ABSTRACT

WILSON, T. ANDER. Advances in Bayesian Methods for High-Dimensional Environmental Data. (Under the direction of Brian J. Reich.)

In many applications there is prior knowledge to support a monotone relationship between the exposure and an outcome. In these situations, statistical models can be improved by incorporating prior knowledge of a monotonic relationship. In this dissertation, we present methodologies to monotone regression in hierarchical and spatial models with applications to environmental statistics. The first case looks at hierarchical dose-response estimation in high-dimensional toxicology experiments. The second case looks at spatially-varying bivariate exposure-response function with monotonicity constraints in the context of multi-city air pollution studies. In observational studies, such as the air pollution example, the results can be sensitive the the choice of confounding variables included in the model. The third part of this dissertation presents methodology to selecting confounders in observational studies.

High-throughput screening (HTS) of environmental chemicals is used to identify chemicals with high potential for adverse human health and environmental effects from among the thousands of untested chemicals. Predicting physiologically-relevant activity with HTS data requires estimating the response of a large number of chemicals across a battery of screening assays based on sparse dose-response data for each chemical-assay combination. Many standard dose-response methods are inadequate because they treat each curve separately and under-perform when there are as few as six to ten observations per curve. We propose a semiparametric Bayesian model for monotone dose-response estimation that borrows strength across chemicals and assays. Our method directly parametrizes the efficacy and potency of the chemicals as well as the probability of response. We use the ToxCast data from the U.S. Environmental Protection Agency (EPA) as motivation. We demonstrate that our hierarchical method provides more accurate estimates of the probability of response, efficacy, and potency than separate curve estimation in a simulation study. We use our semiparametric method to compare the efficacy of chemicals in
the ToxCast data to well-characterized reference chemicals on estrogen receptor \( \alpha \) (ER\( \alpha \)) and peroxisome proliferator-activated receptor \( \gamma \) (PPAR\( \gamma \)) assays, then estimate the probability that other chemicals are active at lower concentrations than the reference chemicals.

Climate change is expected to alter the distribution of ambient ozone levels and temperatures which, in turn, may impact public health. Much research has focused on the effect of short-term ozone exposures on mortality and morbidity while controlling for temperature as a confounder, but less is known about the joint effects of ozone and temperature. The extent of the health effects of changing ozone levels and temperatures will depend on whether these effects are additive or synergistic. In this dissertation we propose a spatial, semi-parametric model to estimate the joint ozone-temperature risk surfaces in 95 US urban areas. Our methodology restricts the ozone-temperature risk surfaces to be monotone in ozone and allows for both non-additive and non-linear effects of ozone and temperature. We use data from the National Mortality and Morbidity Air Pollution Study (NMMAPS) and show that the proposed model fits the data better than additive linear and non-linear models. We then examine the synergistic ozone-temperature effect both nationally and locally and find evidence of a non-linear ozone effect and an ozone-temperature interaction at higher temperatures and ozone concentrations.

When estimating the effect of an exposure or treatment on an outcome, such as the effect of ozone on mortality, it is important to select the proper subset of confounding variables to include in the model. Including too many covariates increases mean square error on the effect of interest while not including confounding variables biases the exposure effect estimate. We propose a decision-theoretic approach to confounder selection and effect estimation. We first estimate the full standard Bayesian regression model and then post-process the posterior distribution with a loss function that penalizes models omitting important confounders. Our method can be fit easily with existing software and in many situations without the use of Markov chain Monte Carlo methods, resulting in computation on the order of the least squares solution. We prove that the proposed estimator has attractive asymptotic properties. In a simulation study we show that our method outperforms existing methods.
Advances in Bayesian Methods for High-Dimensional Environmental Data

by
T. Ander Wilson

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2014

APPROVED BY:

Montserrat Fuentes

David M. Reif

Ana-Maria Staicu

Brian J. Reich
Chair of Advisory Committee
DEDICATION

To my family.
BIOGRAPHY

The author was born in Newton, Massachusetts. He graduated from Newton North High School and matriculated to the University of Vermont. After graduating with a Bachelors of Arts Degree in mathematics and taking a ten month hiatus to kayak and ski, he became a senior programmer analyst at Mathematica Policy Research in Cambridge, Massachusetts. After four years of nutrition and health policy research at Mathematica Policy Research, he enrolled in the North Carolina State University Department of Statistics where he received a Masters of Statistics in 2011 and PhD in 2014.
ACKNOWLEDGEMENTS

I would like to thank my parents and Abby for their unending support and encouragement, and my advisor, Dr. Brian Reich, for his guidance and patience along the way. In addition, I would like to thank my collaborators and EPA mentors: Drs. David Reif, Ana Rappold, and Lucas Neas.

I would also like to thank my funding sources. NIH training grant GM081057: Biostatistics Training in the Omics Era provided funding for my first three years. This includes the funding for material presented in Chapter 2. In addition, the research in Chapter 3 was supported in part by an appointment to the Research Participation Program for the U.S. Environmental Protection Agency, Office of Research and Development, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and EPA.
# TABLE OF CONTENTS

**LIST OF TABLES** ........................................... viii

**LIST OF FIGURES** ........................................... ix

**Chapter 1 Introduction** .................................. 1

1.1 Introduction ............................................. 1

1.1.1 Hierarchical Monotone Dose-Response Modeling for High-Throughput Toxicity Screening of Environmental Chemicals ........................................... 2

1.1.2 Modeling the Effect of Temperature on Ozone-Related Mortality ........... 3

1.1.3 Confounder Selection via Penalized Credible Regions ......................... 4

**Chapter 2 Hierarchical Dose-Response Modeling for High-Throughput Toxicity Screening of Environmental Chemicals** .......... 6

2.1 Introduction ............................................. 6

2.2 The ToxCast Data .......................................... 10

2.3 Model Description ........................................ 12

2.3.1 Hierarchical Structure and Prior Specification ............................... 13

2.3.2 Assay Effects, Chemical Effects, and Prior Knowledge ....................... 14

2.3.3 Posterior Computation .................................. 15

2.4 Model Comparisons ....................................... 16

2.4.1 Simulation Study ....................................... 16

2.4.2 Cross Validation and Model Fit ................................ 16

2.5 ToxCast Data Application .................................. 17

2.5.1 Summary of Active Responses and Assay and Chemical Effects ........... 19

2.5.2 Comparison with Reference Chemicals ..................................... 21

2.6 Discussion ................................................ 23

2.7 Software .................................................. 24

**Chapter 3 Modeling the Effect of Temperature on Ozone-Related Mortality** ... 25

3.1 Spatial monotone surface model ................................ 29

3.1.1 Ozone-temperature surface model .................................. 29

3.1.2 Hierarchical model for monotonicity and spatial smoothing ............... 30

3.1.3 Confounder model ........................................ 31

3.2 A two-stage approach for large datasets .................................. 32

3.2.1 Stage 1: City-specific GLM regression .................................. 33

3.2.2 Stage 2: Bayesian model for stage 1 output ................................ 33

3.2.3 Priors and computational details ..................................... 34

3.3 Analysis of the ozone-temperature log RR surfaces ............................ 35

3.3.1 Cross-validation ........................................... 35

3.3.2 Analysis of the national average log RR surface ............................ 37

3.3.3 Analysis of the city-specific log RR surfaces ................................ 40

3.3.4 Analysis of excess mortality ..................................... 43
Chapter 4 Confounder Selection via Penalized Credible Regions

4.1 Introduction
4.2 Methods
  4.2.1 Modeling approach
  4.2.2 Penalized regression reformulation
  4.2.3 Simplification under flat prior
  4.2.4 Extension to multiple exposures
4.3 Theoretical Results
4.4 Computation and tuning
4.5 Simulation Study
  4.5.1 Simulation with linear model
  4.5.2 Simulation in the ultra high-dimensional setting
  4.5.3 Simulation with binary treatment and logistic confounder model
4.6 Data Analysis
4.7 Discussion
4.8 Software

Chapter 5 Future Directions

5.1 Extensions for Hierarchical Dose-Response Modeling
5.2 Extensions for Multi-Pollutant Modeling
5.3 Extensions for Confounder and Variable Selection

References

Appendices

Appendix A Supplemental Material for Hierarchical Dose-Response Modeling for High-Throughput Toxicity Screening of Environmental Chemicals
  A.1 Computation and Full Conditionals
    A.1.1 Conditional Posterior Distributions
    A.1.2 Priors for ZILL
  A.2 Simulation
  A.3 Additional Figure

Appendix B Supplemental Material for Modeling the Effect of Temperature on Ozone Related Mortality
  B.1 Additional Figures
  B.2 Cross-Validation Results
  B.3 Full Conditional Distribution
    B.3.1 Full Conditionals
  B.4 MCMC algorithm
  B.5 Trace Plots

Appendix C Supplemental Material for Confounder Selection via Penalized Credible Regions
C.1 Tuning ................................................................. 104
C.2 Proof of Theorem 1 ................................................. 105
C.3 Covariates included in data analysis ....................... 107
### LIST OF TABLES

| Table 3.1 | Difference in cross-validation deviance from linear additive model (NMMAPS model). | 37 |
| Table 3.2 | Mean log RR on high and moderate temperature days over the observed and common ozone rates by region (in percent change in mortality per 10 ppb increase in ozone). | 43 |
| Table 3.3 | Percent increase in mortality associated with an increase from the medians of ozone and temperature to the 95th percentiles of ozone and temperature using different models. | 45 |
| Table 4.1 | Simulation results for design 1. Bias, MSE, and coverage (Cover) are for the effect of interest $\hat{\beta}_x$. Coverage is 95% confidence or credible interval coverage. CPU Time is reported in seconds on a MacBook Pro with OS X, 8 GB RAM, and 2 GHz Intel Core i7. SEs for the AUC range from 0.002 to 0.004, and for CPU time from less than 0.001 to 0.498. | 67 |
| Table 4.2 | Simulation results for design 2. Bias, MSE, and coverage (Cover) are for the effect of interest $\hat{\beta}_x$. Coverage is 95% confidence or credible interval coverage. CPU Time is reported in seconds on a MacBook Pro with OS X, 8 GB RAM, and 2 GHz Intel Core i7. SEs for the AUC range from less than 0.001 to 0.006, and for CPU time from less than 0.001 to 0.003. | 68 |
| Table A.1 | Simulation results for mean response and probability of active response. The displayed values are the mean (standard error) across simulated data sets. RMSE is the pointwise root mean square error taken over 50 evenly spaced points on log base 3, the spacing scale for concentrations in the real data. RMSE Top and RMSE AC50 are only calculated for dose-response curves that are active. AUC is the area under the ROC curve. | 92 |
| Table C.1 | Additional covariates included in the data analysis. | 108 |
LIST OF FIGURES

Figure 2.1 Panel (a) shows an annotated example of the four-parameter log-logistic (FPLL) function. The FPLL model is $f(x; t, b, a, w) = t - (t - b) \times \text{Logit} [-w\{\log(x) - \log(a)\}]$, where $x$ is the tested concentration and $(t, b, w, a)$ parameterizes the Emax, Emin, AC50, and $w$ (rate of increase), respectively. Panel (b) shows six sample basis functions with internal knots (−4, −2, 0, 2, 4) marked with gray circles. This figure appears in color in the electronic version of this article.

Figure 2.2 Illustration of the ToxCast data structure showing 20 chemicals tested on 8 assays. This is a sample of the 309 chemicals and 81 assays used in Section 2.5 with curves as reported in the publicly available ToxCast data. All 309 chemicals are tested on all 81 assays, resulting in $309 \times 81 = 25,029$ chemical-assay combinations. This figure appears in color in the electronic version of this article.

Figure 2.3 Example of ZIPPLL and ToxCast estimates for 12 chemicals on the PXRE assay. The ZIPPLL posterior mean (thick solid black line) with 95% posterior intervals (dashed black lines), and the ToxCast fit (thin solid gray line) are shown. The legend shows a binary indicator of an active response from ToxCast (1=active) and the posterior probability of an active response using ZIPPLL. The FPLL fits are redrawn from the parameter estimated in the ToxCast public release data. This figure appears in color in the electronic version of this article.

Figure 2.4 Panel (a) plots the number of assays each chemical responded on with 95% posterior intervals. The simultaneous fitting of all chemical-assay combinations allows for the estimation of the joint distribution of assays and naturally propagates the distribution of the total number of assay responses. The x indicates the number of assay responses reported in the ToxCast data. (The 309 chemical names are omitted on the horizontal axis for readability.) Panel (b) shows posterior mean random intercept for the assay probability of response ($\delta_j$) as discussed in Section 2.3.2 and 95% posterior intervals. This figure appears in color in the electronic version of this article.

Figure 2.5 Chemical ranks by potency with 90% credible intervals. The x-axis shows the posterior mean AC50 (more potent to the left). The y-axis ranks the chemicals by potency (most potent, rank one, at top). All chemicals with at least a 0.50 percent probability of active response are plotted. For PPAR$\gamma$ and ER$\alpha$ reference chemicals are marked for comparison. These chemicals have shown documented activity on these assays. This figure appears in color in the electronic version of this article.

Figure 2.6 Posterior probability that chemicals are more active than selected reference chemicals on PPAR$\gamma$. All chemicals with at least a 0.05 probability of being more active than all three reference chemicals are shown and ordered by their posterior probability.
Figure 3.1 Ozone-temperature distribution for selected cities. The upper and lower boxes contain data for high temperature days and moderate temperature days, respectively, over the observed ozone range. The purple subsections highlight the common ozone range, the intersection of the ozone ranges for high and moderate temperature days. Details on the definition used to identify these days are in Section 3.3.3. Seattle is the only city in the dataset that does not have a common ozone range.

Figure 3.2 Panels (a) and (b) show the pointwise mean and standard deviation of the national log RR surfaces. Panels (c) and (d) show cross section of the log RR surface with ozone fixed at 50 and 100 ppb, respectively, along with 95% posterior intervals. Log RR is in percent change in mortality per one ppb increase in ozone.

Figure 3.3 Comparison of the log RR at the 50th, 75th, 95th, and 99th percentiles of temperature. The estimates extend over the range of ozone values observed in at least 5 cities. Subfigure 3.3a shows the mean log RR for each cross-section and Subfigure 3.3b shows the pointwise posterior probability that log RR is greater at the high temperature. Log RR is in percent change in mortality per one ppb increase in ozone.

Figure 3.4 National interaction surface and pointwise probability of positive interaction. Panel 3.4a shows the national interaction surface which is the cross-derivative of the log risk surface or the derivative of the log RR surface with respect to temperature. This shows how log RR changes with temperature and quantifies the interaction at each point. Panel 3.4b show the probability that the national interaction surface is greater than 0.

Figure 3.5 Log RR surfaces for selected cities (left) and their pointwise standard deviations (right). Surfaces are plotted only over the range of data observed data for that city. The cities were selected for being geographically diverse and having varied ozone and temperature ranges.

