbol{Y} - A \odot \exp(\boldsymbol{\alpha}^\dagger + \boldsymbol{X}\boldsymbol{\beta}^\dagger)\right] = 0$, we then have

$$\boldsymbol{\alpha}^* - \boldsymbol{\alpha}^\dagger = \left[\operatorname{diag}\left(A \odot \exp(\breve{\boldsymbol{\alpha}} + \boldsymbol{X}\boldsymbol{\beta}^\dagger)\right) + \gamma_n \nabla^2_{\tilde{\boldsymbol{\alpha}}} R(\breve{\boldsymbol{\alpha}})\right]^{-1}\left[\gamma_n \nabla_{\tilde{\boldsymbol{\alpha}}} R(\boldsymbol{\alpha}^\dagger)\right]$$

and

$$\begin{aligned}
\|\boldsymbol{\alpha}^* - \boldsymbol{\alpha}^\dagger\|_2 &\le \gamma_n \left\| \left[\operatorname{diag}\left(A \odot \exp(\breve{\boldsymbol{\alpha}} + \boldsymbol{X}\boldsymbol{\beta}^\dagger)\right) + \gamma_n \nabla^2_{\tilde{\boldsymbol{\alpha}}} R(\breve{\boldsymbol{\alpha}})\right]^{-1} \right\|_2 \|\nabla_{\tilde{\boldsymbol{\alpha}}} R(\boldsymbol{\alpha}^\dagger)\|_2 \\
&\le \gamma_n \lambda_{\min}\left[\operatorname{diag}\left(A \odot \exp(\breve{\boldsymbol{\alpha}} + \boldsymbol{X}\boldsymbol{\beta}^\dagger)\right) + \gamma_n (L_n + \delta I_n)\right]^{-1} \frac{1}{2}\|(L_n + \delta I_n)\boldsymbol{\alpha}^\dagger\|_2 \\
&\le \frac{\gamma_n}{2}\|\tilde{L}_n \boldsymbol{\alpha}^\dagger(\boldsymbol{s}^*)\|_2 \left[\max_i A_i \exp(\breve{\boldsymbol{\alpha}}_i + X_i\boldsymbol{\beta}^\dagger)\right]^{-1} \\
&= \gamma_n \|\tilde{L}_n \boldsymbol{\alpha}^\dagger(\boldsymbol{s}^*)\|_2 \cdot O_P(1)
\end{aligned}$$

where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue of a matrix. We have made use of the boundedness of $\max_i |\Omega_i|$, the continuity of $\alpha(\cdot)$, and the fact that $L_n + \delta I_n$ is positive semi-definite. The claim in Lemma 4.1 regarding the ℓ_2 smoothing penalty then follows.

For the ℓ_1 fusion penalty, recall that the gradient of the smoothed penalty is $\nabla_{\tilde{\boldsymbol{\alpha}}} R(\tilde{\boldsymbol{\alpha}}) = B_n^\top S_\infty(\gamma_n B_n \tilde{\boldsymbol{\alpha}}/\xi)$ and that $S_\infty(\cdot)$ represents projection onto the ℓ_∞ unit ball. Continuing from

(C.4) with a Taylor expansion of $\ell(\tilde{\boldsymbol{\alpha}}, \boldsymbol{\beta})$ with respect to $\tilde{\boldsymbol{\alpha}}$,

$$\begin{aligned}
0 &= -\mathbb{E}_0\left[\boldsymbol{Y} - A \odot \exp(\boldsymbol{\alpha}^* + \boldsymbol{X}\boldsymbol{\beta}^*)\right] + \gamma_n \nabla_{\tilde{\boldsymbol{\alpha}}} R(\boldsymbol{\alpha}^*) \\
&= -\mathbb{E}_0\left[\boldsymbol{Y} - A \odot \exp(\boldsymbol{\alpha}^\dagger + \boldsymbol{X}\boldsymbol{\beta}^\dagger)\right] + \gamma_n \nabla_{\tilde{\boldsymbol{\alpha}}} R(\boldsymbol{\alpha}^*) + \left[\operatorname{diag}\left(A \odot \exp(\breve{\boldsymbol{\alpha}} + \boldsymbol{X}\boldsymbol{\beta}^\dagger)\right)\right]\left(\boldsymbol{\alpha}^* - \boldsymbol{\alpha}^\dagger\right) \\
&= -\mathbb{E}_0\left[\boldsymbol{Y} - A \odot \exp(\boldsymbol{\alpha}^\dagger + \boldsymbol{X}\boldsymbol{\beta}^\dagger)\right] \\
&\quad + \gamma_n \nabla_{\tilde{\boldsymbol{\alpha}}} R(\boldsymbol{\alpha}^\dagger) + \left[\operatorname{diag}\left(A \odot \exp(\breve{\boldsymbol{\alpha}} + \boldsymbol{X}\boldsymbol{\beta}^\dagger)\right)\right]\left(\boldsymbol{\alpha}^* - \boldsymbol{\alpha}^\dagger\right) + \gamma_n\left[\nabla_{\tilde{\boldsymbol{\alpha}}} R(\boldsymbol{\alpha}^*) - \nabla_{\tilde{\boldsymbol{\alpha}}} R(\boldsymbol{\alpha}^\dagger)\right]
\end{aligned}$$

which leads to

$$\begin{aligned}
\|\boldsymbol{\alpha}^* - \boldsymbol{\alpha}^\dagger\|_1 &\le \gamma_n \left[\max_i A_i \exp(\breve{\boldsymbol{\alpha}}_i + X_i\boldsymbol{\beta}^\dagger)\right]^{-1} \left[\left\|\nabla_{\tilde{\boldsymbol{\alpha}}} R(\boldsymbol{\alpha}^\dagger)\right\|_1 + \left\|\nabla_{\tilde{\boldsymbol{\alpha}}} R(\boldsymbol{\alpha}^*) - \nabla_{\tilde{\boldsymbol{\alpha}}} R(\boldsymbol{\alpha}^\dagger)\right\|_1\right] \\
&\le \gamma_n \left[\max_i A_i \exp(\breve{\boldsymbol{\alpha}}_i + X_i\boldsymbol{\beta}^\dagger)\right]^{-1} \left[\left\|B_n^\top S_\infty\left(\frac{\gamma_n B_n \boldsymbol{\alpha}^\dagger}{\xi}\right)\right\|_1\right. \\
&\qquad \left. + \left\|B_n^\top\left(S_\infty\left(\frac{\gamma_n B_n \boldsymbol{\alpha}^*}{\xi}\right) - S_\infty\left(\frac{\gamma_n B_n \boldsymbol{\alpha}^\dagger}{\xi}\right)\right)\right\|_1\right] \\
&\le \gamma_n O_P(1)\left[\frac{\gamma_n}{\xi}\|B_n^\top\|_1\|B_n\boldsymbol{\alpha}^\dagger\|_1 + \frac{\gamma_n}{\xi}\|B_n^\top\|_1\|B_n\|_1\|\boldsymbol{\alpha}^* - \boldsymbol{\alpha}^\dagger\|_1\right] \qquad \text{(C.5)}
\end{aligned}$$

where (C.5) holds because $\|S_\infty(u)\|_1 \le \|u\|_1$ and $\|S_\infty(u) - S_\infty(v)\|_1 \le \|u - v\|_1$ for any vectors u and v. Noting that $\|B_n^\top\|_1 = 2$ (which is the maximum row sum of absolute values of B_n) and $\|B_n\|_1 = \max_i d_i$ (which is the maximum column sum of absolute values of B_n), when n is large so that $\gamma_n^2 \max_i d_i < \xi/2$, (C.5) yields

$$\frac{1}{2}\|\boldsymbol{\alpha}^* - \boldsymbol{\alpha}^\dagger\|_1 \le \frac{\gamma_n^2}{\xi} O_P(1)\|B_n\boldsymbol{\alpha}^\dagger\|_1,$$

which completes the proof. □

Proof of Theorem 4.1. Our estimator is given by $\hat{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta}} \mathbb{P}_n \mathcal{L}(\boldsymbol{\theta}) + \tau_n\|\boldsymbol{\beta}\|_1$. We denote the marginal mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ as $\mu_i := \mathbb{E}_\varepsilon \int_{\Omega_i} \exp[\alpha(s) + X_i\boldsymbol{\beta} + \varepsilon(s)]\mathrm{d}s$. Define the empirical process term

$$\nu_n(\boldsymbol{\theta}) = (\mathbb{P}_n - \mathbb{P})\mathcal{L}(\boldsymbol{\theta})$$

and the excess risk

$$\mathcal{E}(\boldsymbol{\theta}) = \mathbb{P}(\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*)).$$