Figure 3.6 Comparison of the log RR at high temperature and moderate temperatures as defined in Section 3.3.3. Panel 3.6a compares the ratio of log RR over the observed ozone range (black) and common ozone range (blue). The posterior mean and 95% interval are shown. Panel 3.6b shows a map of the ratio on days in the common ozone range. Cities outlined in black (Los Angeles and St. Louis) are significant at the 0.8 level.

Figure 4.1 Probability of including covariates for simulation design one with $n = 100$ for the credible region method (●), BAC (○), BMA (△), and adaptive LASSO (+). Left: the proportion of simulated data sets for which each variable is selected. For BMA and BAC a variable is counted as included if its inclusion probability is greater than 0.5. Right: the average inclusion probability for BMA and BAC. Covariates in the first section (1 to 7) are confounders, in the second section (8 to 14) are other explanatory variables, and the far right dot is the average of the non-important covariates (15 to 57).
Figure 4.2  AUC for simulation design one with \( p > n \) for three different rates of growth for \( n \).

Figure 4.3  Solution path for three subgroups and overall. The thick black line is the estimated PM effect, \( \hat{\beta}_x \). The other lines are the estimates for the other covariates in the model. The regression coefficients correspond to centered and scaled variables.

Figure 4.4  Estimates of the PM effect (\( \hat{\beta}_x \)). Top: The point estimates with each model and the 95% interval. Bottom: The ratio of the SE or SD of \( \hat{\beta}_x \) with each model compared to the full OLS model.

Figure A.1  Example of simulated regression functions under using ZILL A.1a and mixture of normals A.1b.

Figure A.2  Comparison of the posterior probability of an active response estimated with ZIPLL (x-axis) and the binary indicator of active response from ToxCast (y-axis) for all 309 chemicals on three assays. The majority of chemicals are considered not active (clustered in bottom left) or active (upper right) with both methods. However, using ZIPLL we estimated that several chemicals have posterior probabilities of response between 0.1 and 0.9, suggesting that there is not conclusive evidence that these chemicals responded or not, but are forced to be classified as either active or not in the ToxCast data.

Figure B.1  Summary of the ozone and temperature distribution.

Figure B.2  Demonstration of the different basis expansions used in the first and second stages. The bottom left shows the observed ozone-temperature distribution in each city. To the right are the basis functions in the temperature direction and to the top are the basis functions in the ozone direction for each stage. The first stage basis functions are different in each city.

Figure B.3  Difference in deviance from the linear, additive model (NMMAPS model) for the spatial monotone model with different numbers of basis functions in the ozone (M1) and temperature (M2) directions.

Figure B.4  Average rank of CV results across Figures B.3a through B.3d, with one being the best performing model. The best model on average was \((M_1, M_2) = (7, 9)\) with an average rank of 3.75; this is the model used for analysis. The next best models were \((9,10), (10,9),\) and \((9,9)\) with average ranks of 5.25, 5.50, and 6.75, respectively.

Figure B.5  Trace plots of the point wise estimates of the risk surface.

Figure B.6  Trace of the posterior sample of \( \log \rho \) (log range in log kilometers). The trace plot only includes the posterior sample and does not show the portion of the burn in portion of the chain. The posterior mean is 922.9, the median is 659.4, and the 95% symmetric interval is \((291.8, 3134.5)\).
Figure C.1 Comparison of model selection methods for simulation design 1. Subfigure C.1a shows the AUC while the remaining subfigures compare model size and performance with different selection criteria. The criteria are forward selection with $\alpha_{fs} = 0.25$ (solid; ———), forward selection with $\alpha_{fs} = 0.15$ (short dash gray; - - - -), $C_p$ (dotted; ······), CV (short dash blue; - - - -), AIC (dash-dot-dash; ·------), and BIC (long dash; — — — —). Model size is the number of confounders selected and does not count the outcome of interest.
Chapter 1

Introduction

1.1 Introduction

In many applications there is scientific evidence that an exposure’s effect on an outcome may be monotonic in nature. In such situations, statistical estimation of the exposure-response relationship can be improved by incorporating prior knowledge of monotonicity into the model. Many methods have been developed for monotonic regression (e.g. Friedman and Tibshirani, 1984; Mukerjee, 1988; Mammen, 1991; Hall and Huang, 2001; Holmes and Heard, 2003; Neelon and Dunson, 2004; Wang and Leng, 2008; Curtis and Ghosh, 2011a). In this dissertation, we develop methodology for monotonic regression in hierarchical and spatial models with applications to environmental exposures. In the first chapter, we use a hierarchical Bayesian model to estimate dose-response relationships in the context of high-dimensional toxicology studies. In this setting, the designed toxicology experiment results in data with a known non-decreasing relationship between dose and multiple in vitro outcomes. In the second chapter, we use a spatial model to estimate bivariate, semi-parametric risk surfaces in different locations. In this case, we are interested in the effect of co-exposures to ozone and temperature on mortality and assume that the ozone effect has a monotonic relationship with mortality. In observational studies such as the air pollution example described in Chapter 3 the estimated exposure effect
can be sensitive to the choice of confounding variables included in the model. In Chapter 4 we present methodology to choose confounding variables in observational studies.

1.1.1 Hierarchical Monotone Dose-Response Modeling for High-Throughput Toxicity Screening of Environmental Chemicals

There are thousands of untested chemicals in common use today (Judson et al., 2009). Comprehensive testing of all chemicals is not feasible due to the financial and temporal costs of traditional chemical testing, which relies heavily on animal testing. Recently, efforts have been made to prioritize the large number of untested chemicals so that available testing resources can be targeted at those chemicals that are most likely to cause adverse health or environmental effects. This approach combines high-throughput screening (HTS) which tests chemicals on a large battery of relatively fast and inexpensive bioassays and predictive models.

The data for each chemical-assay combination consists of dose-response pairs and are often known to have a monotonic relationship. Large designed experiments such as the US Environmental Protection Agency’s (EPA) ToxCast which tests thousands of chemicals on hundreds of HTS assays can result in large datasets (Dix et al., 2007; Kavlock et al., 2012). However, the data for each chemical-assay combination can be sparse. Hence, HTS data consists of a large number of sparse dose-response relationships from which we seek to rank chemicals based on their relative bioactivity as measured on HTS assays as well as estimate key parameters used in existing predictive model. As such, statistical methods are needed that estimate the dose-response relationship in a way that parameterizes key statistics and facilitates ranking of chemicals and other between chemical comparisons. At the same time, these methods must be robust to the sparsity of data observed in each chemical-assay combination, presumably by sharing information across the large number of chemical-assay pairs, and being efficient enough to allow for estimation of large HTS datasets.

In Chapter 2 we present a hierarchical Bayesian model for the estimation of monotone dose-response curves in HTS experiments. The proposed method effectively parametrizes key
summary statistics used in predictive models – the potency, efficacy, and probability of an active response – while allowing for flexibility of the shape of the dose-response relationship. Using the proposed method, we estimate the dose-response relationship of a subset of the ToxCast data and rank chemicals based on their potential potency on key bioassay.

1.1.2 Modeling the Effect of Temperature on Ozone-Related Mortality

The EPA has concluded that current scientific evidence supports a causal relationship between ozone and respiratory health effects and a likely to be causal relationship between ozone and cardiovascular health effects and mortality (US EPA, 2013). Several observational studies have associated ambient ozone exposures with increased rates of mortality while controlling for temperature and other confounding variables (Bell et al., 2004), while clinical research has found that ozone’s effect on lung function is monotonically increasing (McDonnell et al., 2012). With the onset of anticipated climate changes (National Research Council, 2004a) both ozone and temperature distributions are expected to increase in the near future (US EPA, 2009). To anticipate the health burden of these changes it is important to understand the joint health effects of ozone and temperature.

Many multi-city time-series studies have associated ambient ozone exposures with increased risk of non-accidental mortality while controlling for temperature (Bell et al., 2004, 2006; Smith et al., 2009). These results includes evidence of a non-linear ozone effect (Bell et al., 2006) and a potentially greater linear ozone effect at higher temperatures than at lower temperatures (Smith et al., 2009). However, estimates of interaction based on comparing the linear ozone effect at different temperatures may be confounded by the different ozone distributions observed at different temperatures and non-linearity in the ozone effect. To isolate interaction it is essential to compare the ozone effect at different temperatures but for the same ozone level.

One way to isolate interaction is to estimate the full ozone-temperature risk surface. By estimating the bivariate risk surface the rate of change in risk associated with changes in ozone exposures, i.e. the ozone effect, can be evaluated at any ozone and temperature combination.
Hence, comparisons of the ozone effect can be made at different temperatures but for the same ozone level. This approach has recently been employed in the single-city setting (Ren et al., 2008; Chen et al., 2013) and visual inspection of the estimated risk surfaces suggests that there is a non-linear and non-additive joint ozone-temperature effect.

There are several challenges to estimating ozone-temperature risk surfaces in multi-city studies. Chief among them is the substantially different ozone and temperature ranges observed in each city. As a result, it is challenging to find a parameterization of the risk surface that is appropriate to use at all locations, while using a different basis expansion at each city does not allow for the use of existing estimation methods such as those used in Dominici et al. (2002). In addition, there is a computational burden associated with estimating surface models at many locations. These issues are addressed in Chapter 3 which details a model for spatially-varying risk surfaces which are constrained to be monotone in ozone.

1.1.3 Confounder Selection via Penalized Credible Regions

In many observations studies the primary interest lies in estimating the effect of an exposure on an outcome while controlling for a large number of potential confounding variables. In such cases, the exposure effect can be sensitive to the choice of confounding variables included in the model. The sensitivity to confounder selection in multi-city time-series studies, such as the work presented in Chapter 3, has been documented by Sacks et al. (2012). Omitting important confounding variables, those correlated with both the outcome and the exposure, can bias the exposure estimate while including too many covariates independent of both the outcome and exposure can inflate the variance of the effect estimate. In problems where the primary interest is in parameter estimation it is important to select an appropriate subset of the potentially large number of confounding variables.

There are ample variable selection methods available. These include: all subsets, forward selection, an array of penalized regression methods (Tibshirani, 1996; Fan and Li, 2001; Zou and Hastie, 2005; Zou, 2006; Bondell and Reich, 2008), and Bayesian methods (George and
McCulloch, 1993; George and Foster, 2000; Brown et al., 2002; Carvalho et al., 2010; Bondell and Reich, 2012), among others. Variable selection, in general, focuses on balancing parsimony and model fit as estimated by minimizing the sum of squared errors. These methods can be highly effective in selecting variables for prediction, but can be biased for effect estimation. Bayesian model averaging (BMA) attempts to average the effect estimates over the model space giving more weight to models that have greater support from the data. However, BMA also tunes for prediction and can result in biased estimates of the effect of interest when the exposure is correlated with other covariates. Hence, a new class of selection methods are needed that explicitly treat one covariate as the exposure or treatment of interest and focus on estimating the associated regression coefficient.

To address the issue of confounder selection and effect estimation Crainiceanu et al. (2008) proposed a two-stage approach. In the first stage, the exposure of interest is regressed on the covariates to identify potential confounders that are correlated with the exposure of interest. Then, in the second stage, the outcome is regressed on the exposures of interest and the other observed covariates to identify additional covariates correlated with the outcome that were not identified as potential confounders in the first stage. Bayesian adjustment for confounding (BAC; Wang et al., 2012) formalized these two steps into a Bayesian approach that estimates the two models simultaneously with a model-averaging approach. BAC proved effective at large sample sizes; however, it was biased at small sample sizes, lacked optimality properties and is computationally inefficient.

Chapter 4 of this thesis presents a new method for confounder selection. The proposed method takes a decision-theoretic approach following that of Bondell and Reich (2012) but uses an exposure model and a loss function that explicitly penalizes models that omit important confounding variables. Chapter 4 details the methodology and presents simulation studies and theoretical results for the proposed estimator.
Chapter 2

Hierarchical Dose-Response Modeling for High-Throughput Toxicity Screening of Environmental Chemicals

2.1 Introduction

There are thousands of untested chemicals in common use. Comprehensive toxicity testing of all chemicals is infeasible due to high monetary and temporal costs (Judson et al., 2009). To address this problem, a new paradigm in toxicity testing focuses on screening larger numbers of chemicals on a diverse battery of relatively quick and inexpensive high-throughput screening (HTS) assays that measure a variety of cellular and biochemical responses. Each assay measures a single endpoint, such as transcription of a target gene or binding to a specific receptor protein. The aim of HTS is to predict which chemicals are most likely to perturb normal biological processes that lead to adverse human health and environmental effects, and focus scarce testing resources on those chemicals.
To predict potential chemical activity from HTS data requires statistical models to estimate the response of each chemical on each assay and to compare and rank chemicals. There are three main quantities used to compare chemicals: 1) the probability that an active response occurred, 2) the potency or concentration at which a response occurs, and 3) the efficacy or magnitude of the response. These three quantities form the basis of chemical prioritization. With improved estimates of these three quantities, predictive models will be able to better predict which chemicals are most likely to have potentially hazardous effects.

Dose-response modeling for HTS data is unique because there are a large number of curves to estimate but the data for each curve are sparse. For example, the ToxCast project at the US EPA (Dix et al., 2007; Kavlock et al., 2012) has screened nearly 2,000 chemicals on over 700 HTS assays; however, each chemical-assay combination is tested at six to ten unique concentrations and, in most cases, in singlicate at each concentration. Analysis is further complicated by assay and chemical effects, such as assays that are more or less sensitive, correlated assays that measure the same or similar cellular response, and chemicals that are highly active or not active on a variety of assays. Hence, HTS requires a dose-response method that is robust to the sparsity of the data for each chemical-assay combination, takes advantage of the larger number of chemicals and assays, and accurately estimates the efficacy, potency, and probability of an active response.

A variety of parametric models are used for estimating monotonic dose-response curves (Ritz, 2010). The most common method is the four-parameter log-logistic model (FPLL) which directly parameterizes the efficacy and potency with the Emax (maximal response or upper asymptote) and AC50 (concentration at which the half maximal response occurs), respectively. Figure 2.1a shows an annotated example FPLL dose-response curve. The current release of EPA ToxCast results uses FPLL to fit each dose-response curve with least squares (Judson et al., 2010). Fitting six to ten observations with a four-parameter model using least squares results in poor variance estimates (currently not provided in the ToxCast public release) and no estimates of the probability that an active response occurs.
**Figure 2.1:** Panel (a) shows an annotated example of the four-parameter log-logistic (FPLL) function. The FPLL model is $f(x; t, b, w, a) = t - (t - b) \times \text{Logit} \left[ -w \{\log(x) - \log(a)\}\right]$, where $x$ is the tested concentration and $(t, b, w, a)$ parameterizes the Emax, Emin, AC50, and $w$ (rate of increase), respectively. Panel (b) shows six sample basis functions with internal knots $(-4, -2, 0, 2, 4)$ marked with gray circles. This figure appears in color in the electronic version of this article.