Similar to the logic of Haris et al. (2019b), we examine

$$\begin{aligned}\nu_n(\boldsymbol{\theta}) - \nu_n(\boldsymbol{\theta}^*) &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_i)\left[(\alpha_i - \alpha_i^*) + \sum_{j=1}^{p}\left(\beta_j x_{ij} - \beta_j^* x_{ij}\right)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_i)(\alpha_i - \alpha_i^*) + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}(Y_i - \mu_i)(\beta_j x_{ij} - \beta_j^* x_{ij}) := \mathrm{I} + \mathrm{II}, \qquad \text{(C.6)}\end{aligned}$$

and analyze the two terms separately. First, define $a_i = \frac{\alpha_i - \alpha_i^*}{\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*\|_1}$; then for any $\rho > 0$, term I satisfies

$$\Pr\left(\left|\frac{\sum_i (Y_i - \mu_i)(\alpha_i - \alpha_i^*)}{n\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*\|_1}\right| > \rho\right) \leq \frac{\mathbb{E}_0|\sum_i a_i(Y_i - \mu_i)|^2}{\rho^2 n^2} \leq \frac{2\sum_i a_i^2 \mathrm{Var}(Y_i)}{\rho^2 n^2} \leq \frac{2(\sum_i a_i^2)\max_i \mathrm{Var}(Y_i)}{\rho^2 n^2} \qquad \text{(C.7)}$$

where the first "$\leq$" holds by Chebyshev's inequality. Furthermore, by the law of total variance, we have

$$\mathrm{Var}(Y_i) = \mathbb{E}_\varepsilon \mathrm{Var}(Y_i \mid \varepsilon) + \mathrm{Var}_\varepsilon \mathbb{E}[Y_i \mid \varepsilon] = \mathbb{E}_0 \mu_i + \mathbb{E}_0 \mu_i^2,$$

which satisfies

$$\begin{aligned}\mathbb{E}_0 \mu_i^k &= \mathbb{E}_0 \int_{\Omega_i} \exp\left[k(\alpha(s) + X_i\boldsymbol{\beta} + \varepsilon(s))\right] \mathrm{d}s \\ &= \int_{\Omega_i} \exp\left[k(\alpha(s) + X_i\boldsymbol{\beta}\right] \mathbb{E}_0 \exp[k\varepsilon(s)]] \, \mathrm{d}s\end{aligned}$$

where $k = 1, 2$. By Assumption 4.4 along with the fact that $|\Omega_i|$ is bounded, we have $\max_i \mathbb{E}_0 \mu_i^k <$ (some) $M < \infty$. Furthermore, it holds that $\sum_i a_i^2 = \|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*\|_2^2 / \|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*\|_1^2 \leq 1$ by construction.

Returning to (C.7), we now established

$$\Pr\left(\left|\frac{\sum_i (Y_i - \mu_i)(\alpha_i - \alpha_i^*)}{n\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*\|_1}\right| > \rho\right) \leq C_{\alpha,\beta,\varepsilon}(n\rho)^{-2} \qquad \text{(C.8)}$$

for some constant $C_{\alpha,\beta,\varepsilon}$. Likewise, for term II, let

$$b_{ij} = \frac{\beta_j x_{ij} - \beta_j^* x_{ij}}{n|\beta_j - \beta_j^*|}.$$

It then follows that

$$\|b_{\cdot j}\|_2^2 = \frac{\sum_i x_{ij}^2}{n^2} \leq \frac{R^2}{n},$$

and

$$\|b_{\cdot j}\|_\infty \leq \frac{R}{n}$$

by Assumption 4.6. Therefore, by a similar argument as for term I, for any $\rho > 0$,

$$\Pr\left(\left|\frac{\sum_i (Y_i - \mu_i)(\beta_j x_{ij} - \beta_i^* x_{ij})}{n|\beta_j - \beta_j^*|}\right| > \rho\right) \leq \frac{|\sum_i b_{ij} \mathrm{Var}(Y_i)|}{\rho^2 n^2} \leq \frac{2\sum b_{ij}^2 \mathrm{Var}(Y_i)}{\rho^2 n^2} \leq C_{\alpha,\beta,\varepsilon} R^2 n^{-3} \rho^{-2}.$$

Applying a union bound yields

$$\Pr\left(\max_{j\in[p]} \left|\frac{\sum_i (Y_i - \mu_i)(\beta_j x_{ij} - \beta_i^* x_{ij})}{n|\beta_j - \beta_j^*|}\right| > \rho\right) \leq C_{\alpha,\beta,\varepsilon} R^2 (n\rho)^{-2}. \tag{C.9}$$

Plugging into (C.6) yields that with probability $\geq 1 - C_1(n\rho)^{-2}$ for some constant C_1,

$$|\nu_n(\boldsymbol{\theta}) - \nu_n(\boldsymbol{\theta}^*)| \leq \rho\left[\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*\|_1 + \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1\right]. \tag{C.10}$$

Next, note that

$$\lambda_{\min}\left[\nabla^2_{\boldsymbol{\theta}} \mathcal{L}\right] \geq \lambda_{\min}\left\{\begin{bmatrix} I_n \\ \boldsymbol{X}^\top \end{bmatrix} \boldsymbol{D} \begin{bmatrix} I_n & \boldsymbol{X} \end{bmatrix}\right\}$$

where $\boldsymbol{D} = diag(\exp(\alpha_i + X_i\boldsymbol{\beta}))$. Note also that the eigenvalues of $\nabla^2_{\boldsymbol{\theta}}\mathcal{L}$ are determined by those of $\boldsymbol{D}$ and $\boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{X}$, the restricted eigenvalue condition in Assumption 4.6 along with the (lower-)bounded intensity condition in Assumption 4.4 guarantee that

$$\lambda_{\min}\left[\nabla^2_{\boldsymbol{\theta}} \mathcal{L}\right] \geq (some)\ m > 0$$

for all $\boldsymbol{\theta} \in \mathcal{B}$ as defined in Assumption 4.6 (we recall $\mathcal{B} = \{\theta : \|\theta - \theta^*\|_1 \leq 4\tau_n^2 s / c\rho\varphi_s^2\}$), establishing the restricted strong convexity of $\mathcal{L}$.

For convenience, let $M^* = \frac{4\tau_n^2 s}{c\rho\varphi_s^2}$. Define

$$Z_{M^*} := \sup_{\boldsymbol{\theta}\in\mathcal{B}} |\nu_n(\theta) - \nu_n(\theta^*)|.$$

We have just shown that

$$Z_{M^*} \leq \rho\left[\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*\|_1 + \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1\right] \leq \rho M^*$$

with probability $\geq 1 - C_1(n\rho)^{-2}$.

Similar to the approach of Haris et al. (2019b), set

$$t = \frac{M^*}{M^* + \|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1}$$

and let $\breve{\boldsymbol{\theta}} = t\hat{\boldsymbol{\theta}} + (1-t)\boldsymbol{\theta}^*$. Then

$$\|\breve{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 = t\|\hat{\boldsymbol{\theta}}\|_1 \leq \frac{M^*}{\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1}\|\hat{\boldsymbol{\theta}}\|_1 = M^*$$

so that $\breve{\boldsymbol{\theta}} \in \mathcal{B}$. By the basic inequalities due to the optimality of $\hat{\boldsymbol{\theta}}$ and the convexity of $\mathcal{L}$,

$$\begin{aligned}
&\mathcal{E}(\hat{\boldsymbol{\theta}}) + \tau_n\|\hat{\boldsymbol{\theta}}\|_1 \leq -[\nu_n(\hat{\boldsymbol{\theta}}) - \nu_n(\boldsymbol{\theta}^*)] + \tau_n\|\boldsymbol{\theta}^*\|_1, \\
&\mathcal{E}(\breve{\boldsymbol{\theta}}) + \tau_n\|\breve{\boldsymbol{\theta}}\|_1 \leq -[\nu_n(\breve{\boldsymbol{\theta}}) - \nu_n(\boldsymbol{\theta}^*)] + \tau_n\|\boldsymbol{\theta}^*\|_1 \leq Z_{M^*} + \lambda\|\boldsymbol{\theta}^*\|_1 \leq \rho M^* + \lambda\|\boldsymbol{\theta}^*\|_1. \qquad \text{(C.11)}
\end{aligned}$$

Further by the separability of $\|\cdot\|_1$, we have

$$\tau_n\|\boldsymbol{\beta}^*\|_1 \leq \tau_n[\|\boldsymbol{\beta}^*_S - \breve{\boldsymbol{\beta}}_S\|_1 + \|\breve{\boldsymbol{\beta}}_S\|_1],$$

and

$$\tau_n\|\breve{\boldsymbol{\beta}}\|_1 = \tau_n[\|\breve{\boldsymbol{\beta}}_S\|_1 + \|(\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{S^C}\|_1]$$

due to the sparsity of $\boldsymbol{\beta}^*$. Plugging into (C.11), we obtain

$$\mathcal{E}(\breve{\boldsymbol{\theta}}) + \tau_n\left[\|\breve{\boldsymbol{\beta}}_S\|_1 + \|(\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{S^C}\|_1\right] \leq \rho M^* + \tau_n\left[\|\boldsymbol{\beta}^*_S - \breve{\boldsymbol{\beta}}_S\|_1 + \|\breve{\boldsymbol{\beta}}_S\|_1\right].$$

Adding $\tau_n\left[\|\breve{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 + \|(\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_S\|_1\right]$ to both sides,

$$\mathcal{E}(\breve{\boldsymbol{\theta}}) + \tau_n\left[\|\breve{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 + \|\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1\right] \leq \rho M^* + \tau_n\|\breve{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 + 2\tau_n\|(\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_S\|_1. \qquad \text{(C.12)}$$

From here we consider two scenarios for (C.12). In both cases, we set $\rho = O(\sqrt{(\log p)/n})$ and $\tau_n \asymp \rho$ such that $\tau_n \geq 8\rho$. When $\rho M^* \leq \tau_n\|\breve{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 + 2\tau_n\|(\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_S\|_1$, (C.12) becomes

$$\begin{aligned}
&\mathcal{E}(\breve{\boldsymbol{\theta}}) + \tau_n\|(\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{S^C}\|_1 \leq \tau_n\|\breve{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 + 3\tau_n\|(\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_S\|_1 \\
\Rightarrow &\|(\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{S^C}\|_1 \leq \|\breve{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 + 3\|(\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_S\|_1.
\end{aligned}$$

Comparing with Assumption 4.3, we see that $\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \in \mathcal{C}(S)$. Under the compatibility condition,

$$\begin{aligned}
\mathcal{E}(\breve{\boldsymbol{\theta}}) + \tau_n\left[\|\breve{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 + \|\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1\right] &\leq \frac{4\tau_n\|\breve{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2\sqrt{s}}{\varphi_s} \\
&\leq \frac{16\tau_n^2 s}{4c\varphi_s^2} + c\|\breve{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2 \text{ for } c = \frac{4\tau_n\sqrt{s}}{\varphi_s} \\
&\leq \frac{4\tau_n^2 s}{c\varphi_s^2} + \mathcal{E}(\breve{\boldsymbol{\theta}}) \qquad \text{(C.13)}
\end{aligned}$$

where the second line follows from a convex conjugate argument, namely, letting $H(v) = \sup_u\{uv - cu^2\}$ for $v \geq 0$ and some constant c, then $H(v) = v^2/4c$. The third line is implied by the restricted strong convexity of $\mathcal{L}$ since $\breve{\boldsymbol{\theta}} \in \mathcal{B}$.