There are several available frequentist approaches for monotonic curve estimation (e.g. Friedman and Tibshirani, 1984; Mukerjee, 1988; Mammen, 1991; Hall and Huang, 2001; Mammen et al., 2001; Wang and Li, 2008). Recently, several semiparametric Bayesian methods for monotone regression for a single curve have been proposed. Holmes and Heard (2003) proposed using piecewise constant functions with random knots, Neelon and Dunson (2004) utilized a piecewise linear spline model, and Curtis and Ghosh (2011a) developed a Bernstein polynomial model. Also, Shively et al. (2009) modeled the monotonic function as the integral of a positive function. None of these general semiparametric regression methods directly parameterize the efficacy or potency of a dose-response relationship, which are widely used for comparing chemicals and
predictive modeling.

Our problem is different from these because we are estimating the response for several chemicals on multiple assays. Bayesian hierarchical models have been used in many fields where the data takes a natural hierarchical structure. Several in vivo or developmental toxicity studies have used Bayesian hierarchical models to improve estimates when measuring the response of multiple correlated endpoints tested with a single chemical (e.g. Faes et al., 2006; Choi et al., 2010) or used a multivariate model that assumes correlation in residuals for multiple health outcomes (Neelon and Dunson, 2004).

To incorporate dependence in the regression function between four HTS assays and eight nanomaterials, Patel et al. (2012) estimate dose-duration-response surfaces using linear B-splines with two internal knots in both the duration and dose direction. The first knot parameterizes potency with the no observable adverse effect level (NOAEL), an alternative measure to AC50. For each chemical, they model correlation in the knot location and basis coefficients across the assays, but they do not model correlation across chemicals within assay. While the direct parameterization of the NOAEL is appealing, this model does not directly parameterize the probability of a response or the efficacy, two important parameters for prioritization. In addition, the model does not include assay effects which we assume to exist in our data and can potentially improve fitting with the large number of chemicals but small sample size with each chemical-assay. While the simple choice of linear splines with two internal knots provides for a reasonable size model for a generalized additive model, this basis is not realistic for a one-dimensional model (dose only, not dose-duration surfaces).

In this paper, we propose a Bayesian hierarchical model for dose-response curves that is specifically tailored to the high-dimensional, sparse data setting of the ToxCast project, called the zero-inflated piecewise log-logistic model (ZIPLL). ZIPLL is a mixture between a non-active response and an active response that extends FPLL to a more flexible spline formulation while maintaining direct parameterization of the efficacy and potency of each chemical-assay combination. Our Bayesian approach naturally estimates the three key summary statistics and measures
of uncertainty for ranks of efficiency and potency which should allow for decision-makers to use the results appropriately when deciding which chemicals to consider for future, more comprehensive, testing. We use a hierarchical framework that borrows strength across chemicals and assays. This adds robustness, incorporates assay and chemical effects, and allows for estimation of joint distributions of responses across multiple assays. In addition, prior information and covariates can be included to exploit known relationships between chemicals, between assays, and between chemical-assay combinations.

2.2 The ToxCast Data

The ToxCast project uses a diverse battery of HTS assays and informatic models to rapidly characterize the activity of thousands of chemicals. These chemical activity profiles are used to support decisions regarding prioritization for further testing (Reif et al., 2010), predict in vivo activity (Martin et al., 2011), and inform risk assessments (Judson et al., 2011). In support of these goals, the ToxCast project has tested over 2,000 chemicals on over 700 HTS assay endpoints for which analysis is ongoing. The data for the first 309 chemicals tested for Phase I are publicly available (http://www.epa.gov/ncct/toxcast/data.html). Figure 2.2 illustrates the unique structure of the data.

In this paper we use the 309 chemicals in the publicly available data and a subset of 81 assays comprising the multiplexed transcription factor reporter platform (Romanov et al., 2008, www.attagene.com). This platform enables high-content, functional assessment of transcription factor activity, which is a core component of cellular gene regulatory networks. Both cis-regulating response element constructs (CIS) and trans-activating (TRANS) potential of multiple nuclear hormone receptors are measured. These 48 CIS and 25 TRANS assays (plus 8 negative control assays) address relevant cellular processes including response to xenobiotics, genotoxic stress, hypoxia, oxidative damage, immune-modulation, and endocrine disruption. Martin et al. (2010) evaluated these assays’ response to the 309 Phase I chemicals.

Chemicals were diluted in dimethyl sulfoxide (DMSO) at, in general, six to ten unique
Figure 2.2: Illustration of the ToxCast data structure showing 20 chemicals tested on 8 assays. This is a sample of the 309 chemicals and 81 assays used in Section 2.5 with curves as reported in the publicly available ToxCast data. All 309 chemicals are tested on all 81 assays, resulting in $309 \times 81 = 25,029$ chemical-assay combinations. This figure appears in color in the electronic version of this article.

concentrations on each HTS assay. The concentrations typically ranged from $0.046 \mu M$ to $100 \mu M$ or from $0.091 \mu M$ to $200 \mu M$ with each concentration three times the previous concentration. In cases of overt cytotoxicity, the concentration ranges were shifted up or down by a multiple of 3 in an attempt to recover the concentration range with a chance to show specific assay effects (Martin et al., 2010). Of the 309 Phase I chemicals four were tested in duplicate and one was tested in triplicate. The remaining chemicals were tested once at each concentration. The responses at each concentration are recorded in fold change over DMSO solution; hence, a response of 1 indicate no response. To reduce the inherent heteroskedasticity of the data, we log transformed the data before curve fitting, but return the data to the original scale before analyzing and plotting results.
2.3 Model Description

Our primary focus is understanding the relationship between the tested doses, \(x_{ijk}\), and the measured responses, \(y_{ijk}\), where \(i\) indexes chemical, \(j\) indexes assay, and \(k\) indexes tested concentrations within a chemical-assay combination. We assume a univariate Gaussian regression model \(y_{ijk} = f_{ij}(x_{ijk}) + \epsilon_{ijk}\) and \(\epsilon_{ijk} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)\).

The ZIPLL regression function is a mixture between an active and non-active response

\[
f(x_{ijk}; \theta_{ij}, w_{ij}, Z_{ij}) = \begin{cases} 
    t_{ij} - (t_{ij} - b_{ij}) \times \text{Logit}\{g(x_{ijk}; a_{ij}, w_{ij})\} & \text{if } Z_{ij} = 1 \\
    b_{ij} & \text{if } Z_{ij} = 0.
\end{cases}
\]  

In (A.1), \(\theta_{ij} = (t_{ij}, b_{ij}, a_{ij})\) parameterizes the Emax (upper asymptote), Emin (lower asymptote), and AC50 of the active response, \(w_{ij}\) controls the shape of an active response, and the latent \(Z_{ij}\) indicates if an active response occurred.

The function \(g(x; a, w)\) in (A.1) can be any monotone decreasing function in \(x\) with location parameter \(a\) and shape parameter \(w\), such that \(g(a; a, w) = 0\) for all \(w\). When \(t \geq b\), the active response is nondecreasing with upper and lower horizontal asymptotes \(t\) and \(b\), respectively.

The constraint that \(g(a; a, 0) = 0\) insures that the response is \((t + b)/2\) when \(x = a\). Hence, \(a\) is the AC50 of the active response. When \(g(x; a, w) = -w\{(\log(x) - \log(a))\}\) active responses follow FPLL (Figure 2.1a), we refer to this model as this is a zero-inflated log-logistic model (ZILL).

In ZIPLL, we replace the log-linear function with a piecewise log-linear spline to add robustness to misspecification of the shape of the dose-response curve. The new function is

\[
g(x_{ijk}; a_{ij}, w_{ij}) = \Psi(x_{ijk}; a_{ij})^T w_{ij} = \sum_{l} w_{ijl} \Psi_l(x_{ijk}, a_{ij}),
\]

where \(\Psi(x_{ijk}, a_{ij})\) is a vector of continuous linear basis functions with \(\Psi(a_{ij}, a_{ij}) = 0\) and \(w_{ij}\) is a \((p + 2)\)-vector of unknown basis coefficients.

We construct the linear basis using \(2p + 1\) internal knots with one knot at 0. We choose
symmetric fixed internal knots at \{-\xi_{p/2}, \ldots, -\xi_1, \xi_0, \xi_1, \ldots, \xi_{p/2}\}. The \(l\)th basis function is 

\[
\Psi_l(x_{ijk}, a_{ij}) = \begin{cases} 
\max \{ \log(x_{ijk}) - \log(a_{ij}), \xi_l \} - \xi_{l-1} & \text{if } l = -p/2 - 1, \ldots, -1 \\
\min \{ \log(x_{ijk}) - \log(a_{ij}), \xi_l \} - \xi_{l-1} & \text{if } l = 1, \ldots, p/2 + 1, 
\end{cases}
\]

with center knot \(\xi_0 = 0\) and external knots \(\xi_{-p/2 - 1} = -\infty\) and \(\xi_{p/2 + 1} = \infty\). Figure 2.1b shows a set of basis functions. This basis ensures a monotone nondecreasing response as long as each element of \(w\) is non-negative, and that \(a_{ij}\) is the AC50 (since \(\Psi(a_{ij}, a_{ij}) = 0\)). This basis has attractive limiting cases. As \(p \to \infty\) we get the full span of monotonic functions, \(g(x)\), through the origin. When the basis coefficients \((w_{ij,1}, \ldots, w_{ij,p+2})\) are all equal, ZIPLL reduces to ZILL, and when \(Z_{ij} = 1\) it further reduces to FPLL.

2.3.1 Hierarchical Structure and Prior Specification

We use a Bayesian approach to estimate the regression function and choose a prior that restricts the parameter space to increasing functions and induces a hierarchical structure between curves. ZIPLL is monotone nondecreasing when \(t_{ij} \geq b_{ij}\) and \(w_{ij} \geq 0\) and is identifiable if \(a_{ij} \in [x_{ij1}, x_{ijn_{ij}}]\). These constraints are met by introducing unconstrained latent parameters \(\theta_{ij}^* = (t_{ij}^*, b_{ij}, a_{ij}^*)^T\) and \(w_{ij}^*\), and mapping these unconstrained parameters to their constrained counterparts by \(a_{ij} = \min(x_{ij}) + \{\max(x_{ij}) - \min(x_{ij})\}/\{1 + \exp(-a_{ij}^*)\}\), \(t_{ij} = \max(t_{ij}^*, b_{ij})\), and \(w_{ij} = \exp(w_{ij}^*)\). This formulation allows \(\theta_{ij}\) and \(w_{ij}\) to conform with the restricted parameter space but \(\theta_{ij}^*\) and \(w_{ij}^*\) to take any real values.

The chemical-assay specific parameters have normal priors

\[
\theta_{ij}^* \ iid \sim N_3(\mu, \Sigma) \\
w_{ij}^* \ iid \sim N_{p+2}(S, \Sigma_s).
\]

To encourage smoothness in the slopes, we put an autoregressive hyperprior on \(S\) similar to Neelon and Dunson (2004). We let \(S_0\) be the a priori belief of the average slope and \(S_k \sim \ldots\)
\(N(S_{k-1}, \lambda^{-1})\) for \(k = 1, \ldots, p + 2\). To allow for uncertainty in the smoothness of \(S\) we put a Gamma\((g_1, g_2)\) hyperprior on \(\lambda\).

### 2.3.2 Assay Effects, Chemical Effects, and Prior Knowledge

It is reasonable to expect that some assays may be more or less sensitive than others and in some cases practitioners may want to incorporate prior knowledge about chemicals and assays such as covariates or known groups of similar chemicals and assays. For example, in the ToxCast data Martin et al. (2010) reported that the number of chemicals active on each assay ranged from 0 to 225 and the expected range of potencies and efficacies varied between assays. This information can be modeled as either fixed or random effects in the prior mean of \(\theta^*_{ij}\) or the prior mean of \(Z_{ij}, \psi_{ij} = \text{Pr}(Z_{ij} = 1)\). We include both the assay level random effects and a probit model on the \(\psi_{ij}\) in our analysis in Section 2.5. To account for between assay differences in efficacy and potency we put a random effects model on \(\theta^*_{ij}\),

\[
\theta^*_{ij} | \mu_j, \Sigma_j \sim N_3(\mu_j, \Sigma_j) \\
\mu_j \overset{iid}{\sim} N_3(\mu, \Sigma).
\]

For the analysis of ToxCast data in Section 2.5 we fit the data with log transformed responses to reduce heterogeneity. Our prior reflects strong confidence that the Emin will be near 1 (or 0 on log scale), the baseline response for DMSO solution, but allows more freedom in the other three parameters. We assume \(\mu\) is normal with mean \((2, 0, -3)^T\) and variance \(\text{diag}(3, 1, 3)\). The prior for \(\Sigma\) is inverse Wishart prior with scale parameter 6 and shape parameter \(\text{diag}(60, 2, 60)\).

We also include the chemical level covariate LogP, the log of the partition coefficient. LogP is a measure of solubility and relates to a chemicals’ ability to permeate a membrane, a prerequisite to a cellular response. LogP was calculated using Leadscope (Leadscope Inc., Columbus, OH). We use LogP to model the prior for \(\psi_{ij}\) with a linear fixed effect and a probit link. To account for the varying sensitivity of the 81 assays, we include an assay level random intercept, \(\psi_{ij} = \ldots\)
\( \Phi(\delta_j + \delta \log P_i) \). The hyperpriors these parameters used to fit the ToxCast data are \( \delta_0 \sim N(0, 10^2) \) and \( \delta_j \sim N(\delta_0, \sigma_{\delta}^2) \) with \( \delta_0 \sim N(0, 10^2) \) and \( \sigma_{\delta}^{-2} \sim \text{Gamma}(0.1, 0.1) \). The remaining hyperpriors are \( \sigma^{-2} \sim \text{Gamma}(1, 0.5) \), \( S_0 = 2 \) and \( \lambda \sim \text{Gamma}(1, .5) \).

### 2.3.3 Posterior Computation

Our MCMC algorithm is a hybrid Gibbs and Metropolis-Hastings sampler. The full conditional posterior distributions of \( \boldsymbol{\mu}, \boldsymbol{\Sigma}, \sigma^{-2}, Z_{ij}, \psi, \lambda, \boldsymbol{\Sigma}_s, \) and \( \boldsymbol{S} \) have simple conjugate forms. The full conditional distributions for \( t^*_{ij} \) and \( b_{ij} \) are both mixtures of truncated normals. All full conditionals are detailed in Appendix A.

The remaining parameters, \( a^*_{ij} \) and \( w^*_{ij} \), do not have closed form posterior distributions. To reduce autocorrelation we use a resolvant transition kernel based on the Metropolis-Hastings kernel (Robert and Casella, 2004). We provide details on sampling these parameters in Appendix A. An R package to implement ZIPLL is provided in the supplemental material.