Continuing from (C.13),

$$\mathcal{E}(\breve{\boldsymbol{\theta}}) + \tau_n \left[\|\breve{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 + \|\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1\right] \leq \rho M^* + \mathcal{E}(\breve{\boldsymbol{\theta}}) \leq \frac{\tau_n M^*}{8} + \mathcal{E}(\breve{\boldsymbol{\theta}}), \tag{C.14}$$

leading to $\|\breve{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \leq M^*/8$. Then, by construction of t,

$$\begin{aligned}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 &= \frac{1}{t}[\|\breve{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1] \\ &\leq \left[1 + \frac{\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1}{M^*}\right] \frac{M^*}{8} \leq \frac{1}{8}[M^* + \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1].\end{aligned}$$

we thus have $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \leq M^*$ so that $\hat{\boldsymbol{\theta}} \in \mathcal{B}$ as well.

We then return to (C.12) to examine the second scenario where $\rho M^* > \tau_n\|\breve{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 + 2\tau_n\|(\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_S\|_1$. In this case, (C.12) simply becomes

$$\mathcal{E}(\breve{\boldsymbol{\theta}}) + \tau_n \left[\|\breve{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 + \|\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1\right] \leq 2\rho M^* \leq \frac{\tau_n M^*}{4}$$

which leads to $\|\breve{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \leq M^*/4$. Similar to the first scenario, we obtain $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \leq M^*$ so that $\hat{\boldsymbol{\theta}} \in \mathcal{B}$. Namely,

$$\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq M^*.$$

holds from (C.12) in both cases. Consequently, we can apply all claims from (C.11) onward involving $\breve{\theta}$ to $\hat{\theta}$. In particular, we have

$$\mathcal{E}(\hat{\boldsymbol{\theta}}) + \tau_n \left[\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1\right] \leq \mathcal{E}(\hat{\boldsymbol{\theta}}) + \rho M^*$$

from (C.14) in scenario one and

$$\mathcal{E}(\hat{\boldsymbol{\theta}}) + \tau_n \left[\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1\right] \leq \mathcal{E}(\hat{\boldsymbol{\theta}}) + 2\rho M^*$$

in scenario two. Thus, it must hold that

$$\tau_n \left[\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1\right] \leq 2\rho M^* \leq \frac{8\tau_n^2 s}{c\varphi_s^2},$$

establishing that

$$\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \frac{8s}{c\varphi_s^2}\tau_n \asymp \sqrt{\frac{\log p}{n}}$$

with probability $\geq 1 - C_1(n\rho)^{-2}$. □

C.3.2 Inference

Proof of Theorem 4.2. Recall that the de-biased estimator is defined as

$$\hat{\boldsymbol{b}} = \hat{\boldsymbol{\beta}} + \frac{1}{n} M X^\top \left[\boldsymbol{Y} - \boldsymbol{B} \odot \exp\left(\hat{\boldsymbol{\alpha}} + \boldsymbol{X}\hat{\boldsymbol{\beta}} \right) \right].$$

We could decompose

$$\begin{aligned}
\hat{\boldsymbol{b}} - \boldsymbol{\beta}^0 &= \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 + \frac{1}{n} M \nabla_{\boldsymbol{\beta}} \ell(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) \\
&= \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 + \frac{1}{n} M \nabla_{\boldsymbol{\beta}} \ell(\hat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^0) + \frac{1}{n} M \nabla^2_{\boldsymbol{\beta}} \ell(\hat{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}})(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}) \\
&= \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 + \frac{1}{n} M \nabla_{\boldsymbol{\beta}} \ell(\hat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^0) - \frac{1}{n} M \nabla^2_{\boldsymbol{\beta}} \ell(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) + \frac{1}{n} M \left[\nabla^2_{\boldsymbol{\beta}} \ell(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) - \nabla^2_{\boldsymbol{\beta}} \ell(\hat{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}) \right] (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \\
&= -(M\hat{\boldsymbol{H}} - I)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) + \frac{1}{n} M \nabla_{\boldsymbol{\beta}} \ell(\hat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^0) + \frac{1}{n} M \left[\nabla^2_{\boldsymbol{\beta}} \ell(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) - \nabla^2_{\boldsymbol{\beta}} \ell(\hat{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}) \right] (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)
\end{aligned} \tag{C.15}$$

by a Taylor expansion of $\ell(\boldsymbol{\alpha}, \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, where $\tilde{\boldsymbol{\beta}}$ is between $\boldsymbol{\beta}^0$ and $\hat{\boldsymbol{\beta}}$.

We analyze each term in (C.15) individually. For the first term, we have

$$\begin{aligned}
\left\| (M\hat{\boldsymbol{H}} - I)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \right\|_\infty &= \max_j \left| (\hat{\boldsymbol{H}} m_j^\top - e_j)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \right| \\
&\le \max_j \left\| \hat{\boldsymbol{H}} m_j^\top - e_j \right\|_\infty \left\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \right\|_1 \\
&= o\left(\frac{1}{\sqrt{s \log p}} \right) \cdot O_P\left(\sqrt{\frac{s \log p}{n}} \right) = o_P\left(n^{-1/2} \right)
\end{aligned} \tag{C.16}$$

by construction of M, Assumption i) under Theorem 4.2 along with the conclusion of Theorem 4.1.

Next, the third term in (C.15) can be bounded as

$$\begin{aligned}
&\frac{1}{n} \left\| M \left[\nabla^2_{\boldsymbol{\beta}} \ell(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) - \nabla^2_{\boldsymbol{\beta}} \ell(\hat{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}) \right] (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \right\|_\infty \\
&\le \left\| \left[\frac{1}{n} M \nabla^2_{\boldsymbol{\beta}} \ell(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) - I \right] (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \right\|_\infty + \left\| \left[\frac{1}{n} M \nabla^2_{\boldsymbol{\beta}} \ell(\hat{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}) - I \right] (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \right\|_\infty,
\end{aligned} \tag{C.17}$$

where the first term is $o_P(n^{-1/2})$ as in (C.16). Also, by Assumption ii-a) in Theorem 4.2, since $\tilde{\boldsymbol{\beta}}$ is between $\boldsymbol{\beta}^0$ and $\hat{\boldsymbol{\beta}}$ between which the gap is shrinking towards 0, we have that $(0, \tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) \in \mathcal{N}(\delta_{\boldsymbol{\alpha}}, \delta_{\boldsymbol{\beta}})$ for large enough n and p, and consequently

$$\left\| \left[\frac{1}{n} M \nabla^2_{\boldsymbol{\beta}} \ell(\hat{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}) - I \right] (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \right\|_\infty \le \max_j \left\| \frac{1}{n} \nabla^2_{\boldsymbol{\beta}} \ell(\hat{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}) m_j^\top - e_j \right\|_\infty \left\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \right\|_1 = o_P(n^{-1/2}).$$

Thus we have showed that the third term in (C.15) is $o_P(n^{-1/2})$.

We finally analyze the second term in (C.15), which can be rewritten as

$$\begin{aligned}\frac{1}{n}M\nabla_{\boldsymbol{\beta}}\ell(\hat{\boldsymbol{\alpha}},\boldsymbol{\beta}^0) &= \frac{1}{n}M\nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\alpha}^\dagger,\boldsymbol{\beta}^0) + \frac{1}{n}M\left[\nabla_{\boldsymbol{\beta}}\ell(\hat{\boldsymbol{\alpha}},\boldsymbol{\beta}^0) - \nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\alpha}^*,\boldsymbol{\beta}^0)\right] + \frac{1}{n}M\left[\nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\alpha}^*,\boldsymbol{\beta}^0) - \nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\alpha}^\dagger,\boldsymbol{\beta}^0)\right] \\ &= \frac{1}{n}M\nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\alpha}^\dagger,\boldsymbol{\beta}^0) + \frac{1}{n}M\left[\nabla_{\boldsymbol{\beta}}\ell(\hat{\boldsymbol{\alpha}},\boldsymbol{\beta}^0) - \nabla_{\boldsymbol{\beta}}\ell(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})\right] + \frac{1}{n}M\left[\nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\alpha}^*,\hat{\boldsymbol{\beta}}) - \nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\alpha}^*,\boldsymbol{\beta}^0)\right] \\ &\quad + \frac{1}{n}M\left[\nabla_{\boldsymbol{\beta}}\ell(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}}) - \nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\alpha}^*,\hat{\boldsymbol{\beta}})\right] + \frac{1}{n}M\left[\nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\alpha}^*,\boldsymbol{\beta}^0) - \nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\alpha}^\dagger,\boldsymbol{\beta}^0)\right].\end{aligned} \tag{C.18}$$

Noting that $(0, \boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}) \in \mathcal{N}(\delta_{\boldsymbol{\alpha}}, \delta_{\boldsymbol{\beta}})$ and $(\boldsymbol{\alpha}^* - \hat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}) \in \mathcal{N}(\delta_{\boldsymbol{\alpha}}, \delta_{\boldsymbol{\beta}})$ for large enough n and p, by Assumption ii-a), we have for the second and third terms in (C.18) that

$$\begin{aligned}&\frac{1}{n}\left\|M\left[\nabla_{\boldsymbol{\beta}}\ell(\hat{\boldsymbol{\alpha}},\boldsymbol{\beta}^0) - \nabla_{\boldsymbol{\beta}}\ell(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})\right] + M\left[\nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\alpha}^*,\hat{\boldsymbol{\beta}}) - \nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\alpha}^*,\boldsymbol{\beta}^0)\right]\right\|_\infty \\ &\le \left\|\left[\frac{1}{n}M\nabla^2_{\boldsymbol{\beta}}\ell(\hat{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}}) - I\right](\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\right\|_\infty + \left\|\left[\frac{1}{n}M\nabla^2_{\boldsymbol{\beta}}\ell(\boldsymbol{\alpha}^*,\breve{\boldsymbol{\beta}}) - I\right](\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\right\|_\infty \\ &\le \max_j \frac{1}{n}\left\|\nabla^2_{\boldsymbol{\beta}}\ell(\hat{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}})m_j^\top - e_j\right\|_\infty \cdot \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 + \max_{j'} \frac{1}{n}\left\|\nabla^2_{\boldsymbol{\beta}}\ell(\boldsymbol{\alpha}^*,\breve{\boldsymbol{\beta}})m_{j'}^\top - e_j\right\|_\infty \cdot \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 \\ &= o\left(\frac{1}{\sqrt{s\log p}}\right) O_P\left(\sqrt{\frac{s\log p}{n}}\right) = o_P(n^{-1/2}).\end{aligned} \tag{C.19}$$

where $\tilde{\boldsymbol{\beta}}, \breve{\boldsymbol{\beta}}$ are both between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^0$.

Also, by Assumption ii-b), it holds for some $\breve{\alpha}$ between $\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}^*$ that

$$\begin{aligned}\frac{1}{n}\left\|M\left[\nabla_{\boldsymbol{\beta}}\ell(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}}) - \nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\alpha}^*,\hat{\boldsymbol{\beta}})\right]\right\|_\infty &= \frac{1}{n}\left\|M\nabla^2_{\boldsymbol{\beta},\boldsymbol{\alpha}}\ell(\breve{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\right\|_\infty \\ &\le \frac{1}{n}\left\|M\nabla^2_{\boldsymbol{\beta},\boldsymbol{\alpha}}\ell(\breve{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})\right\|_\infty \cdot \|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_\infty = o_P(n^{-1/2})\end{aligned} \tag{C.20}$$

and for some $\boldsymbol{\alpha}^\ddagger$ between $\boldsymbol{\alpha}^*$ and $\boldsymbol{\alpha}^\dagger$,

$$\begin{aligned}\frac{1}{n}\left\|M\left[\nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\alpha}^*,\boldsymbol{\beta}^0) - \nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\alpha}^\dagger,\boldsymbol{\beta}^0)\right]\right\|_\infty &= \frac{1}{n}\left\|M\nabla^2_{\boldsymbol{\beta},\boldsymbol{\alpha}}\ell(\boldsymbol{\alpha}^\ddagger,\boldsymbol{\beta}^0)(\boldsymbol{\alpha}^* - \boldsymbol{\alpha}^\dagger)\right\|_\infty \\ &\le \frac{1}{n}\left\|M\nabla^2_{\boldsymbol{\beta},\boldsymbol{\alpha}}\ell(\boldsymbol{\alpha}^\ddagger,\boldsymbol{\beta}^0)\right\|_\infty \cdot \left\|\boldsymbol{\alpha}^* - \boldsymbol{\alpha}^\dagger\right\|_\infty = o_P(n^{-1/2}).\end{aligned} \tag{C.21}$$

Combining (C.18)-(C.21) leads to

$$\frac{1}{n}M\nabla_{\boldsymbol{\beta}}\ell(\hat{\boldsymbol{\alpha}},\boldsymbol{\beta}^0) = \frac{1}{n}M\nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\alpha}^\dagger,\boldsymbol{\beta}^0) + o_P(n^{-1/2}). \tag{C.22}$$

Plugging (C.16), (C.17) and (C.22) into (C.15) yields

$$\hat{\boldsymbol{b}} - \boldsymbol{\beta}^0 = \frac{1}{n} M \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\alpha}^\dagger, \boldsymbol{\beta}^0) + o_P(n^{-1/2}),$$

and the asymptotic linearity of $\hat{\boldsymbol{b}}$ therefore establishes for each $j = 1, \ldots, p$ that

$$\frac{\sqrt{n}(\hat{b}_j - \beta_j)}{\left[M \mathbb{E}_0 \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\alpha}^\dagger, \boldsymbol{\beta}^0) \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\alpha}^\dagger, \boldsymbol{\beta}^0)^\top M^\top\right]_{jj}} \xrightarrow{d} N(0,1).$$

To see why the covariance estimator $\hat{\boldsymbol{\Sigma}}$ defined in (4.7) is a valid conservative estimate for $\mathbb{E}_0 \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\alpha}^\dagger, \boldsymbol{\beta}^0) \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\alpha}^\dagger, \boldsymbol{\beta}^0)^\top$, note that

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n X_i^\top X_i \mathrm{Var}(Y_i \mid X_i) &= \frac{1}{n} \sum_{i=1}^n X_i^\top X_i \mathbb{E}_{\varepsilon_i^*} \mathbb{E}_{Y_i \mid \varepsilon_i^*} (Y_i - |\Omega_i| P_i \exp(\tilde{\alpha}_i + X_i \boldsymbol{\beta}^0 + \varepsilon_i^*))^2 \\
&\preceq \frac{2}{n} \sum_{i=1}^n X_i^\top X_i \Big[\Big(Y_i - |\Omega_i| P_i \exp(\hat{\alpha}_i + X_i \hat{\boldsymbol{\beta}}) \Big)^2 \\
&\qquad + \mathbb{E}_{\varepsilon_i^*} \Big(|\Omega_i| P_i \exp(\hat{\alpha}_i + X_i \hat{\boldsymbol{\beta}}) - |\Omega_i| P_i \exp(\tilde{\alpha}_i + X_i \boldsymbol{\beta}^0 + \varepsilon_i^*) \Big)^2 \Big]
\end{aligned}$$

where $\varepsilon_i^* = \varepsilon(s_i)$ for the location s_i defined in Lemma 4.1, and we recall that $\alpha_i^\dagger = \tilde{\alpha}_i + \log \mathbb{E}_0[\exp(\varepsilon_i^*)]$.

When the error random field is stationary, independent and Gaussian with variance $\hat{\sigma}^2$, we could alternatively adopt the estimator $\tilde{\boldsymbol{\Sigma}}$ defined in (4.9). By the law of total variance,

$$\begin{aligned}
\mathrm{Cov}(Y) &= \mathbb{E}_0[\mathrm{Cov}(Y \mid \varepsilon)] + \mathrm{Cov}\left(\mathbb{E}_0(Y \mid \varepsilon)\right) \\
&= \mathrm{diag}\left[|\Omega_i| P_i \exp(\boldsymbol{\alpha}_i^\dagger + X_i \boldsymbol{\beta}^0) + (|\Omega_i| P_i)^2 \exp(2\alpha^0(s_i^*) + 2 X_i \boldsymbol{\beta}^0) \mathrm{Var}(\exp \varepsilon(s_i^*)) \right] \\
&= \mathrm{diag}\left[|\Omega_i| P_i \exp(\boldsymbol{\alpha}_i^\dagger + X_i \boldsymbol{\beta}^0) + (|\Omega_i| P_i)^2 \exp(2\alpha^0(s_i^*) + 2 X_i \boldsymbol{\beta}^0) [\exp(\sigma^2) - 1] \exp(\sigma^2)) \right] \\
&= \mathrm{diag}\left[|\Omega_i| P_i \exp(\boldsymbol{\alpha}_i^\dagger + X_i \boldsymbol{\beta}^0) + (|\Omega_i| P_i)^2 \exp(2\boldsymbol{\alpha}^\dagger + 2 X_i \boldsymbol{\beta}^0) [\exp(\sigma^2) - 1]) \right]
\end{aligned}$$

which leads to $\tilde{\boldsymbol{\Sigma}}$ as a plug-in estimator.

□

Model	Method	Model Specification	Theoretical Guarantees
Cox process	Minimal contrast estimation (Diggle, 2003; Møller and Waagepetersen, 2003)	Parametric	N/A
	Bayesian estimation	Parametric	Convergence of LGCP posteriors under discretization (Waagepetersen, 2004)
	Bayesian estimation with INLA (Rue et al., 2009)	Parametric	N/A
	Bayesian estimation with variational approximation	Parametric	Convergence of posteriors to KL minimizer of a normal distribution (Wang and Blei, 2019)
	Bayesian estimation with basis function approximation of the random field (Simpson et al., 2016)	Random field with a basis expansion form	Convergence of basis function and discrete approximation, but not the full posterior
	Poisson maximum likelihood estimation (Schoenberg, 2005)	Known form for marginal means of the intensity	Consistency of parameters
	Composite likelihood estimation (Guan, 2006)	Stationarity or known form of second order intensity	Consistency and asymptotic normality of parameters
	Covariate-based kernel smoothing (Guan, 2008)	Nonparametric	Consistency
	Two-step estimation (Waagepetersen and Guan, 2009)	Known form of first- and second-order intensity functions	Consistency and asymptotic normality of parameters
Areal data model, e.g. Besag-York-Mollié (BYM) or BYM2 (Diggle, 2003; Møller and Waagepetersen, 2003; Riebler et al., 2016)	Bayesian estimation	Spatially correlated errors induced by connectivity, and independent non-spatial heterogeneity	N/A

Table C.1: Comparison of existing models and estimation methods for doubly-stochastic spatial processes.