MCMC sampling returns posterior samples for \( \boldsymbol{\theta}_{ij} \) and \( Z_{ij} \) which provide the estimates of potency \( (a_{ij}) \), efficacy \( (t_{ij}) \), and activity. The posterior probability of an active response is the proportion of samples with \( t_{ij} > b_{ij} + \kappa \) and \( Z_{ij} = 1 \). We assume that the minimum clinically important response is a one fold increase a baseline measure by using \( \kappa = 1 \). This is the same assumption used in previous ToxCast analyses (Martin et al., 2010). We use this “clinically important” definition to define active responses in Section 2.5 in order to be consistent with the current ToxCast practices.

This algorithm performed well on our simulated and real data and by using the resolvant kernel there is a reasonably small level of autocorrelation. For the full 309 chemicals and 81 assays we ran the chain for 50,000 iterations and discarded the first 20,000 as burnin. The smaller simulation with 100 curves was run for 20,000 iterations with 5,000 discarded for burnin. We assessed convergence by inspecting trace plots, and comparing multiple chains.

MCMC sampling is carried out in C called from R (R Development Core Team, 2011) with .C. Runtime for simulated data set of 100 curves of eight observations for 20,000 iterations is
42 seconds with ZIPLL. Analysis of all 309 chemicals and 81 assays with 50,000 iterations of ZIPLL including random assay effects and probit model for covariates as specified in Section 2.5 took 19.9 hours. Both computation times are on a DELL Dual Processor Xeon Six Core 3.6 GHz machine with 60GB RAM.

2.4 Model Comparisons

2.4.1 Simulation Study

To evaluate the performance of ZIPLL we conducted a simulation study based on the design of the ToxCast data. We summarize the simulation results here and provide a full discussion in Appendix A. We fit ZIPLL, ZILL, FPLL using nonlinear least squares, and Bayesian, monotone, piecewise linear spline model proposed by Neelon and Dunson (2004) that fits each chemical-assay combination separately.

When data was simulated from ZILL, the four methods performed similarly, with the hierarchical methods (ZIPLL and ZILL) slightly outperforming the others with respect to pointwise root mean square error (RMSE). When we used an asymmetric response pattern that violated the assumptions of FPLL and ZILL, ZIPLL had smaller pointwise RMSE as well as RMSE on the AC50 and Emax. ZIPLL also had better credible interval coverage with smaller interval widths. Finally, the three Bayesian methods estimated active responses with high probability, while FPLL did not.

2.4.2 Cross Validation and Model Fit

To determine if ZIPLL provides a better fit specifically for the ToxCast data we performed a cross validation study and compared several model fit statistics for the four methods described in Section 2.4.1 fit to the ToxCast data for 309 chemicals and 81 assays. For ZIPLL, we included the full random effects model and probit model on the probability of response. We determined that one interior knot at 0, two linear segments, performed best based on the cross validation
predictive MSE. For cross validation we removed one observation from each chemical-assay combination, fit the remaining data, and predicted the removed response. This was repeated 8 times, leaving out a different observation each time.

ZIPLL had the lower predictive MSE, 0.2177 compared to 0.2200 for ZILL, 0.3041 for monotone, linear splines, and 0.3080 for LS. Whereas ZIPLL had noticeable advantages over ZILL in the simulation, the two methods were similar in cross validation; however, this small improvement is significant with a paired $t$-test. We also compared the three Bayesian methods using DIC (Spiegelhalter et al., 2002), log psuedo-marginal likelihood (LPML) (Geisser and Eddy, 1979), continuous ranked probability score (CRPS) (Matheson and Winkler, 1976), and the method of Gelfand and Ghosh (1998) using equal weights for the two components. In each case, ZIPLL outperformed ZILL by a small margin. Therefore, we present the results below assuming the ZIPLL model.

2.5 ToxCast Data Application

We fit the 309 chemicals and 81 assays with ZIPLL, including assay random effects and probit model as specified in Section 2.3.2. Figure 2.3 shows the dose-response estimates for 12 chemicals on pregnane X receptor response element (PXRE) fit with ZIPLL and the reported fits from the ToxCast public use files. The first row shows three chemicals where the ZIPLL posterior mean is similar to the FPLL fits reported in ToxCast. The second and third rows show chemicals where ZIPLL better fits the data by adapting to an asymmetric response pattern.

The bottom row of Figure 2.3 highlights the importance of probabilistic estimation of an active response. The three chemicals shown have similar response patterns; however, using the current ToxCast methodology one is marked active, having increased by at least one fold change, on PXRE while the other two are not. With ZIPLL, the estimated probabilities of response are between 0.26 and 0.87. Appendix A Figure 2 compares the ZIPLL probability with ToxCast indicator for all 309 chemicals on three assays. The majority of chemicals are considered not active or active with both methods. However, using ZIPLL we estimated that
Figure 2.3: Example of ZIPLL and ToxCast estimates for 12 chemicals on the PXRE assay. The ZIPLL posterior mean (thick solid black line) with 95% posterior intervals (dashed black lines), and the ToxCast fit (thin solid gray line) are shown. The legend shows a binary indicator of an active response from ToxCast (1=active) and the posterior probability of an active response using ZIPLL. The FPLL fits are redrawn from the parameter estimated in the ToxCast public release data. This figure appears in color in the electronic version of this article.
several chemicals have posterior probabilities of response between 0.1 and 0.9, suggesting that there is not conclusive evidence that these chemical responded or not, but are forced to be classified as either active or not in the ToxCast data. The set of chemicals having high non-zero posterior probabilities of response via ZIPLL yet a ToxCast call of no response include several with evidence of PXR activity from other ToxCast assays (e.g. Flumiclorac-pentyl) and/or independent structure-activity models (e.g. Butafenacil) (Kortagere et al., 2010).

### 2.5.1 Summary of Active Responses and Assay and Chemical Effects

A natural result of the hierarchical analysis is estimation of the joint distribution of responses across assays. Figure 2.4a shows the number of active assay responses for each chemical. The number of assay responses reported in ToxCast tends to be around the lower bound of the ZIPLL posterior interval and is similar to the number of assay responses estimated if we consider anything with a ZIPLL posterior probability of 0.75 to be active. Overall, there are 2667 (2616, 2718) active assay-chemical combinations estimated with ZIPLL compared to 1887 reported in ToxCast. This suggests that some assay responses may be missed using the current ToxCast methods, potentially hindering prioritization efforts.

Figure 2.4b shows the posterior of the assay random intercept for the probit model of probability of response. The most and least sensitive assays had statistically significant random effects. At the lower end, the eight assays with the prefix “M” are negative controls and all had effects around -10, while more potent assays like PXRE and PPARγ had large positive effects. The posterior mean of the coefficient for LogP is -0.0005, and this effect was not significant. This may be due to selection bias. Solubility (low logP) was part of the selection criteria for the first 309 chemicals in order to accommodate solubility in Dimethyl sulfoxide (DMSO); however, this restriction was relaxed for chemicals included in forthcoming ToxCast phases, so LogP may prove to be an important factor in future samples.

The simultaneous fitting of all chemicals allows for simple rankings of chemical by potency as well as a measure of uncertainty in the rankings. Figure 2.5 shows the ranking of chemicals
Figure 2.4: Panel (a) plots the number of assays each chemical responded on with 95% posterior intervals. The simultaneous fitting of all chemical-assay combinations allows for the estimation of the joint distribution of assays and naturally propagates the distribution of the total number of assay responses. The x indicates the number of assay responses reported in the ToxCast data. (The 309 chemical names are omitted on the horizontal axis for readability.) Panel (b) shows posterior mean random intercept for the assay probability of response ($\delta_j$) as discussed in Section 2.3.2 and 95% posterior intervals. This figure appears in color in the electronic version of this article.
by posterior mean AC50 on three assays: PXRE, peroxisome proliferator-activated receptor 
γ (PPARγ), and estrogen receptor α (ERα). Among chemicals with at least a 0.5 posterior 
probability of being active, the mean AC50 was 32.7 for PXRE, 64.5 for PPARγ, and 50.6 for 
ERα. This supports the inclusion of an assay random effect in the model.

### 2.5.2 Comparison with Reference Chemicals

A useful way to summarize the results for each assay is to compare chemicals with reference 
chemicals known to be active on the assay. For example, PPARγ is a commonly used assay 
that has a plausible connection with neoplastic pathology (see Peters et al., 1997; Aoki, 2007). 
Figure 2.5b highlights the response of four reference chemicals for PPARγ: perfluorooctane 
sulfonic acid (PFOS), Diethylhexyl phthalate (DEHP), Phthalic acid, mono-2-ethylhexyl ester 
(PAMEHP), and perfluorooctanoic acid (PFOA) (Casals-Casas and Desvergne, 2011). Because 
the reference chemicals have known biological effects, other chemicals with a high probability of 
being more potent than the reference chemicals on a given assay may have greater potential for 
similar biological effects to the reference chemicals, and thus may be higher priority candidates 
for additional testing than chemicals that are not as potent as the reference chemicals. The four 
reference chemicals’ posterior mean potencies rank (with 1 being the most potent) 42.9 (35.0, 
49.0), 105.7 (57.0,154.0), 114.0 (68.0,156.0), and 122.6 (72.0,159.0), respectively, on this assay 
among 161 chemicals with at least 0.5 probability of activity, indicating there are many good 
candidates for further testing.

Another commonly studied assay is ERα. Figure 2.5c shows results for ERα with reference 
chemicals Bisphenol A (BPA) and Methoxychlor highlighted. These two reference chemicals 
have mean posterior rank 1 (1.0,1.0) and 7.4 (4.0,10.0), respectively, among the 103 chemicals 
with posterior probability of an active response of 0.5 or more. This implies there is at least 
0.95 probability that BPA is the most potent chemical among the 309 and very few ToxCast 
chemicals are more potent than Methoxychlor.

These rankings allow us to estimate the posterior probability that chemicals are more active
Figure 2.5: Chemical ranks by potency with 90% credible intervals. The x-axis shows the posterior mean AC50 (more potent to the left). The y-axis ranks the chemicals by potency (most potent, rank one, at top). All chemicals with at least a 0.50 percent probability of active response are plotted. For PPARγ and ERα reference chemicals are marked for comparison. These chemicals have shown documented activity on these assays. This figure appears in color in the electronic version of this article.

than the reference chemicals both marginally for each assay and jointly across assays. The ability to estimate this probability jointly across biologically related assays provides an important capability—pathway based prioritization. For assays measuring distinct biological targets, such as ERα and PPARγ, no chemical had more than a trivial posterior probability of being more active than the reference chemicals on both, which is expected. As an example of prioritization based upon single assays, Figure 2.6 shows the 66 chemicals with at least a 0.05 posterior probability of being more potent than the four PPAR-specific reference chemicals on PPARγ. With the more diverse set of reference chemicals available in the forthcoming ToxCast data, comparisons across several assays will be feasible and can provide a probabilistic ranking of chemicals based on the potential for bioactivity on these pathways.
2.6 Discussion

In this paper we propose a new model for estimating the dose-response relationship for HTS data when a large number of chemicals are tested across several assays. ZIPLL directly parametrizes the AC50 and Emax. As a result, the efficacy and potency are easily interpretable. Through our simulation study, we demonstrated that ZIPLL accurately estimates the AC50, Emax, and the probability of response curve. Further, ZIPLL is robust to assumptions about the shape of the response. Overall, this hierarchical approach to analyzing HTS data outperformed methods that treat each curve as independent and ignore correlation between assays.

Our proposed MCMC algorithm takes about 20 hours to fit 309 chemicals on 81 assays, longer than the published ToxCast method. However, for HTS projects like ToxCast, data are analyzed in large batches, so real-time updates are not necessary. As a result, emphasis is on model performance over efficient computation. In the case that a small number of chemicals were added, all hyperparameters could be fixed based on the full run and the posterior computed for the new chemicals in a few minutes. For larger batches, computation time for ZIPLL scales linearly for both the number of chemicals and number of assays, making runs on larger...
We applied ZIPLL to the ToxCast data and showed that the probabilities of response were largely consistent with the binary classification in the ToxCast public release data. However, in borderline cases ZIPLL added useful information by quantifying the uncertainty in the presence of a response. We also demonstrated the advantage of estimating the posterior distribution of the AC50. This allowed us to rank chemicals and estimate the posterior probability that a chemical is more potent than reference chemicals, which provides a useful tool for prioritization.

Ultimately, a comprehensive risk assessment must include not only coverage of all relevant exposure and hazard factors, but thorough characterization of individual factors as well. The dose-response model provided by ZIPLL will prove especially useful in such a scenario, where the more informative results characterize HTS hazard in a manner that can be quantitatively combined with other risk factors. With the addition of data from future HTS projects having expanded assay coverage and reference chemical sets, these rankings can be extended to estimate the joint probability that chemicals are more active than reference chemicals on multiple assays, thus providing a physiologically relevant, pathway-based hazard assessment.

### 2.7 Software

The R package ZIPLL (Wilson et al., 2013) implements the ZIPLL model and is available at http://www4.stat.ncsu.edu/~tawilso3/software.html.
Chapter 3

Modeling the Effect of Temperature on Ozone-Related Mortality

The US Environmental Protection Agency (EPA) has concluded that current scientific evidence supports a “causal relationship” between ozone and respiratory health effects and a “likely to be causal” relationship between ozone and cardiovascular health effects and mortality (US EPA, 2013). Extreme temperatures, especially heat waves, have also shown adverse associations with respiratory and cardiovascular health (Bhaskaran et al., 2009; Turner et al., 2012). As a photochemical air pollutant, ozone and temperature are both driven by solar radiation and their association is enhanced by the temperature dependence of other ozone precursors (Bloomer et al., 2009). Both ambient temperatures and ground-level ozone are expected to increase in the near future (US EPA, 2009) in response to anticipated climate changes (IPCC, 2007). Under these circumstances, a better understanding of the joint effects of temperature and ozone on human health is essential for public health.

Most multi-city, time-series studies of ozone have focused on a main ozone effect while controlling for potential confounding variables in a generalized additive model. The highly influential analysis of Bell et al. (2004) estimated a 0.25% (0.12%-0.39%) increase in mortality associated with a 10 ppb increase in same-day mean ozone at the national level while controlling
for confounders such as temperature and weather conditions. Recent studies have found evidence that the joint effects of ozone and temperature may not be additive (Bell et al., 2006; Smith et al., 2009; Chen et al., 2013). Future climate change scenarios predict increases in the high ends of the distributions of both the ozone and temperature. Under these future conditions, the disease burden of high ozone and temperature days may be different if the joint ozone-temperature effect is synergistic instead of additive.