Appendix D

SUPPLEMENTARY MATERIALS FOR CHAPTER 5

D.1 Proofs

In this section, we provide proofs for Theorems 5.1 and 5.2. We begin by introducing some notations.

First, recall that our solution to the optimization problem (5.6) is defined as $(\hat{f}_1, \ldots, \hat{f}_K, \hat{\beta})$ for which we adopt the shorthand notation $(\hat{f}, \hat{\beta})$ (when it does not cause confusion), and that under the same shorthand notation, the true functions and parameters are (f^*, β^*). Also, we let $\tilde{f}_1, \ldots, \tilde{f}_K \in \mathcal{F}$ be such that $\tilde{f}_k(x_{ik}) - f_k^*(x_{ik}) = 0$ for $i = 1, \ldots, n$ and $k = 1, \ldots, K$ and define the *target causal parameter* $\tilde{\beta}$ as.

$$\tilde{\beta} := \arg\min_{\beta} \mathbb{P}_0 \ell \left(y,\ \sum_{k=1}^{K} \tilde{f}_k(x_k) + w\beta_0 + \sum_{s=1}^{n-1} \frac{M_s}{N_s} \beta_s \right)$$

i.e. the minimizer of the population risk with respect to β, with each nuisance function f_k fixed at $\tilde{f}_k$. For simplicity, we also introduce the shorthand expression for the loss function, $L(f, \beta) := \ell(y, \sum_k f_k(x_k) + w\beta_0 + \sum_s \frac{M_s}{N_s}\beta_s)$. We also recall that the *empirical process term* is

$$\nu_n(f, \beta) := (\mathbb{P}_n - \mathbb{P}_0) L(f, \beta)$$

and the *excess risk* is

$$\mathcal{E}(f, \beta) := \mathbb{P}_0[L(f, \beta) - L(\tilde{f}, \tilde{\beta})].$$

D.1.1 Consistency of $\hat{\beta}$

It is straightforward to verify that the target causal parameter $\tilde{\beta}$ coincides with the true causal parameter β^*, stated in the following Lemma.

Lemma D.1 (Target versus true causal parameter)**.** *If the loss function $\ell(\cdot)$ (and hence $L(\cdot)$) is convex, then $\tilde{\beta} = \beta^*$.*

Proof. Comparing the β-optimality conditions

$$0 = \nabla_\beta \mathbb{E}[L(\tilde{f}, \tilde{\beta})]$$
$$0 = \nabla_\beta \mathbb{E}[L(f^*, \beta^*)]$$

and noting that $\tilde{f}(x_{ik}) = f^*(x_{ik})$ for all i and k, the convexity of $L(\cdot)$ immediately leads to the conclusion. □

We follow a similar outline as in Haris et al. (2019a) to prove Theorem 5.1. First, from a similar logic as Haris et al. (2019a), we have the following *basic inequality*:

$$\mathcal{E}(\hat{f}, \hat{\beta}) + \lambda_\beta R_\beta(\hat{\beta}) + \sum_k \lambda_k R_{\mathcal{F}}(\hat{f}_k) \leq -\left[\nu_n(\hat{f}, \hat{\beta}) - \nu_n(\tilde{f}, \tilde{\beta})\right] + \mathcal{E}(\tilde{f}, \tilde{\beta}) + \lambda_\beta R_\beta(\tilde{\beta}) + \sum_k \lambda_k R_{\mathcal{F}}(\tilde{f}_k). \tag{D.1}$$

Also, given the convexity of $L(\cdot)$, R_β and $R_{\mathcal{F}}$, for any $t \in (0,1)$ and $(f^\dagger, \beta^\dagger) := (t\hat{f} + (1-t)\tilde{f}, t\hat{\beta} + (1-t)\tilde{\beta})$, we also have

$$\mathcal{E}(f^\dagger, \beta^\dagger) + \lambda_\beta R_\beta(\beta^\dagger) + \sum_k \lambda_k R_{\mathcal{F}}(f_k^\dagger) \leq -\left[\nu_n(f^\dagger, \beta^\dagger) - \nu_n(\tilde{f}, \tilde{\beta})\right] + \mathcal{E}(\tilde{f}, \tilde{\beta}) + \lambda_\beta R_\beta(\tilde{\beta}) + \sum_k \lambda_k R_{\mathcal{F}}(\tilde{f}_k). \tag{D.2}$$

Lemma D.2 below provides a bound on the empirical process term, which is needed for establishing the consistency of β.

Lemma D.2. *Under Assumptions 5.1,5.3-5.5, the following inequality holds*

$$\nu_n(f, \beta) - \nu_n(\tilde{f}, \tilde{\beta}) \leq \rho \left[\|\beta - \tilde{\beta}\|_1 + \lambda_{\mathcal{F}} \sum_k R_{\mathcal{F}}(f_k - \tilde{f}_k)\right]$$

with probability at least $1 - 2\exp(-C_1 n\rho^2) - C_2 \exp(-C_2' n\rho^2) - 2\exp(-C_3 n\rho^2)$ *for positive constants* C_1, C_2, C_2', C_3 *not depending on* n.

Proof. Under item 1) of Assumption 5.4, recall that the loss function $\ell(\cdot)$ takes the form $\ell(y, \eta) = \zeta \cdot y\eta + h(\eta)$. Denote $\mu_i := \mathbb{E}_{W,G_n}\mathbb{E}_{Y|W,G_n} Y_i$, $\mu_{W,G_n,i} := E_{Y|W,G_n}[Y_i \mid W, G_n]$ and

$Z := \left[w, \frac{M_1}{N_1}, \ldots, \frac{M_{n-1}}{N_{n-1}}\right]$, we can then write

$$\begin{aligned}\nu_n(f,\beta) - \nu_n(\tilde{f},\tilde{\beta}) = &\underbrace{\frac{\zeta}{n}\sum_i \left[(Y_i - \mu_i)\sum_k (f_k - \tilde{f}_k)(x_{ik})\right]}_{\text{I}} + \underbrace{\frac{\zeta}{n}\sum_i \left[(Y_i Z_i^\top - \mathbb{E}[Y_i Z_i^\top])(\beta - \tilde{\beta})\right]}_{\text{II}} \\ &+ \underbrace{\frac{1}{n}\sum_i \left[h(\sum_k f_k(x_{ik}) + Z_i^\top \beta) - h(\sum_k \tilde{f}_k(x_{ik}) + Z_i^\top \tilde{\beta})\right]}_{\text{III}} \\ &- \underbrace{\frac{1}{n}\sum_i \mathbb{E}\left[h(\sum_k f_k(x_{ik}) + Z_i^\top \beta) - h(\sum_k \tilde{f}_k(x_{ik}) + Z_i^\top \tilde{\beta})\right]}_{\text{IV}}. \end{aligned} \tag{D.3}$$

We bound the four terms in (D.3) separately, starting with term II. For any $j = 0, \ldots, n-1$, letting $\rho = \kappa\sqrt{\frac{\log n}{n}}$ for some κ that will be chosen later, we have

$$\Pr\left(\left|\frac{\zeta \sum_i (Y_i Z_{ij} - \mathbb{E}[Y_i Z_{ij}])(\beta_j - \tilde{\beta}_j)}{n|\beta_j - \tilde{\beta}_j|}\right| > \frac{\rho}{2}\right) = \mathbb{E}_{W,G_n}\left[\Pr\left(\left|\frac{\zeta \sum_i (Y_i Z_{ij} - \mu_{W,G_n,i} Z_{ij})(\beta_j - \tilde{\beta}_j)}{n|\beta_j - \tilde{\beta}_j|}\right| > \frac{\rho}{2}\,\middle|\, Z\right)\right]. \tag{D.4}$$

Since the Y_i's are independent conditioning on Z_i, and $Y_i \mid Z_i$ are uniformly sub-Gaussian under Assumption 5.3, by Lemma 8.2 of van de Geer (2000), (D.4) continues as

$$\Pr\left(\left|\frac{\zeta \sum_i (Y_i Z_{ij} - \mathbb{E}[Y_i Z_{ij}])(\beta_j - \tilde{\beta}_j)}{n|\beta_j - \tilde{\beta}_j|}\right| > \frac{\rho}{2}\right) \leq 2\exp\left(-\frac{n\rho^2}{4 \cdot 8(K_0^2 + \sigma_0^2)\zeta^2}\right). \tag{D.5}$$

We apply a union bound across all $j = 0, \ldots, n-1$ and (D.5) then becomes

$$\begin{aligned}\Pr(\text{II} > \frac{\rho}{2}\|\beta - \tilde{\beta}\|_1) &\leq 2n\exp\left(-\frac{n\rho^2}{32(K_0^2 + \sigma_0^2)\zeta^2}\right) = 2\exp\left(-\frac{n\rho^2}{32(K_0^2 + \sigma_0^2)\zeta^2} + \log n\right) \\ &= 2\exp\left[-n\rho^2\left(\frac{1}{32(K_0^2 + \sigma_0^2)\zeta^2} - \frac{1}{\kappa}\right)\right] := 2\exp(-C_1 n\rho^2) \end{aligned} \tag{D.6}$$

and $C_1 > 0$ as long as $\kappa > 32(K_0^2 + \sigma_0^2\zeta^2)$.