Estimating the existence and nature of an interactive effect is non-trivial as temperature and ozone are both correlated with health outcomes and with each other. In multi-city, time-series studies the analysis is further complicated by the different ozone and temperature ranges observed in each city which makes pooling estimates across cities challenging. Several studies of ozone-related daily mortality have used a stratified model to examine the potential differences in ozone effect by temperature and found a larger average ozone effect on high temperature days compared to moderate temperature days (Bell and Dominici, 2008; Ren et al., 2008; Smith et al., 2009). However, average ozone levels are higher on high temperature days (see Figure 3.1 for example and Appendix B Figure 1 for additional details) and the ozone effect may be larger at higher ozone concentrations (Smith et al., 2009). This makes it unclear if the larger average ozone effect on high temperature days is due to the higher ozone on those days or an ozone-temperature interaction.

To estimate interaction, the concentration-response gradient for ozone must be evaluated at different temperature ranges for the same ozone range. Matching on ozone isolates interaction from the potential confounding caused by higher ozone levels at higher temperatures and non-linear ozone effects. One way to do this is to estimate the full ozone-temperature risk surface and evaluate the rate of change of the risk surface in the ozone direction at different temperatures for a constant ozone value. The gradient of the risk surface with respect to ozone describes the ozone effect for each temperature and ozone concentration. A changing gradient as a function of temperature for a fixed ozone value signifies interaction between ozone and temperature. We will refer to the gradient in the ozone direction as the log relative risk (log RR) of ozone.
Figure 3.1: Ozone-temperature distribution for selected cities. The upper and lower boxes contain data for high temperature days and moderate temperature days, respectively, over the observed ozone range. The purple subsections highlight the common ozone range, the intersection of the ozone ranges for high and moderate temperature days. Details on the definition used to identify these days are in Section 3.3.3. Seattle is the only city in the dataset that does not have a common ozone range.
throughout, which approximates excess RR.

Recent studies have estimated the ozone-temperature risk surface in single-city and independently in multi-city analyses (Chen et al., 2013; Ren et al., 2008). The risk surfaces appear to be non-linear and non-additive, but inference was limited to visual inspection. Instead, these studies resorted to a stratified model with different linear effects in each temperature stratum for a formal analysis instead of making quantitative inference from the risk surfaces. In addition, the multi-city approach used by Ren et al. (2008) estimates the risk surfaces independently in each city, not allowing for sharing of information between cities. The resulting risk surface estimates are highly variable and do not allow for combining city-specific surfaces to estimate a national risk surface.

In this paper, we propose a spatial monotone surface model to estimate the ozone-temperature risk and log RR surfaces in multi-city time-series studies. We model the ozone-temperature risk surface with the outer-product of Bernstein polynomial basis functions and restrict the surfaces to be monotone in ozone. The Bernstein polynomial formulation enables closed-form representation of the risk surfaces and log RR surfaces and facilitates comparisons between cities. It also allows for sharing strength between nearby cities with a spatial model for the basis coefficients.

We also present a two-stage approach to this model that reduces its computational burden for large data (following the framework of Dominici et al., 2002). In the two-stage approach, we introduce a transformation that allows a different local basis expansion to be used in each city that adapts to the city-specific ozone and temperature ranges. This allows for the first-stage surface estimates to be estimated with a basis expansion specifically tailored to the ozone-temperature distribution in each city. However, the second stage surfaces use a common basis expansion spanning the national ozone-temperature distribution.

We use the proposed spatial monotone surface model to estimate the national and city-specific log RR surfaces in 95 US urban areas using data from the National Morbidity and Mortality Air Pollution Study (NMMAPS; Samet et al., 2000a,b). Our results show evidence of a synergistic ozone-temperature effect. At higher temperatures and ozone values, the log
RR of ozone tends to increase with temperature. We find the ozone-temperature interaction increases estimates of excess mortality at higher temperature.

3.1 Spatial monotone surface model

We assume the mortality count $Y_{ct}$ in city $c$ at time $t$ is Poisson with log mean

$$\log E(Y_{ct}) = f_c(\text{ozone}_{ct}, \text{temp}_{ct}) + g_c(\text{confounders}_{ct})$$ (3.1)

where $f_c$ is the city-specific ozone-temperature risk surface, and $g_c$ controls for potential confounding variables. The model for daily mortality is adopted from Bell et al. (2004) by moving daily mean temperature from the confounder model $g_c$ to the risk model $f_c$ and relaxing the assumptions on $f_c$ to include interaction and non-linear effects. The bivariate risk surface $f_c$ is modeled with a spatial prior that imposes monotonicity in the ozone direction. The confounder model $g_c$ includes linear and non-linear functions of potential confounders. In the subsections below we specify models for the components of (3.1).

3.1.1 Ozone-temperature surface model

We model $f_c$ as the outer-product of Bernstein polynomial basis expansions of ozone and temperature (Lorentz, 1986; Tenbusch, 1997). This allows for a flexible regression surface including non-linear and non-additive effects, but includes additive linear or polynomial effects as special cases. In addition, both the Bernstein polynomial regression function (the log risk surface) and its derivatives (the log RR surface) can be expressed in closed form. Closed-form representation of the derivative facilitates analysis of the log RR of ozone.

The $k$th Bernstein polynomial basis function of order $M$ is $b_k(x, M) = \binom{M}{k}x^k(1-x)^{M-k}$ for $x \in [0, 1]$. Denote ozone as $x_1$, temperature as $x_2$, and $\mathbf{x} = (x_1, x_2)^T$. To scale the data to the unit interval we define the basis function $B_{l,k}(x_l, M) = b_k[(x_l - \min_{ct} x_{l,ct})/r_l, M]$, where $r_l = \max_{ct} x_{l,ct} - \min_{ct} x_{l,ct}$ and $l = 1, 2$ indicates ozone and temperature, respectively. For
notational simplicity let $B_{j,k}(x, M_1, M_2) = B_{1,j}(x_1, M_1) \times B_{2,k}(x_2, M_2)$. The bivariate regression function is

$$f_c(x) = \sum_{j=0}^{M_1} \sum_{k=0}^{M_2} \psi_{j,k,c} B_{j,k}(x, M_1, M_2),$$

(3.2)

where $j$ and $k$ index the ozone and temperature basis expansions, respectively. The first derivative of (3.2) with respect to ozone is

$$\frac{\partial f_c(x)}{\partial x_1} = M_1 \sum_{j=0}^{M_1-1} \sum_{k=0}^{M_2} (\psi_{j+1,k,c} - \psi_{j,k,c}) B_{j,k}(x, M_1-1, M_2),$$

(3.3)

the log RR of ozone.

For simplicity, we write the $(M_1 + 1)(M_2 + 1)$-vector of unknown coefficients as

$$\psi_c = \left( \psi_{0,0,c}, \ldots, \psi_{M_1,0,c}, \psi_{0,1,c}, \ldots, \psi_{M_1,M_2,c} \right)^T.$$  

(3.4)

Also, denote the $n_c \times (M_1 + 1)(M_2 + 1)$ basis expansion of ozone and temperature in city $c$ over days $t = 1, \ldots, n_c$ as

$$B(X_c) = \left[ B_{0,0}(X_{c,M_1,M_2}), \ldots, B_{M_1,M_2}(X_{c,M_1,M_2}) \right]^T.$$  

(3.5)

### 3.1.2 Hierarchical model for monotonicity and spatial smoothing

Evidence from clinical research of lung function suggests that ozone-related risk is monotonic increasing (McDonnell et al., 2012) while temperature is often described to have an “inverted J” or “U-shaped” risk function (Curriero et al., 2002). To reflect this prior knowledge we constrain the ozone effect to be monotone, but leave the temperature effect unconstrained.

Bernstein polynomials are well suited for shape-restricted regression (see Chang et al., 2007; Curtis and Ghosh, 2011b; Wang and Ghosh, 2012). From (3.3), a sufficient condition for monotonicity in the ozone direction is $\psi_{j+1,k,c} \geq \psi_{j,k,c}$ for all $j$ and $k$. We reparameterize the
coefficients as $\theta_{0,k,c} = \psi_{0,k,c}$ and $\theta_{j,k,c} = \psi_{j+1,k,c} - \psi_{j,k,c}$ for $j > 0$ using the matrix

$$
\mathbf{T} = \mathbf{I}_{M_2+1} \otimes \begin{pmatrix}
1 & 0 & 0 & \ldots & 0 & 0 \\
-1 & 1 & 0 & \ldots & 0 & 0 \\
0 & -1 & 1 & \ldots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \ldots & -1 & 1 \\
\end{pmatrix}_{(M_1+1) \times (M_1+1)}.
$$

(3.6)

With this parameterization $\theta_c = \mathbf{T} \psi_c$ and $f(\mathbf{x}; \theta_c)$ is monotone in $x_1$ if $\theta_{j,k} \geq 0$ for $j > 0$.

The ozone-temperature surface can vary across cities. To borrow strength across nearby cities while ensuring a monotone risk surface we model the basis coefficients with a truncated multivariate Gaussian process (GP). To do this, we define the latent vector $\theta^*_c$ and let $\theta_{0,k,c} = \theta^*_{0,k,c}$ and $\theta_{j,k,c} = \max(0, \theta^*_{j,k,c})$ if $j > 0$ for all $k$.

The prior on $\theta^*_c$ is a multivariate Gaussian process with mean $\mathbb{E}(\theta^*_c) = \mu$ and separable covariance function

$$
\text{cov}(\theta^*_c, \theta^*_c) = \exp \left[ -\frac{d(c, c')}{\rho} \right] \times \mathbf{S}_2 \otimes \mathbf{S}_1
$$

(3.7)

where $\mathbf{S}_1$ is a $(M_1+1) \times (M_1+1)$ matrix capturing covariance in the ozone direction, $\mathbf{S}_2$ is a $(M_2+1) \times (M_2+1)$ matrix capturing the covariance in the temperature direction, and the exponential function captures spatial dependence. The distance function $d(c, c')$ is the great circle distance between cities $c$ and $c'$ in kilometers. The mean vector $\mu$ has prior $N(1 \mu_0, \tau \mathbf{S}_2 \otimes \mathbf{S}_1)$. While each component of the separable covariance is not identifiable on its own, the product is identifiable.

### 3.1.3 Confounder model

We define $g_c$ as a generalized additive model that includes linear and non-linear effects for potential confounders. This is the confounder model used by Bell et al. (2004), with the same degrees of freedom, excluding daily mean temperature which is now in $f_c$. The confounder model includes an age-specific intercept (<65, 65-74, $\geq$75), categorical variables for day of
week, smooth functions of time interacted with age group with seven degrees of freedom per year (natural splines), natural cubic spline of the running mean of temperature with six degrees of freedom, natural cubic spline of dewpoint with 3 degrees of freedom and of the running mean of dewpoint with 3 degrees of freedom. The confounder model can be represented as the linear model \( g_c(Z_c) = Z_c \gamma_c \). The prior for the confounder regression coefficients is \( \pi(\gamma_c) \propto 1 \). While the prior for the risk surface \( f_c \) includes a spatial component, the prior on the confounder model is independent between cities.

### 3.2 A two-stage approach for large datasets

For large datasets, estimating the ozone-temperature risk surfaces is computationally intensive. To ease computation, we break the model into two stages, similar to the approach used by Dominici et al. (2002) and Bell et al. (2004). The two-stage approach approximates the spatial monotone model presented in Section 3.1 but with reduced computational burden. In the first stage, we estimate (3.1) separately in each city with no monotonicity constraint or spatial smoothing using computationally efficient quasi-likelihood estimation. Then, we use the first-stage parameter estimates as data for a Bayesian hierarchical model that spatially smooths the risk surfaces and constrains the estimates to be monotone. The results given in Section 3.3 are produced using this two-stage approach.

A natural approach for the city-specific estimates is to use the same basis expansion in each city. However, the observed ozone and temperature ranges vary dramatically from city to city. As a result, some basis functions are well supported at the national level but not supported in some cities individually. This can yield unstable first-stage estimates. To extract as much information as possible from each city in the first stage we use a local basis expansion in each city that spans only the observed ozone and temperature range in that city. In the second stage a common basis expansion is used for all cities to approximate the full model described in Section 3.1. Appendix B Figure 2 shows the first and second stage basis expansions for six cities.
3.2.1 Stage 1: City-specific GLM regression

The first-stage basis expansions are scaled to the city-specific ozone and temperature ranges. In city \( c \) the first-stage basis functions are

\[
b_{c,l,k}(x_l, M^c) = b_k[(x_l - \min_l x_{l,ct})/r^c_l, M^c],
\]

where \( r^c_l = \max_l x_{l,ct} - \min_l x_{l,ct} \), for \( l = 1, 2 \). The first-stage estimates can also vary in the order of the basis expansion, allowing for smaller \( M_1^c \) and \( M_2^c \) in cities with smaller ozone and temperature ranges, respectively. Using this basis expansion the first-stage model of \( f_c \) is

\[
f_c(x) = \sum_{j=0}^{M_1^c} \sum_{k=0}^{M_2^c} \beta_{c,j,k} b_{1,j}^c(x_1, M_1^c) b_{2,k}^c(x_2, M_2^c), \tag{3.8}
\]

with unknown first-stage parameters \( \beta_{c,j,k}, j = 0, \ldots, M_1^c \) and \( k = 0, \ldots, M_2^c \). The confounder model remains as specified in Section 3.1.3. Denote the first-stage estimate as \( (\hat{\beta}_c^T, \hat{\gamma}_c^T)^T \) and \n\[
\text{cov}[(\hat{\beta}_c^T, \hat{\gamma}_c^T)^T] = V_c \text{ where } V_c \text{ can be partitioned as}
\]

\[
V_c = \begin{bmatrix}
V_{c,11} & V_{c,12} \\
V_{c,21} & V_{c,22}
\end{bmatrix}, \tag{3.9}
\]

and \( b_c(X_c) \) is the \( n_c \times (M_1^c + 1)(M_2^c + 1) \) matrix of first-stage basis expansions.

The first-stage parameters \( \beta_c \) correspond to different basis functions than those used in other cities and in the global expansion in (3.2). Hence, \( \beta_c \) are not directly comparable across cities or with \( \psi_c \). This difference is resolved in the second stage (Section 3.2.2).

3.2.2 Stage 2: Bayesian model for stage 1 output

In the second stage, we reparameterize the first-stage risk surface estimates in terms of the global basis expansion (3.2). This provides a common set of parameters \( \theta_c \) to estimate \( f_c, c = 1, \ldots, n \), with our spatial monotone model described in Section 3.1.2. Setting the first-stage parameterization of the risk function \( f_c(X_c) = b_c(X_c)\beta_c \) equal to the second-stage parameter-
\begin{equation}
A_c = \left[ b_c^T(X_c)b_c(X_c) \right]^{-1} b_c^T(X_c)B(X_c)T^{-1}.
\end{equation}

The quantity $A_c\theta_c$ is the projection of the second-stage log risk surface onto the column space of the first-stage basis expansion.