We then analyze term I making use of the entropy bound specified in Assumption 5.5. Specifically, we note that if the entropy bound hold for a function class $\mathcal{F}$, then the same entropy bound also holds, up to a constant, for the class

$$\left\{\frac{f_k - \tilde{f}_k}{\lambda_{\mathcal{F}} \sum_k R_{\mathcal{F}}(f_k - \tilde{f}_k)} :\ f_k \in \mathcal{F}\right\}$$

for all $k = 1, \ldots, K$.

Using the bound for Dudley's integral,

$$A_0^{1/2} T_n^{1/2} \int_0^Q \sqrt{\log\left(1 + \frac{1}{u}\right)} \mathrm{d}u \leq \tilde{A}_0 T_n^{1/2},$$

and denoting $n^{-1} \sup_{f_k \in \mathcal{F}} \sum_i (f_k(x_{ik}) - \tilde{f}_k(x_{ik}))^2 / \lambda_{\mathcal{F}} \sum_k R_{\mathcal{F}}(f_k - \tilde{f}_k) := Q^2$, by Corollary 8.3 of van de Geer (2000), we have for all $\delta > 2K\tilde{C}_2\tilde{A}_0\sqrt{T_n/n}$ that for all k,

$$\begin{aligned}
\Pr\left(\sup_{f_k \in \mathcal{F}} \left| \frac{\frac{\zeta}{n}\sum_i (Y_i - \mu_i)(f_k - \tilde{f}_k)(x_{ik})}{\lambda_{\mathcal{F}} \sum_k R_{\mathcal{F}}(f_k - \tilde{f})} \right| \geq \frac{\delta}{K}\right) &= \mathbb{E}_{W,G_n}\left[\Pr\left(\sup_{f_k \in \mathcal{F}} \left| \frac{\frac{\zeta}{n}\sum_i (Y_i - \mu_{W,G_n,i})(f_k - \tilde{f}_k)(x_{ik})}{\lambda_{\mathcal{F}} \sum_k R_{\mathcal{F}}(f_k - \tilde{f}_k)} \right| \geq \frac{\delta}{K} \middle| Z\right)\right] \\
&\leq \mathbb{E}_{W,G_n}\left[\tilde{C}_2 \exp\left(-\frac{n\delta^2}{4K^2\tilde{C}_2^2 Q^2}\right)\right] = \tilde{C}_2 \exp\left(-\frac{n\delta^2}{4K^2\tilde{C}_2^2 Q^2}\right).
\end{aligned} \tag{D.7}$$

We then apply a union bound across $k = 1, \ldots, K$ to (D.7) and set $\delta = \rho$, which yields

$$\begin{aligned}
\Pr(\mathrm{I} > \rho\lambda_{\mathcal{F}} \sum_k R_{\mathcal{F}}(f_k - \tilde{f}_k)) &\leq K \cdot \Pr\left(\sup_{f_k \in \mathcal{F}} \left| \frac{\frac{\zeta}{n}\sum_i (Y_i - \mu_i)(f_k - \tilde{f}_k)(x_{ik})}{\lambda_{\mathcal{F}} \sum_k R_{\mathcal{F}}(f_j - \tilde{f}_k)} \right| \geq \frac{\delta}{K}\right) \\
&\leq K\tilde{C}_2 \exp\left(-\frac{n\delta^2}{4K^2\tilde{C}_2^2 Q^2}\right) := C_2 \exp(-C_2' n\rho^2)
\end{aligned} \tag{D.8}$$

as long as $\kappa > 2K\tilde{C}_2\tilde{A}_0\sqrt{T_n}$.

For terms III and IV, define $\varphi_i(z) := h(\sum_k f_k(x_{ik}) + z^\top \beta) - h(\sum_k \tilde{f}_k(x_{ik}) + z^\top \tilde{\beta})$. Since $h(\cdot)$ has bounded first order derivative, i.e. $|h'(\eta)| < M$ under Assumption 5.4, we have that each $\varphi_i(\cdot)$ satisfies the bounded difference property, i.e. for any z, z', we have

$$|\varphi_i(z) - \varphi_i(z')| \leq M \sum_j |\beta_j - \tilde{\beta}_j| \cdot I\{z_j \neq z_j'\}.$$

We note that the random vectors $\{(z_{i0}, \ldots, z_{i(n-1)})\}_{i=1}^n$ are independent condition on the random graph G_n (i.e., when the only source of randomness is in the treatment assignment vector W). Hence by McDiarmid's inequality (McDiarmid, 1989) along with a union bound across all $i = 1, \ldots, n$, we

have

$$\begin{aligned}
\Pr\left(|\mathrm{III}+\mathrm{IV}|>\frac{\rho}{2}\|\beta-\tilde{\beta}\|_1\right) &= \mathbb{E}_{G_n}\left[\Pr\left(|\mathrm{III}+\mathrm{IV}|>\frac{\rho}{2}\|\beta-\tilde{\beta}\|_1 \;\middle|\; G_n\right)\right] \\
&\leq \mathbb{E}_{G_n}\left[\Pr\left(\frac{\left|\frac{1}{n}\sum_i \varphi_i(z_i)-\mathbb{E}\varphi_i(z_i)\right|}{\|\beta-\tilde{\beta}\|_1}>\frac{\rho}{2} \;\middle|\; G_n\right)\right] \\
&\leq n\cdot\mathbb{E}_{G_n}\left[\Pr\left(\max_i \frac{|\varphi_i(z_i)-\mathbb{E}\varphi_i(z_i)|}{\|\beta-\tilde{\beta}\|_1}>\frac{\rho}{2} \;\middle|\; G_n\right)\right] \\
&\leq 2n\exp\left(-\frac{2\rho^2\|\beta-\tilde{\beta}\|_1^2}{4\|\beta-\tilde{\beta}\|_2^2}\right)\leq 2\exp\left(-\frac{n\rho^2}{2}+\log n\right) \\
&\leq 2\exp\left(-C_3 n\rho^2\right) \qquad \text{(D.9)}
\end{aligned}$$

where C_3 is positive as long as $\kappa^2>2$. Combining this and the requirement on κ from terms I and II, it suffices to set $\kappa>\max\{32(K_0^2+\sigma_0^2\zeta^2), 2K\tilde{C}_2\tilde{A}_0\sqrt{T_n}, \sqrt{2}\}$.

Finally, combining (D.6), (D.8) and (D.9), and applying a union bound again, we obtain

$$\begin{aligned}
&\Pr\left(\nu_n(f,\beta)-\nu_n(\tilde{f},\tilde{\beta})>\rho\left[\|\beta-\tilde{\beta}\|_1+\lambda_{\mathcal{F}}\sum_k R_{\mathcal{F}}(f_k-\tilde{f}_k)\right]\right) \\
&\leq 2\exp(-C_1 n\rho^2)+C_2\exp(-C_2' n\rho^2)+2\exp(-C_3 n\rho^2),
\end{aligned}$$

establishing the bound in Lemma D.2. □

Lemma D.2 implies that our solution $(\hat{f},\hat{\beta})$ falls into a specific set with probability approaching 1 as the sample size n grows. We are now ready to prove Theorem 5.1 based on such fact, by showing that our stated bound on the estimation error holds over this set.

Proof of Theorem 5.1. Let $M^* := \frac{4\lambda_{\tilde{\beta}}^2 m}{c\phi^2(m)\rho}+\frac{\lambda_\beta R(\tilde{\beta}_{S^C})}{\rho}$. By Lemma D.2, we have with high probability that

$$\begin{aligned}
Z_{M^*} :&= \sup_{\|\beta-\tilde{\beta}\|_1+\sum_k R_{\mathcal{F}}(f_k-\tilde{f}_k)\leq M^*} |\nu_n(f,\beta)-\nu_n(\tilde{f},\tilde{\beta})| \\
&\leq \sup_{\|\beta-\tilde{\beta}\|_1+\sum_k R_{\mathcal{F}}(f_k-\tilde{f}_k)\leq M^*} \rho\left[\|\beta-\tilde{\beta}\|_1+\sum_k R_{\mathcal{F}}(f_k-\tilde{f}_k)\right]\leq \rho M^*.
\end{aligned}$$

Set $t=\frac{M^*}{M^*+\|\hat{\beta}-\tilde{\beta}\|_1+\sum_k R_{\mathcal{F}}(\hat{f}_k-\tilde{f}_k)}$ and $(f^\dagger,\beta^\dagger):=(t\hat{f}+(1-t)\tilde{f}, t\hat{\beta}+(1-t)\tilde{\beta})$. Then we have by construction that

$$\|\beta^\dagger-\tilde{\beta}\|_1+\sum_k R_{\mathcal{F}}(f_k^\dagger-\tilde{f}_k)\leq M^*$$

and that

$$\|\beta - \tilde{\beta}\|_1 + \sum_k R_{\mathcal{F}}(f_k - \tilde{f}_k) = \frac{1}{t}\left[\|\beta^\dagger - \tilde{\beta}\|_1 + \sum_k R_{\mathcal{F}}(f_k^\dagger - \tilde{f}_k)\right]. \tag{D.10}$$

Let $S := \{0, \ldots, m\}$ for the index m defined in Assumption 5.2), and $S^C = \{m+1, \ldots, n-1\}$. We write $\beta_S = (\beta_0, \ldots, \beta_m)$ and β_{S^C} analogously. Recall the basic inequality (D.2),

$$\mathcal{E}(f^\dagger, \beta^\dagger) + \lambda_\beta R_\beta(\beta^\dagger) + \sum_k \lambda_k R_{\mathcal{F}}(f_k^\dagger) \leq -\left[\nu_n(f^\dagger, \beta^\dagger) - \nu_n(\tilde{f}, \tilde{\beta})\right] + \mathcal{E}(\tilde{f}, \tilde{\beta}) + \lambda_\beta R_\beta(\tilde{\beta}) + \sum_k \lambda_k R_{\mathcal{F}}(\tilde{f}_k),$$

and that $\mathcal{E}(\tilde{f}, \tilde{\beta}) = 0$. We have, for its left-hand side (LHS), by triangular inequality that

$$\begin{aligned}
\mathcal{E}(f^\dagger, \beta^\dagger) + \lambda_\beta R_\beta(\beta^\dagger) + \sum_k \lambda_k R_{\mathcal{F}}(f_k^\dagger) &\geq \lambda_\beta\left[R(\beta_S^\dagger) + R(\beta_{S^C}^\dagger)\right] + \sum_k \lambda_k\left[R_{\mathcal{F}}(f_k^\dagger - \tilde{f}_k) - R_{\mathcal{F}}(\tilde{f}_k)\right] \\
&\geq \lambda_\beta\left[R(\beta_S^\dagger) + R((\beta^\dagger - \tilde{\beta})_{S^C}) - R(\tilde{\beta}_{S^C})\right] + \sum_k \lambda_k\left[R_{\mathcal{F}}(f_k^\dagger - \tilde{f}_k) - R_{\mathcal{F}}(\tilde{f}_k)\right].
\end{aligned} \tag{D.11}$$

Meanwhile, for the right-hand side (RHS), we have with high probability that

$$-\left[\nu_n(f^\dagger, \beta^\dagger) - \nu_n(\tilde{f}, \tilde{\beta})\right] + \lambda_\beta R_\beta(\tilde{\beta}) + \sum_k \lambda_k R_{\mathcal{F}}(\tilde{f}_k) \leq \rho M^* + \lambda_\beta\left[R_\beta(\tilde{\beta}_S - \beta_S^\dagger) + R(\beta_S^\dagger)\right] + \sum_k \lambda_k R_{\mathcal{F}}(\tilde{f}_k). \tag{D.12}$$

Combining (D.11) and (D.12) and rearranging yields

$$\mathcal{E}(f^\dagger, \beta^\dagger) + \lambda_\beta R(\beta^\dagger - \tilde{\beta}) + \sum_k \lambda_k R_{\mathcal{F}}(f_k^\dagger - \tilde{f}_k) \leq \rho M^* + 2\lambda_\beta R((\beta^\dagger - \tilde{\beta})_S) + \lambda_\beta R(\tilde{\beta}_{S^C}) + 2\sum_k \lambda_k R_{\mathcal{F}}(\tilde{f}_k). \tag{D.13}$$

We consider two cases regarding the RHS of (D.13):

Case 1: $2\lambda_\beta R((\beta^\dagger - \tilde{\beta})_S) \leq 2\sum_k \lambda_k R_{\mathcal{F}}(\tilde{f}_k) + \rho M^*$

By Assumption 5.6, we have that $\mathcal{E}(f^\dagger, \beta^\dagger) \geq 0$. Also, recall that $\lambda_\beta \asymp \lambda_k \asymp \rho$ for all k. We therefore have

$$\begin{aligned}
\|\beta^\dagger - \tilde{\beta}\|_1 + \sum_k R_{\mathcal{F}}(f_k^\dagger - \tilde{f}_k) &\leq \frac{R((\beta^\dagger - \tilde{\beta})_S)}{\phi(m)} + \|(\beta^\dagger - \tilde{\beta})_{S^C}\|_1 + \sum_k R_{\mathcal{F}}(f_k^\dagger - \tilde{f}_k) \\
&\leq \frac{R(\beta^\dagger - \tilde{\beta})}{\phi(m)} + \sum_k R_{\mathcal{F}}(f_k^\dagger - \tilde{f}_k) + 2o_P(1) \\
&\leq C_1\sqrt{\frac{n}{\log n}}\left[\lambda_\beta R(\beta^\dagger - \tilde{\beta}) + \sum_k \lambda_k R_{\mathcal{F}}(f_k^\dagger - \tilde{f}_k)\right] \\
&\leq C_1\sqrt{\frac{n}{\log n}}\left[2\rho M^* + \lambda_\beta R(\tilde{\beta}_{S^C}) + 4\sum_k \lambda_k R_{\mathcal{F}}(\tilde{f}_k)\right]
\end{aligned} \tag{D.14}$$

for some constant $C_1 > 0$. By Assumption 5.2, $C_1\sqrt{n/\log n}\lambda_\beta R(\tilde{\beta}_{S^C}) < \frac{M^*}{6}$ when n is large enough. Thus, as long as $\rho \leq \frac{1}{12C_1}\sqrt{\log n/n}$ and $\max_k \lambda_k \leq \frac{1}{24C_1}\sqrt{\log n/n}$, (D.14) continues as

$$\|\beta^\dagger - \tilde{\beta}\|_1 + \sum_k R_{\mathcal{F}}(f_k^\dagger - \tilde{f}_k) \leq \frac{M^*}{6} + \frac{M^*}{6} + \frac{M^*}{6} = \frac{M^*}{2}.$$

Also, by (D.10), we have that

$$\begin{aligned}\|\hat{\beta} - \tilde{\beta}\|_1 + \sum_k R_{\mathcal{F}}(\hat{f}_k - \tilde{f}_k) &= \frac{1}{t}\left[\|\beta^\dagger - \tilde{\beta}\|_1 + \sum_k R_{\mathcal{F}}(f_k^\dagger - \tilde{f}_k)\right] \\ &\leq \frac{M^*}{2}\left[1 + \frac{\|\hat{\beta} - \tilde{\beta}\|_1 + \sum_k R_{\mathcal{F}}(\hat{f}_k - \tilde{f}_k)}{M^*}\right]\end{aligned}$$

i.e.,

$$\|\hat{\beta} - \tilde{\beta}\|_1 + \sum_k R_{\mathcal{F}}(\hat{f}_k - \tilde{f}_k) \leq M^*.$$

As a consequence, we can redo the calculations up to (D.14) with $(f^\dagger, \beta^\dagger)$ replaced by $(\hat{f}, \hat{\beta})$.

Case 2: $2\lambda_\beta R((\beta^\dagger - \tilde{\beta})_S) > 2\sum_k \lambda_k R_{\mathcal{F}}(\tilde{f}_k) + \rho M^*$

In this case, (D.13) can be rearranged as

$$\begin{aligned}&\mathcal{E}(f^\dagger, \beta^\dagger) + \lambda_\beta R(\beta^\dagger - \tilde{\beta}) + \sum_k \lambda_k R_{\mathcal{F}}(f_k^\dagger - \tilde{f}_k) \leq 4\lambda_\beta R((\beta^\dagger - \tilde{\beta})_S) + \lambda_\beta R(\tilde{\beta}_{S^C}) \\ \Rightarrow &\mathcal{E}(f^\dagger, \beta^\dagger) + \lambda_\beta R((\beta^\dagger - \tilde{\beta})_{S^C}) + \sum_k \lambda_k R_{\mathcal{F}}(f_k^\dagger - \tilde{f}_k) \leq 3\lambda_\beta R((\beta^\dagger - \tilde{\beta})_S) + \lambda_\beta R(\tilde{\beta}_{S^C}) \\ \Rightarrow &\lambda_\beta R((\beta^\dagger - \tilde{\beta})_{S^C}) \leq 3\lambda_\beta R((\beta^\dagger - \tilde{\beta})_S) + \lambda_\beta R(\tilde{\beta}_{S^C}). \qquad \text{(D.15)}\end{aligned}$$

Therefore, applying the compatibility condition in Assumption 5.7 yields

$$\begin{aligned}\mathcal{E}(f^\dagger, \beta^\dagger) + \lambda_\beta R(\beta^\dagger - \tilde{\beta}) + \sum_k \lambda_k R_{\mathcal{F}}(f_k^\dagger - \tilde{f}_k) &\leq 4\lambda_\beta \frac{\|\beta^\dagger - \tilde{\beta}\|\sqrt{m}}{\phi(m)} + \lambda_\beta R(\tilde{\beta}_{S^C}) \\ &\leq H\left(\frac{4\lambda_\beta\sqrt{m}}{\phi(m)}\right) + \mathcal{E}(f^\dagger, \beta^\dagger) + \lambda_\beta R(\tilde{\beta}_{S^C}) \\ &\leq \frac{16\lambda_\beta^2 m}{4c\phi^2(m)} + \mathcal{E}(f^\dagger, \beta^\dagger) + \lambda_\beta R(\tilde{\beta}_{S^C}) \\ &\leq \rho M^* + \mathcal{E}(f^\dagger, \beta^\dagger)\end{aligned}$$

where $H(\cdot)$ is the convex conjugate of $G(\cdot)$ as defined in Haris et al. (2019a). Recall that $\lambda_\beta \asymp \lambda_k \asymp \rho$, we have

$$\|\beta^\dagger - \tilde{\beta}\|_1 + \sum_k R_{\mathcal{F}}(f_k^\dagger - \tilde{f}_k) \leq \frac{M^*}{2}$$

as long as λ_β, λ_k and ρ are scaled properly such that $\lambda_\beta \geq 2\rho$ and $\min_k \lambda_k \geq 2\rho$. Consequently, we can repeat the steps in Case 1 and show that

$$\|\hat{\beta} - \tilde{\beta}\|_1 + \sum_k R_{\mathcal{F}}(\hat{f}_k - \tilde{f}_k) \leq M^*.$$

We therefore have, in both cases, that

$$\begin{aligned}
&\lambda_\beta \|\beta^\dagger - \tilde{\beta}\|_1 + \sum_k \lambda_k R_{\mathcal{F}}(f_k^\dagger - \tilde{f}_k) \asymp \rho M^* \\
\Rightarrow &\|\beta^\dagger - \tilde{\beta}\|_1 + \sum_k R_{\mathcal{F}}(f_k^\dagger - \tilde{f}_k) = O_P\left(\sqrt{\frac{m \log n}{n}}\right),
\end{aligned}$$

establishing the consistency of the causal estimates $\hat{\beta}$. □

D.1.2 Asymptotic Normality of $\hat{b}$

Proof of Theorem 5.2. We consider the case where the optimization problem (5.6) is specified by the matrix form (5.7), and let $(\tilde{\alpha}, \tilde{\beta})$ be the target parameter induced by $(\tilde{f}, \tilde{\beta})$, i.e., $\tilde{\alpha}$ satisfies $\Psi_k \tilde{\alpha}_k = (\tilde{f}_k(x_{1k}), \ldots, \tilde{f}_k(x_{nk}))$ for all k. We then recognize that $\tilde{\alpha} = \alpha^*$ for the population risk minimizer α^* defined in Assumption 5.9.