Most two-stage approaches use the likelihood $\hat{\beta}_c \sim N(\beta_c, V_{c,11})$ in the second stage (e.g. Dominici et al., 2002; Bell et al., 2004). For our model $\beta_c$ does not have the same meaning in each city. We instead replace $\beta_c$ with $A_c\theta_c$ and use the second-stage likelihood

\begin{equation}
\begin{bmatrix}
\hat{\beta}_c \\
\hat{\gamma}_c
\end{bmatrix}
\mid \theta_c, \gamma_c, \tilde{V}_c 
\sim N
\begin{bmatrix}
A_c\theta_c \\
\gamma_c
\end{bmatrix}, \tilde{V}_c
\end{equation}

To complete the second-stage model we put a flat prior on $\gamma$ and use the prior model described in Section 3.1.2 for $\theta^\ast$ (and thus for $\theta$).

3.2.3 Priors and computational details

We estimate (3.1) with quasi-likelihood methods and the \texttt{glm} function in \texttt{R}. We use an offset proportional to the log population to account for different population sizes. To complete the Bayesian specification of the second-stage model we use the hyperpriors:

\begin{equation}
\begin{align*}
\mu_0 &\sim N(0, \tau_0^2) \\
S_2 &\sim \text{IW}(M_2 + 2, I), \\
S_1 &\sim \text{IW}(M_1 + 2, I), \\
\tau &\sim \text{Ga}(a_{\tau}, b_{\tau}), \text{ and} \\
\log(\rho) &\sim N(\mu_\rho, \sigma_\rho^2).
\end{align*}
\end{equation}

We are interested in the posterior of $(\theta_c^T, \gamma_c^T)^T$. To expedite computation we perform MCMC
sampling on the marginal posterior of $\theta_c$, marginalizing over $\gamma_c$ which is immediate from (3.11).

The parameters $\mu_c$, $\mu_0$, $\tau_c$, $S_1$, and $S_2$ have simple conjugate forms and are updated with Gibbs sampling. The latent $\theta^*_c$ are sampled with a Gibbs sampler using a mixture of truncated normals. The range does not have a closed-form full conditional. We sample $\log(\rho)$ with random-walk Metropolis-Hastings sampler. The full conditional for all parameters, acceptance ratio for $\rho$, and MCMC algorithm are provided in Appendix B.

To get the predicted values used for cross-validation in Section 3.3.1 we need the posterior mean of $\gamma_c$,

$$E\left(\gamma_c|\theta_c, \hat{\beta}_c, V_c\right) = \hat{\gamma}_c + V_{c,21}V^{-1}_{c,11}\left(A_c\hat{\theta}_c - \hat{\beta}_c\right),$$

where $\hat{\theta}_c$ is the posterior mean of $\theta_c$. Hence, no MCMC is required to get the posterior mean of $\gamma_c$, rather, it can be computed in closed form using the marginal posterior estimate.

### 3.3 Analysis of the ozone-temperature log RR surfaces

In this section, we estimate the city-specific and national average log RR surfaces using the two-stage approach presented in Section 3.2 and examine the nature of the ozone effect at different temperatures and ozone levels. We use the NMMAPS data and estimate the risk surfaces using same-day 1-hour maximum ozone and mean temperature, and estimate the surfaces for the same 95 US urban areas used in Bell et al. (2004). The NMMAPS data contains time-series data for 1987 through 2000 with daily mortality counts by age group and daily measurements of ozone, meteorological conditions, and co-pollutants. We limit the analysis to April through October, or ozone season.

#### 3.3.1 Cross-validation

We performed cross-validation to determine if the spatial monotone risk surface model fits the data better than alternative models and to determine the order of polynomials to use. For each city we used 80% of the data as a training set and fit each model to those data. We compared
models with the deviance of the 20% holdout sample using the predicted values. The deviance is \(2[\hat{Y}_{ct} - Y_{ct} \log(\hat{Y}_{ct}) - \log(Y_{ct})]\) where \(\hat{Y}_{ct}\) is the predicted mortality count on holdout sample day \(t\) in city \(c\).

The comparison models are:

1. A non-spatial monotone risk surface model that replaces the multivariate Gaussian process on \(\theta^*_c\) with an independent multivariate normal model at each site. Hence, \(\theta^*_c\) are independent \(N(\mu, S_2 \otimes S_1)\).

2. A spatial unconstrained risk surface version that models \(f(x; \theta_c)\) directly with a (non-truncated) multivariate Gaussian process prior for \(\theta_c\); hence, removing the monotonicity constraint.

3. A non-spatial unconstrained model that combines the two previously described simplifications of the monotone risk surface model.

4. The NMMAPS model presented in Bell et al. (2004) with a linear ozone effect and natural spline for temperature.

5. An additive model that replaces the linear ozone effect in the NMMAPS model with an unconstrained spline.

For the risk surface models we use the noninformative hyper-parameters \(\tau_0 = 100\), \(a_\tau = 0.001\), and \(b_\tau = 0.001\). For the range parameter we let \(\mu_\rho = 7\) and \(\sigma_\rho = 10\) to provide a diffuse prior centered around 1000 km which is about the range estimate reported in Smith et al. (2009). For the additive model we use the same priors formulation as the non-spatial, unconstrained model but only use a basis expansion of ozone in \(f_c\) since temperature is accounted for in the confounder model. The second-stage model was run for 500,000 iterations and the first half were discarded as burn-in. We thinned the posterior sample keeping every 50th draw due to high autocorrelation, primarily in the range parameter. Appendix B Figures 4 and 5 show trace plots of the posterior sample used for analysis.
Table 3.1: Difference in cross-validation deviance from linear additive model (NMMAPS model).

<table>
<thead>
<tr>
<th>Model</th>
<th>Above 95th Percentile</th>
<th>Overall</th>
<th>Ozone</th>
<th>Temperature</th>
<th>Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial Monotone</td>
<td>-473.2</td>
<td>-42.3</td>
<td>-38.7</td>
<td>-21.1</td>
<td></td>
</tr>
<tr>
<td>Non-spatial Monotone</td>
<td>-459.9</td>
<td>-40.2</td>
<td>-38.5</td>
<td>-20.7</td>
<td></td>
</tr>
<tr>
<td>Spatial unconstrained</td>
<td>-461.4</td>
<td>-38.0</td>
<td>-35.7</td>
<td>-18.9</td>
<td></td>
</tr>
<tr>
<td>Non-spatial unconstrained</td>
<td>-447.9</td>
<td>-38.0</td>
<td>-39.5</td>
<td>-21.4</td>
<td></td>
</tr>
<tr>
<td>Additive ($M_1 = 4$)</td>
<td>-353.9</td>
<td>-24.6</td>
<td>-28.7</td>
<td>-16.5</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1 shows the difference in deviance from the NMMAPS model for the other five models. For the spatial monotone model the best performing model had second-stage expansion of order $M_1 = 7$ and $M_2 = 9$ and first-stage city-specific expansion orders $M_1^c = \max(r_1^c M_1/r_1, 6)$ and $M_2^c = \max(r_2^c M_2/r_2, 4)$. Hence, the first-stage expansions are smaller and proportional to the ratio of the city-specific range ($r_1^c$) to the national range ($r_1$) down to a minimum order. CV results for this order are presented in Table 3.1. Additional results for other order expansions are provided in Appendix B.

Overall, the spatial monotone model has the smallest deviance. All four risk surface models have deviances well below those of the linear and non-linear additive models. Hence, the data support using a non-additive, non-linear model. Because we are interested in the tail behavior we compare the deviance on holdout sample days with ozone above each city’s 95th percentile, days with temperature above the 95th percentile, and the intersection of the two. The spatial monotone model had the lowest deviance for higher ozone but the non-spatial unconstrained model fit slightly better for high temperature and the intersection. We proceed to analyze the data with the spatial monotone model with $M_1 = 7$ and $M_2 = 9$.

### 3.3.2 Analysis of the national average log RR surface

The national ozone-temperature log RR surface shows a statistically significant association between daily 1-hour maximum ozone and mortality. The ozone effect is greater at higher ozone concentrations and at higher temperatures. Hence, a one ppb increase in ozone is associated
with a larger increase in mortality on high temperature days and days with higher ozone levels. Figure 3.2a shows the national log RR surface. This is the pointwise average over the city-specific log RR surfaces weighted by the pointwise city-specific precision. Figures 3.2c and 3.2d show the log RR of ozone at 50 ppb and 100 ppb, respectively, as functions of temperature, along with 95% posterior intervals. All results for log RR are the gradient (3.3) multiplied by 1000. This can be interpreted as the expected percent change in mortality associated with a 10 ppb increase in ozone, where 10 ppb was chosen to be consistent with other publications.

The posterior mean of $\mu$ used in the two-stage normal-normal models is often interpreted as the estimate of the national average risk (e.g. Bell et al., 2004). As such the national average risk is a precision weighted average of the city-specific estimates and the prior mean. In our model
\( \mu \) represents the mean of the latent process \( \theta_c \) and not the process defining the shape restricted surfaces of interest and thus does not have analogous interpretation. To obtain estimate of the national average log RR surface we use the precision weighted average of the city-specific log RR surfaces which are realizations of a truncated process \( \theta_c \) defining the modeled shape restricted surfaces.

Other two-stage analyses that have used a two-level normal model in the second stage have used the posterior mean of \( \mu \) to estimate the national risk. For two-level normal models the posterior mean of \( \mu \) is a precision weighted average of the city-specific estimates and the prior mean, which with flat priors is a precision weighted average of the city effects which we use here. The precision weighted estimate gives more weight in the tails to cities with data extending to those regions of the surface, whereas the population weighted and unweighted cities do not reflect the different ozone-temperature distribution in each city.

At higher ozone concentrations, temperature has a larger modifying effect. Figure 3.3 shows the log RR at the 50th, 75th, 95th, and 99th percentiles of temperature as a function of ozone. For each temperature, log RR increases at higher ozone concentrations. As temperature increases from the median, the log RR increases monotonically for all ozone values; however, the difference between the cross-sections at low ozone values is very small. The probability that RR is greater at the 99th percentile than the 50th percentile is about 0.92 (Figure 3.3b).

The cross-derivative of the risk surface provides a more complete picture of a departure from additivity. The cross-derivative surface is

\[
\frac{\partial^2 f_c(x)}{\partial x_1 \partial x_2} = M_1 M_2 \sum_{j=0}^{M_1-1} \sum_{k=0}^{M_2-1} (\theta_{j,k+1,c} - \theta_{j,k,c}) B_{j,k}(x, M_1 - 1, M_2 - 1). \tag{3.14}
\]

We refer to (3.14) as the interaction surface as it is the rate of change in the log RR surface with respect to temperature. With an additive model the cross-derivative is zero. Figure 3.4 shows the national interaction surface and the pointwise probability that the interaction is greater than zero. The national interaction surface shows a synergistic effect at higher ozone
Figure 3.3: Comparison of the log RR at the 50th, 75th, 95th, and 99th percentiles of temperature. The estimates extend over the range of ozone values observed in at least 5 cities. Subfigure 3.3a shows the mean log RR for each cross-section and Subfigure 3.3b shows the pointwise posterior probability that log RR is greater at the high temperature. Log RR is in percent change in mortality per one ppb increase in ozone.

and temperatures. This synergism is statistically significant over the part of the surface with temperature equal to 90°F. At lower temperatures, the mean interaction is negative, although not statistically significant in this summer-only analysis.

### 3.3.3 Analysis of the city-specific log RR surfaces

We now examine the city-specific surfaces and interaction at the city level. Figure 3.5 shows examples of four city-specific log RR surfaces and their pointwise standard deviations. While the national log RR surface shows a statistically significant ozone effect, the city-specific ozone effect, averaged over the observed days in each city, is significant at the 0.05 level in six cities (Los Angeles, CA; Dallas/Ft. Worth, TX; Houston, TX; San Jose, CA; Oakland, CA; and St. Louis, MO).

To evaluate the interaction effect we compare high and moderate temperature days in each city. We define high temperature days as those with temperatures between the 95th and 99th city-specific temperature percentiles and moderate temperature days to have temperatures be-
Figure 3.4: National interaction surface and pointwise probability of positive interaction. Panel 3.4a shows the national interaction surface which is the cross-derivative of the log risk surface or the derivative of the log RR surface with respect to temperature. This shows how log RR changes with temperature and quantifies the interaction at each point. Panel 3.4b shows the probability that the national interaction surface is greater than 0.

The days between the 50th and 75th percentiles. Within each range we include only days between the 10th and 90th percentiles of ozone in order to minimize the influence of days with extreme ozone values in either direction. Figure 3.1 outlines these days in black for four cities. By using city-specific percentiles this definition of high and moderate temperature days adapts to each city’s weather; however, like previous studies that used a stratified model to compare log RR at different temperatures it does not account for the different ozone distributions of high and moderate temperature days. To remove the effect of different ozone ranges we limit the high and moderate temperature regions to a common ozone range, indicated by the purple box in Figure 3.1.

On high temperature days the log RR is larger than on moderate temperature days over the observed ozone range in most cities, but the difference is greatly reduced when comparing only over the common ozone range. Figure 3.6a compares the ratio of mean log RR on high temperature days to mean log RR on moderate temperature days for both ozone ranges. The large reduction in the ratio when limiting to a common ozone range suggests that much of the
Figure 3.5: Log RR surfaces for selected cities (left) and their pointwise standard deviations (right). Surfaces are plotted only over the range of data observed data for that city. The cities were selected for being geographically diverse and having varied ozone and temperature ranges.
Table 3.2: Mean log RR on high and moderate temperature days over the observed and common ozone rates by region (in percent change in mortality per 10 ppb increase in ozone).

<table>
<thead>
<tr>
<th>Region</th>
<th>Observed Ozone Range</th>
<th>Common Ozone Range</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High Temp</td>
<td>Moderate Temp</td>
</tr>
<tr>
<td>Indust. Midwest</td>
<td>0.25 (0.10)</td>
<td>0.17 (0.06)</td>
</tr>
<tr>
<td>Northeast</td>
<td>0.12 (0.09)</td>
<td>0.06 (0.05)</td>
</tr>
<tr>
<td>Northwest</td>
<td>0.14 (0.06)</td>
<td>0.07 (0.03)</td>
</tr>
<tr>
<td>Southern CA</td>
<td>0.13 (0.07)</td>
<td>0.08 (0.05)</td>
</tr>
<tr>
<td>Southeast</td>
<td>0.17 (0.39)</td>
<td>0.14 (0.33)</td>
</tr>
<tr>
<td>Southwest</td>
<td>0.14 (0.08)</td>
<td>0.08 (0.05)</td>
</tr>
<tr>
<td>Upper Midwest</td>
<td>0.10 (0.07)</td>
<td>0.06 (0.04)</td>
</tr>
<tr>
<td>National</td>
<td>0.15 (0.04)</td>
<td>0.09 (0.02)</td>
</tr>
</tbody>
</table>

Note: The regions from the NMMAPS data are used (see Samet et al., 2000a,b).

difference in log RR between high and moderate temperature days is due to the higher ozone levels on high temperature days in conjunction with the non-linearity of the ozone effect. Over the common ozone range, the ratio ranged from about 1 to 2.5. Most of the cities with larger levels of interaction over the common ozone range are in the north where there is a larger difference between the temperatures on high and moderate temperature days (Figure 3.6b and Table 3.2).