We have by applying Taylor's expansion that

$$\begin{aligned}
\sqrt{n} v^\top \hat{I}_{\beta|\alpha}^{1/2}(\hat{b} - \beta^*) &= \sqrt{n} v^\top \hat{I}_{\beta|\alpha}^{1/2}(\hat{\beta} - \tilde{\beta} - \hat{I}_{\beta|\alpha}^{-1} \hat{S}(\hat{\alpha}, \hat{\beta})) \\
&= \sqrt{n} v^\top \hat{I}_{\beta|\alpha}^{1/2}\left(\hat{\beta} - \tilde{\beta} - \hat{I}_{\beta|\alpha}^{-1} \hat{S}(\hat{\alpha}, \tilde{\beta}) - \hat{I}_{\beta|\alpha}^{-1} \bar{I}_{\beta|\alpha}(\hat{\beta} - \tilde{\beta})\right) \\
&= -\sqrt{n} v^\top \hat{I}_{\beta|\alpha}^{-1/2} \hat{S}(\hat{\alpha}, \tilde{\beta}) - \sqrt{n} v^\top \hat{I}_{\beta|\alpha}^{1/2}(\hat{I}_{\beta|\alpha}^{-1} \bar{I}_{\beta|\alpha} - 1)(\hat{\beta} - \tilde{\beta}) \qquad \text{(D.16)}
\end{aligned}$$

where $\bar{I}_{\beta|\alpha}$ is defined analogously as $\hat{I}_{\beta|\alpha}$ but with $\beta^\dagger := u\hat{\beta} + (1-u)\tilde{\beta}$ plugged in instead of $\hat{\beta}$, for some $u \in [0, 1]$. We will handle the two terms in (D.16) separately.

First, for the decorrelated score function $\hat{S}(\alpha, \beta)$, we have

$$\begin{aligned}
\hat{S}(\hat{\alpha}, \tilde{\beta}) &= \nabla_\beta L(\hat{\alpha}, \tilde{\beta}) - \hat{\xi}^\top \nabla_\alpha L(\hat{\alpha}, \tilde{\beta}) \\
&= \nabla_\beta L(\tilde{\alpha}, \tilde{\beta}) - \hat{\xi}^\top \nabla_\alpha L(\tilde{\alpha}, \tilde{\beta}) + \left[\nabla^2_{\beta\alpha} L(\alpha^\dagger, \tilde{\beta}) - \hat{\xi}^\top \nabla^2_{\alpha\alpha} L(\alpha^\dagger, \tilde{\beta})(\hat{\alpha} - \tilde{\alpha})\right]
\end{aligned}$$

where $\alpha^\dagger = u\tilde{\alpha} + (1-u)\hat{\alpha}$ for some $u \in [0, 1]$. This leads to

$$\begin{aligned}
\left\|\hat{S}(\hat{\alpha}, \tilde{\beta}) - S(\tilde{\alpha}, \tilde{\beta})\right\|_1 &\leq \|(\xi^* - \hat{\xi})^\top \nabla_\alpha L(\tilde{\alpha}, \tilde{\beta})\|_1 + \left\|\left[\nabla^2_{\beta\alpha} L(\alpha^\dagger, \tilde{\beta}) - \hat{\xi}^\top \nabla^2_{\alpha\alpha} L(\alpha^\dagger, \tilde{\beta})\right](\hat{\alpha} - \tilde{\alpha})\right\|_1 \\
&= O_P(\varepsilon_1(n)) \cdot o_P(1) + O_P(\varepsilon_2(n)) \cdot o_P(1) = o_P(n^{-1/2}) \qquad \text{(D.17)}
\end{aligned}$$

by Assumptions 5.9 and 5.10. In addition,

$$\begin{aligned}\|\hat{I}_{\beta|\alpha} - I^*_{\beta|\alpha}\|_1 &\leq \|\nabla^2_{\beta\beta} L(\hat{\alpha}, \hat{\beta}) - I^*_{\beta\beta}\|_1 + \|\xi^{*\top}(I^*_{\alpha\beta} - \nabla^2_{\alpha\beta} L(\hat{\alpha}, \hat{\beta}))\|_1 + \|(\hat{\xi} - \xi^*)^\top \nabla^2_{\alpha\beta} L(\hat{\alpha}, \hat{\beta})\|_1 \\ &= O_P(\varepsilon_3(n)) + [O_P(\|\xi^*\|_1 \varepsilon_3(n)) + O_P(\varepsilon_1(n)\|I^*_{\alpha\beta}\|_1)] + O_P(\varepsilon_1(n)\varepsilon_3(n)) = o_P(1)\end{aligned} \tag{D.18}$$

by Assumptions 5.9 and 5.11 and the assumptions specified under Theorem 5.2.

By the definition of $\tilde{\alpha}_k$ and $\tilde{f}$, we have that $\Psi_k \tilde{\alpha}_k = (f^*_k(x_{1k}), \ldots, f^*_k(x_{nk}))$. Recall also that $\tilde{\beta} = \beta^*$ by Lemma D.1. We thus have

$$\sqrt{n} v^\top {I^*_{\beta|\alpha}}^{-1/2} S(\tilde{\alpha}, \tilde{\beta}) \xrightarrow{d} N(0, 1) \tag{D.19}$$

for any v such that $v^\top v = 1$.

We are now ready to analyze the first term of (D.16), which is

$$\begin{aligned}-\sqrt{n} v^\top \hat{I}^{-1/2}_{\beta|\alpha} \hat{S}(\hat{\alpha}, \tilde{\beta}) &= -\sqrt{n} v^\top {I^*_{\beta|\alpha}}^{-1/2} S(\tilde{\alpha}, \tilde{\beta}) - \sqrt{n} v^\top {I^*_{\beta|\alpha}}^{-1/2} \left[\hat{S}(\hat{\alpha}, \tilde{\beta}) - S(\tilde{\alpha}, \tilde{\beta})\right] \\ &\quad - \sqrt{n} v^\top \left(\hat{I}^{-1/2}_{\beta|\alpha} - {I^*_{\beta|\alpha}}^{-1/2}\right) \hat{S}(\hat{\alpha}, \tilde{\beta}) \\ &= -\sqrt{n} v^\top {I^*_{\beta|\alpha}}^{-1/2} S(\tilde{\alpha}, \tilde{\beta}) + o_P(1) O_P(1) + o_P(1) \\ &= -\sqrt{n} v^\top {I^*_{\beta|\alpha}}^{-1/2} S(\tilde{\alpha}, \tilde{\beta}) + o_P(1)\end{aligned} \tag{D.20}$$

where the first term converges to $N(0, 1)$ distribution, and the last two terms are controlled as consequences of (D.17) and (D.18).

The second term of (D.16) can be rearranged as

$$-\sqrt{n} v^\top \hat{I}^{1/2}_{\beta|\alpha} (\hat{I}^{-1}_{\beta|\alpha} \bar{I}_{\beta|\alpha} - 1)(\hat{\beta} - \tilde{\beta}) = -\sqrt{n} v^\top \hat{I}^{-1/2}_{\beta|\alpha} (\bar{I}_{\beta|\alpha} - \hat{I}_{\beta|\alpha})(\hat{\beta} - \tilde{\beta}). \tag{D.21}$$

Furthermore, similar to (D.18), we have

$$\|\bar{I}_{\beta|\alpha} - \hat{I}_{\beta|\alpha}\|_1 \leq \|\nabla^2_{\beta\beta} L(\hat{\alpha}, \bar{\beta}) - \nabla^2_{\beta\beta} L(\hat{\alpha}, \hat{\beta})\|_1 + \|\hat{\xi}^\top (\nabla^2_{\alpha\beta} L(\hat{\alpha}, \bar{\beta}) - \nabla^2_{\alpha\beta} L(\hat{\alpha}, \hat{\beta}))\|_1 = O_P(\|\xi^*\| \varepsilon_3(n))$$

which, when plugged back into (D.21), yields

$$\begin{aligned}|-\sqrt{n} v^\top \hat{I}^{1/2}_{\beta|\alpha} (\hat{I}^{-1}_{\beta|\alpha} \bar{I}_{\beta|\alpha} - 1)(\hat{\beta} - \tilde{\beta})| &\leq O_P(\sqrt{n}) \cdot \|\bar{I}_{\beta|\alpha} - \hat{I}_{\beta|\alpha}\|_1 \|\hat{\beta} - \tilde{\beta}\|_1 \\ &= O_P(\sqrt{n}) o_P(1/\sqrt{\log n}) O_P\left(\sqrt{\frac{\log n}{n}}\right) = o_P(1).\end{aligned} \tag{D.22}$$

Combining (D.20) and (D.22), we have established the asymptotic linearity of $\hat{b}$ and showed its asymptotic distribution as claimed. □