3.3.4 Analysis of excess mortality

Overall we see a trend of a synergistic ozone-temperature effect, both in the national log RR surface and interaction surface, and in the city-specific estimates. This interaction at higher ozone levels and temperatures leads to a larger estimate of excess mortality. Table 3.3 compares the expected change in mortality associated with an increase from median ozone and temperature to the 95th percentile of ozone and temperature for the additive linear model (NMMAPS), the additive non-linear model, and the monotone, spatial risk surface model by region. The general trends are the same, with the largest increases in the northeast and industrial midwest, but the risk surface estimates are larger in most regions.
Figure 3.6: Comparison of the log RR at high temperature and moderate temperatures as defined in Section 3.3.3. Panel 3.6a compares the ratio of log RR over the observed ozone range (black) and common ozone range (blue). The posterior mean and 95% interval are shown. Panel 3.6b shows a map of the ratio on days in the common ozone range. Cities outlined in black (Los Angeles and St. Louis) are significant at the 0.8 level.
Table 3.3: Percent increase in mortality associated with an increase from the medians of ozone and temperature to the 95th percentiles of ozone and temperature using different models.

<table>
<thead>
<tr>
<th>Region</th>
<th>Additive Linear</th>
<th>Additive non-linear</th>
<th>Surface</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indust. Midwest</td>
<td>4.57 (0.77)</td>
<td>3.27 (1.65)</td>
<td>4.13 (0.42)</td>
</tr>
<tr>
<td>Northeast</td>
<td>5.61 (0.94)</td>
<td>5.88 (1.92)</td>
<td>5.31 (0.48)</td>
</tr>
<tr>
<td>Northwest</td>
<td>3.92 (0.82)</td>
<td>0.90 (1.84)</td>
<td>2.42 (0.58)</td>
</tr>
<tr>
<td>Southern CA</td>
<td>2.77 (0.80)</td>
<td>3.89 (2.57)</td>
<td>3.88 (0.77)</td>
</tr>
<tr>
<td>Southeast</td>
<td>0.70 (0.52)</td>
<td>3.15 (1.20)</td>
<td>3.23 (0.38)</td>
</tr>
<tr>
<td>Southwest</td>
<td>2.89 (0.84)</td>
<td>4.71 (2.04)</td>
<td>4.49 (0.64)</td>
</tr>
<tr>
<td>Upper Midwest</td>
<td>1.33 (1.21)</td>
<td>2.91 (2.08)</td>
<td>4.82 (0.62)</td>
</tr>
<tr>
<td>National</td>
<td>3.06 (0.30)</td>
<td>3.54 (0.75)</td>
<td>3.98 (0.24)</td>
</tr>
</tbody>
</table>

3.4 Discussion

In this paper, we propose a two-stage procedure to estimate city-specific ozone-temperature risk surfaces. To accommodate different temperature and ozone ranges in different cities, we use local basis expansions in the first stage. The first-stage results are combined in the second-stage model using a global basis expansion and spatially-varying coefficients to allow for different ozone-temperature effects by city. Cross-validation provides evidence that the effects of ozone and temperature are non-linear, and depart from simple additivity between ozone and temperature. Specifically, the national average log RR surface indicates that the RR of ozone is higher on high temperature days and this interaction is most pronounced in the northern US, where summer temperatures have the most variability.

The results of this study have important implications toward understanding the nature of the joint effects of ozone and temperature on mortality. To compare the ozone effect between different temperature strata it is important that the distributions of ozone be similar within each strata. High correlation between temperature and ozone violate this requirement and thus call for careful consideration and interpretation of methods. Indeed, the results suggest that the higher ozone estimates at high temperatures in stratified studies is primarily due to non-linear effect of ozone coupled with higher concentrations of ozone formed at high temperatures.
However, the results also suggest that statistically significant modification of risk is present at higher temperature levels.

IPCC (2007) raised concerns whether current air quality management practices can adequately protect public health under the future climate regimes. Climate projections show increases in extreme ozone and temperature days over a significant portion of the US (Hogrefe et al., 2004; Kunkel et al., 2008; Tagaris et al., 2007; Wu et al., 2008). The synergistic ozone-temperature effect estimated in this paper implies that the disease burden of these extreme weather days is greater than would be estimated with additive models. This is despite the relatively small size of the interaction effect compared to the non-linear main effect at higher ozone levels.

In our paper we have developed methodology to capture non-linear interactions between temperature and ozone effects. This methodology could be used in other health effects analyses, and potentially beyond. For example, there is increasing interest in identifying joint effects of multiple pollutants (National Research Council, 2004b), such as ozone and fine particulate matter (Kalendra, 2010). Another application is studying cumulative effects using various lagged pollution or temperature variables (Schwartz, 2000; Welty and Zeger, 2005; Heaton and Peng, 2012, 2013). For any two predictors our method would apply directly; extending the model to include a high-dimensional surface for the joint effect of several predictors will be challenging. To accommodate several predictors, our approach could be modified to have an additive structure (Hastie and Tibshirani, 1990) with main effect curves for each predictor and two-dimensional interaction surfaces for pairs of variables thought to interact.
Chapter 4

Confounder Selection via Penalized Credible Regions

4.1 Introduction

In many applications, the effect of an exposure or treatment $X$ on an outcome $Y$ is estimated while controlling for other explanatory or confounding variables. These additional covariates are often selected from a potentially large number of observed variables $U$, and the estimated exposure effect can be sensitive to the set of covariates included. In such cases, omitting confounding variables, those correlated with both $X$ and $Y$, can bias the estimate of the exposure effect. On the other hand, including variables that are correlated with neither $X$ nor $Y$ increases the variance of the exposure effect. Estimating the exposure effect is particularly difficult when the number of potential confounders is large relative to the sample size. However, accounting for all confounding variables while minimizing the number of variables uncorrelated with the outcome that are included in the model is essential for inference on the exposure effect.

There are numerous variable selection methods that can be used to select a model that balances model fit and parsimony. Popular methods include a variety of penalized regression models: the least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996), smoothly
clipped absolute deviation (SCAD; Fan and Li, 2001), the elastic net (Zou and Hastie, 2005), the adaptive LASSO (Zou, 2006), and octagal shrinkage and clustering algorithm for regression (OSCAR; Bondell and Reich, 2008). Bayesian methods have also been developed for variable selection (George and McCulloch, 1993; George and Foster, 2000; Brown et al., 2002; Carvalho et al., 2010; Bondell and Reich, 2012). However, these variable selection methods emphasize prediction, not estimation of the effect of one exposure of interest while treating the other predictors as confounders. Applying these general variable selection methods without specifically treating covariates as confounders could be problematic. For example, if a confounder is highly correlated with the exposure and the exposure is forced to be included in the model, the confounder will likely be dropped, leading to bias.

Rather than selecting a single model, Bayesian model averaging (BMA) attempts to account for effect uncertainty by averaging the effect estimate over the entire model space, giving higher weights to models that have greater support from the data. However, to insure an unbiased effect estimate the model should contain all confounding variables. Crainiceanu et al. (2008) noted that BMA averages over the subspace of models that includes all confounding variables and the subspace that does not, and demonstrated the bias in the effect estimate that can result. To address this, Crainiceanu et al. (2008) proposed a two-step approach. In the first step, the exposure is regressed on the other covariates to identity potential confounders. In the second step, the outcome is regressed on the exposure and the confounders identified in the first step. Additional explanatory variables not identified as confounders are selected in this second step.

Wang et al. (2012) combine the exposure and outcome models into a one-step Bayesian model. For both models, a set of indicator variables parameterize which covariates are included, and those included in the exposure model are included in the outcome model with probability one. Conceptually, this is BMA that only averages over the subspace of exposure models that include all confounders while simultaneously accounting for uncertainty in confounder selection. This method worked well for larger sample sizes, but is biased in simulations for smaller sample sizes and no optimality properties have been identified. In addition, it is computationally
In this paper we take a decision-theoretic approach to confounder selection and effect estimation. Our approach extends that of Bondell and Reich (2012) from the usual variable selection setting to the important problem of confounder adjustment. We first fit the standard Bayesian regression model and then post-process the posterior distribution in a decision-theoretic way using a loss function that penalizes models that do not include confounding variables and other important covariates. Thus, the proposed estimator is the Bayes rule associated with the proposed confounder-specific loss function. The distinction between the different losses associated with errors in the estimates of the variable of interest and confounders are made explicit in our loss function that leads to the final estimate. This approach allows the priors on the regression coefficients to represent the state of knowledge before seeing the data as in the usual subjective Bayesian model, rather than assigning priors designed to avoid errors caused by failing to include important confounders, which is less natural in the Bayesian framework.

The proposed method has several appealing properties. Under general conditions it exhibits the oracle properties (Fan and Li, 2001) and simulation studies show good finite sample performance. Our method can be easily fit with existing software for a wide variety of outcome and exposure models, including all generalized linear models. In some cases Markov chain Monte Carlo methods are not required, for example, the normal linear model with normal or flat priors. We also establish a connection between our method and the adaptive LASSO (Zou, 2006), but with weights tailored to confounder adjustment rather than variable selection.

4.2 Methods

4.2.1 Modeling approach

The idea behind the proposed method is to find the simplest model among all feasible models. We define a feasible model as one with coefficients in the $(1 - \alpha) \times 100\%$ posterior region of the full model. Within this credible region we select the model that maximizes parsimony. In
this case, parsimony is defined as the model containing all confounders and other covariates associated with the outcome, but no other covariates.

Consider the normal linear outcome model

\[
Y = \beta_0 + X \beta_x + U \beta_u + \epsilon_y, \quad \epsilon_y \sim N(0, \sigma^2_y I) \tag{4.1}
\]

and exposure model

\[
X = \gamma_0 + U \gamma_u + \epsilon_x, \quad \epsilon_x \sim N(0, \sigma^2_x I). \tag{4.2}
\]

For notational simplicity, let \( \beta = (\beta_0, \beta_x, \beta^T_u) \), \( \beta_u = (\beta_1, \ldots, \beta_p)^T \), \( \gamma = (\gamma_0, \gamma^T_u) \), and \( \gamma_u = (\gamma_1, \ldots, \gamma_p)^T \) and assume \( W = (1, X, U) \) is \( n \times (p + 2) \) with a column of ones for the intercept.

Let \( C^\beta_\alpha \) and \( C^\gamma_\alpha \) be \((1 - \alpha) \times 100\% \) posterior credible regions of \( \beta \) and \( \gamma \), respectively. In the case of flat or \( N(0, \sigma^2_y / \tau_y I) \) on \( \beta \) (with \( \tau_y \) fixed) and inverse-gamma prior on \( \sigma^2_y \), the highest posterior density region is of the form \( C^\beta_\alpha = \{ \beta : (\beta - \hat{\beta})^T \Sigma_y^{-1} (\beta - \hat{\beta}) < C_\alpha \} \) where \( \hat{\beta} \) and \( \Sigma_y \) are the posterior mean and covariance. For regions of this form, there is a one-to-one relationship between the chosen \( \alpha \) level and the scaler \( C_\alpha \). For the exposure model, \( C^\gamma_\alpha \) has a similar form.

In general, elliptical credible regions of this form exist for other priors on \( \beta \) and for other likelihoods such as generalized linear models, but may not be the highest posterior density regions.

The set of feasible exposure and confounder coefficients are \( \{ \beta : \beta \in C^\beta_\alpha \} \) and \( \{ \gamma : \gamma \in C^\gamma_\alpha \} \) for a given probability level \( \alpha \in (0, 1) \). This feasible space will potentially include coefficient vectors with some parameters equal to zero, and thus a reduced model.

The model of interest is \( Y = \beta_0 + X \beta_x + U_{A_u} \beta_{A_u} + \epsilon_y \), where \( A_u = \{ j : \beta_j \neq 0 \} \cup \{ j : \gamma_j \neq 0 \} \) and \( U_{A_u} \) and \( \beta_{A_u} \) are the corresponding subsets of \( U \) and \( \beta \). This includes all variables correlated with the outcome or exposure. It is possible that this may include some variables correlated with \( X \) but not \( Y \); however, the errors resulting from including these variables are less severe than errors from omitting variables that are correlated with both the outcome and exposure.

To find a sparse estimate we are interested in the feasible coefficient vector that minimizes
the cardinality of \( A \). The proposed estimator is

\[
\tilde{\beta} = \arg\min_{\beta} \epsilon (\beta_0^2 + \beta_x^2) + \|\beta_u + |\gamma_u|\|_0, \quad \text{subject to } \beta \in C_\alpha^\beta \text{ and } \gamma \in C_\alpha^\gamma
\]  

(4.3)

where \( \| \cdot \|_0 \) denotes the \( L_0 \) norm of a vector, that is, the number of nonzero elements, and \( \epsilon \) is fixed and small so the intercept and exposure effect are essentially unpenalized.

### 4.2.2 Penalized regression reformulation

The solution to (4.3) can be hard to find in practice as it requires a search over potentially high-dimensional posterior regions. To ease computation, we use the smooth homotopy between \( L_0 \) and \( L_1 \) proposed by Lv and Fan (2009) and used by Bondell and Reich (2012) for the usual linear regression setting, \( \rho_a(t) = \{(a+1)t)/(a+t) \). The proposed criterion is

\[
\sum_{j=1}^p \rho_a(|\hat{\beta}_j| + |\hat{\gamma}_j|) + \left[\frac{a(a+1)}{(a+|\hat{\beta}_j| + |\hat{\gamma}_j|)^2} + \frac{a(a+1)}{(a+|\hat{\beta}_j| + |\hat{\gamma}_j|)^2}\right].
\]  

(4.4)

Since both credible regions are convex sets of the form \( \{ \beta : \beta \in C_\alpha^\beta \} \) and \( \{ \gamma : \gamma \in C_\alpha^\gamma \} \), the proposed estimator in (4.3) with the local linearized penalty (4.4) is equivalent to the Lagrangian optimization problem

\[
\tilde{\beta} = \arg\min_{\beta} (\beta - \hat{\beta})^T \Sigma_y^{-1}(\beta - \hat{\beta}) + (\gamma - \hat{\gamma})^T \Sigma_x^{-1}(\gamma - \hat{\gamma})
\]

\[
+ \lambda \epsilon (\beta_0^2 + \beta_x^2) + \lambda \sum_{j=1}^p \left[ \frac{(|\hat{\beta}_j| + |\hat{\gamma}_j|)^2}{(a + |\hat{\beta}_j| + |\hat{\gamma}_j|)^2} + \frac{(|\hat{\beta}_j| + |\hat{\gamma}_j|)^2}{(a + |\hat{\beta}_j| + |\hat{\gamma}_j|)^2} \right].
\]  

(4.5)

In (4.5), the Lagrangian multiplier \( \lambda \) absorbs the \( a(a+1) \) in the numerator of (4.4).
Because our primary interest is in the effect of $X$, we let $\epsilon = 0$ to completely remove the penalty from the intercept and exposure. The second and fifth terms on the right hand side of (4.5) and the $\gamma_j$ from the numerator of the fourth term do not effect the minimization with respect to $\beta$ and can be removed. Finally, by letting $a \to 0$ we get the objective function

$$\tilde{\beta} = \arg\min_{\beta} (\beta - \hat{\beta})^T \Sigma_y^{-1} (\beta - \hat{\beta}) + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{(|\hat{\gamma}_j| + |\hat{\beta}_j|)^2}.$$  

For any given dataset, there is a one-to-one and decreasing relationship between $\alpha$ and $\lambda$ because of the one-to-one relationships between $\alpha$ and $C_\alpha$ and between $C_\alpha$ and $\lambda$. The path of estimates obtained by varying $\alpha$ is identical to the path obtained by varying $\lambda$. This converts (4.3) into a convex optimization problem with a single tuning parameter that can be easily solved with existing software.

### 4.2.3 Simplification under flat prior

Assuming the linear model in (4.1) and a flat prior for $\beta$, $\text{pr}(\beta) \propto 1$, the posterior mean is the least squares estimate $\hat{\beta} = (W^T W)^{-1} W^T Y$ and the posterior covariance is proportional to $\Sigma_y \propto (W^T W)^{-1}$. In this case, $(\beta - \hat{\beta})^T \Sigma_y^{-1} (\beta - \hat{\beta})$ is a linear function of the sum of squared errors and the estimate can be written

$$\tilde{\beta} = \arg\min_{\beta} (Y - W\beta)^T (Y - W\beta) + \lambda \sum_{j=1}^{p} \hat{w}_j |\beta_j|$$  

(4.7)

where $\hat{w}_j = 1/(|\hat{\gamma}_j| + |\hat{\beta}_j|)^2$ for confounders $j = 1, \ldots, p$. This is a special case of the adaptive LASSO solution (Zou, 2006) with data-driven weights $w_j$ tailored to confounder adjustment so that variables associated with either the exposure (large $|\hat{\gamma}_j|$) or the response (large $|\hat{\beta}_j|$) have small penalty and are thus encouraged to be included in the outcome model.
4.2.4 Extension to multiple exposures

There is often interest in the effect of multiple exposures on an outcome. The penalized credible
region confounder selection method naturally extends to the multiple exposure case. For a
second exposure $X_2$ there is a second exposure model, for example

$$X_2 = \delta_0 + U\delta_u + \epsilon_{x_2}, \quad \epsilon_{x_2} \sim N(0, \sigma_{x_2}^2 I).$$ (4.8)

Using a similar approach to the single exposures setting the penalized regression estimator for
multiple exposures is

$$\tilde{\beta} = \arg\min_{\beta} (\beta - \hat{\beta})^T \Sigma_y^{-1} (\beta - \hat{\beta}) + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{(|\beta_j| + |\hat{\gamma}_j| + |\hat{\delta}_j|)^2}. \quad (4.9)$$

All the computation remains the same with (4.9). Additional exposure models can be used
when there are more than two exposures of interests. In addition, more complex relationships
can be estimated when it is appropriate. For example, $X_2$ could be a covariate in (4.2) if it
is believed that confounders are correlated with the first exposure conditionally on $X_2$, i.e.
Simpson’s paradox.

4.3 Theoretical Results

Although motivated by Bayesian decision theory, the proposed estimator can take the form of
a standard penalized regression objective function. Therefore, it is of interest to evaluate this
new estimate using the techniques of penalized regression. Theorem 1 demonstrates that the
credible region confounder method has the oracle property for a properly chosen $\lambda_n$. Without
loss of generality let $A = \{0, x, 1, \ldots, p_0\}$ and make the following assumptions:

Assumption 1. The posterior covariance $\Sigma_y^{-1}/n \xrightarrow{d} C$, where $C$ is a $(p + 2) \times (p + 2)$ positive
definite matrix that can be partitioned as

\[ C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}. \]  

(4.10)

**Assumption 2.** The posterior means are asymptotically normal, \( \sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, C^{-1}) \), and analogously for \( \gamma \).

**Theorem 1.** Under conditions 1 and 2, if \( \frac{\lambda_n}{\sqrt{n}} \to 0 \) and \( \lambda_n \sqrt{n} \to \infty \) the penalized credible region confounder method is consistent in variable selection, \( \lim_{n \to \infty} \text{pr}(A_n = A) = 1 \), and asymptotically normal, \( \sqrt{n}(\tilde{\beta}_A - \beta_A) \xrightarrow{d} N(0, C^{-1}_{11}) \).

**Remark 1.** Theorem 1 requires that the posterior mean is consistent for \( \beta \), which holds if \( \beta \)'s prior does not change with the sample size and has positive mass in the neighborhood of the truth, and that the asymptotic covariance matrix of \( W \) is full rank.

Theorem 1 assumes a rate on \( \lambda_n \). However, for the Bayesian credible regions approach it would be more natural to study asymptotics as a function of the confidence level \( \alpha_n \). For a given dataset, there is a one-to-one correspondence between \( \lambda \) and \( \alpha \). However, the limiting behavior as a function of \( \alpha_n \) can be quite different than those for \( \lambda_n \) (Gunes and Bondell, 2012; Bondell and Reich, 2012). Theorem 1 implies path consistency as a function of \( \alpha_n \). Because the solution paths obtained by varying \( \alpha \) and \( \lambda \) are identical for each dataset, the selection consistency result in Theorem 1 implies that the correct model will be included in the solution path with probability tending to one for the Bayesian credible regions approach. A proof of Theorem 1 is in Appendix C.

### 4.4 Computation and tuning

We first fit the full exposure and outcome models separately. This can be done with MCMC or with a closed form solution when it is known to exist. Given the posterior means \( \hat{\beta} \) and \( \hat{\gamma} \) and
posterior covariances $\Sigma_x$ and $\Sigma_y$, we obtain the solution to (4.6) using least angle regression (LARS; Efron et al., 2004). We then refit the outcome model using only confounders with nonzero coefficients for our final estimates.

Let $X^* = \Sigma_y^{-1/2} D$ where $D = \text{diag}\{\delta^{-1}, \delta^{-1}, (|\hat{\beta}_1| + |\hat{\gamma}_1|)^2, \ldots, (|\hat{\beta}_p| + |\hat{\gamma}_p|)^2\}$ and $\delta$ is a very small number that effectively removes the penalty on $\beta_0$ and $\beta_x$. If $\delta$ is 0 then the model fit is exactly (4.6); however, approximation with a very small number results in the same solution and a nonzero value is required for LARS. Then let $Y^* = \Sigma_y^{-1/2} \hat{\beta}$. Solving the $L_1$ penalized regression function with $Y^*$ and $X^*$ gives the solution $\beta^*$. The solution to (4.3) is $\tilde{\beta} = D\beta^*$.

The method must be tuned by selecting either the confidence level $\alpha \in (0, 1)$ or the penalty parameter $\lambda > 0$. For any given dataset, there is a one-to-one relationship between $\alpha$ and $\lambda$ and the path of estimates obtained by varying $\alpha$ is identical to the path obtained by varying $\lambda$. Gunes and Bondell (2012) discuss this relationship in detail and note that most tuning methods (AIC, BIC, and cross-validation) can result in similar models but very different $\alpha$ levels for different data sets. Further, the solution corresponding to a certain $\alpha$ level is not related to controlling the false positive rate at that level. As such, tuning based on a pre-specified $\alpha$ level can result in poor selection properties, while using the solution path obtained from varying $\alpha$ or $\lambda$ performs well. For convenience, we tune based on $\lambda$ since this is the approach taken in the LARS package in R. This gives the entire solution path.

There are several approaches to selecting a model from that path. We found that forward selection performed best in our simulation study. We performed forward selection on the exposure and outcome model simultaneously using the same path. Forward selection stops when the minimum of the $\chi^2$ tests for adding an additional covariate to the outcome and exposure model fails to reject at a particular $\alpha$ level denoted by $\alpha_{fs}$. Larger $\alpha_{fs}$ values will include more covariates and can be considered more conservative. The LARS path may have a variable appear in a model and then be removed before reappearing (Efron et al., 2004). For forward selection, we modified the LARS path to include a covariate in every model after it first appears in order to ensure nested models for testing. Alternative selection methods include cross-validation,
Mallow’s $C_p$ (Mallows, 1973), Akaike information criterion (Akaike, 1973), and Bayesian information criterion (Schwarz, 1978). Appendix C contains a simulation comparing these methods.

4.5 Simulation Study

4.5.1 Simulation with linear model

Our first simulation design follows from Wang et al. (2012). We assume $Y \sim N(W\beta, I)$ where $W$ is an $n \times (p+1)$ design matrix with the variable of interest in the first column and $p = 57$ potential confounders. We let the covariates for the $i$th observation, $W_i$, be independent $N(0, \Sigma)$. The covariance matrix $\Sigma = \{\sigma_{jk}\}$ has diagonal elements $\sigma_{jj} = 1$ and nonzero off-diagonal elements $\sigma_{jk} = 0.7^{j+k-2}$ for $j \neq k = 0, \ldots, 7$. The remaining off-diagonal elements are zero; hence, $w_{ik}$ for $k > 7$ are independent standard normal. The coefficients are $\beta_x = \beta_1 = \cdots = \beta_{14} = 0.1$ and $\beta_{15} = \cdots = \beta_{57} = 0$. Hence, we have seven confounding variables correlated with $Y$ and $X$, seven additional explanatory variables correlated with $Y$ but not $X$, and 43 variables uncorrelated with both $X$ and $Y$.

We present the results for the penalized credible regions using three different priors on $\beta$ and $\gamma$: flat as assumed in Section 4.2.3; $\beta \sim N(0, \sigma_y^2/\tau_y)$ and $\gamma \sim N(0, \sigma_x^2/\tau_x)$ with $\tau_y$ and $\tau_x$ estimated from a linear model and fixed (empirical Bayes); and with independent Gamma(0.001,0.001) priors on $\tau_y$ and $\tau_x$ for a fully Bayesian approach with proper priors. For the empirical Bayes approach we let $\hat{\tau}_y = \bar{\sigma}_y^2/((p + 1)^{-1} \sum_{j=0}^p \hat{\beta}_j^2)$ and $\hat{\tau}_x = \bar{\sigma}_x^2/((p^{-1} \sum_{j=1}^p \hat{\gamma}_j^2)$ where $\bar{\sigma}$, $\hat{\beta}$, and $\hat{\gamma}$ are estimated from the full linear model. In all cases we put independent Gamma(0.001,0.001) priors on $\sigma_y^{-2}$ and $\sigma_x^{-2}$.

Table 4.1 compares model performance for the credible region method with the three priors. For comparison we evaluate the simulated data with the true frequentist linear model that includes only the first 14 covariates; the full frequentist linear model that includes all 57 covariates; BMA with $\beta_j = \eta_j \alpha_j$, $\eta_j \sim Bern(0.5)$, and $\alpha_j \sim N(0, 10^2)$; Bayesian adjustment for confounding (BAC; Wang et al., 2012) with $\omega = \infty$ using the R package BEAU; and adaptive
LASSO with $\gamma = 2$.

In general, the fully Bayesian credible region approach performed similarly to the true model in terms of bias, MSE, and interval coverage for $\hat{\beta}_x$, with the exception of the smallest sample size where it was biased and had lower coverage. At the larger sample sizes, the empirical Bayes and flat prior versions perform well and provide faster computation. These methods (and the adaptive LASSO) require a large enough sample size to get good initial least squares estimates $\hat{\beta}$ and $\hat{\gamma}$, as well as $\hat{\tau}_y$ and $\hat{\tau}_x$ for the empirical Bayes priors, and thus perform poorly for small $n$. BMA, BAC, and adaptive LASSO had higher MSE, were more biased, and had lower interval coverage compared to the fully Bayesian credible region method, although the MSE for BAC and adaptive LASSO approaches that of the credible region at higher sample sizes. At the larger sample sizes, BAC was notably slower than the other methods. For the moderate and larger sample sizes the interval coverage for $\hat{\beta}_x$ achieved the nominal 95% level indicating that proper inference is being made with the credible region method.

We computed the AUC for the credible region approach and adaptive LASSO by calculating the sensitivity and specificity as additional variables are added along the path formed by varying $\lambda$. The credible region approach had larger AUC showing the benefit of weights specifically tailored to confounder adjustment. For BAC and BMA the AUC is estimated by adding variables to the model according to their inclusion probability. BAC had slightly higher AUC than the credible region method at the smallest sample size but smaller AUC at the largest sample size.

Despite the higher AUC at small sample sizes, BAC tended to give more weight to smaller models. For $n = 100$, Figure 4.1 compares the inclusion probability for each of the important covariates and the average inclusion probability of the other covariates, using the empirical Bayes priors for the confounder method. This shows that the credible region approach and BMA had a higher inclusion rate for the confounding variables than BAC and adaptive LASSO.

While variables 1 through 14 have the same effect size, their correlation with the exposure varies. The credible region approach includes confounding variables with greater frequency. The same is true for BAC but the pattern is not as strong. Adaptive LASSO and BMA do not
account for the correlation with the outcome resulting in similar inclusion probabilities for all true predictors, not an emphasis on those correlated with the exposure that can impact the exposure effect estimation. The inclusion rate for the true confounders is below one because confounders that do not reduce the sum of squared errors for either model are excluded. By excluding true confounders that do not reduce the sum of squared errors the credible region approach can have lower MSE than the true model due to the larger error degrees of freedom.

These simulation results highlight the advantage of the penalized credible region approach rather than BMA or BAC which use zero-inflated priors for selection. Specifically, at smaller sample sizes the credible regions approach has smaller bias and MSE and higher interval coverage. The credible regions approach with flat priors has a similar objective function to the adaptive LASSO, except the weights in the penalty are augmented for confounder selection. Hence, the comparison of the results for the adaptive LASSO and the penalized credible region approach with flat priors highlights the advantage gained by including the confounder specific
weights. These advantages include smaller bias and MSE as well as better interval coverage and AUC. Using the approach of Bondell and Reich (2012) with flat priors results in an estimator identical to the adaptive LASSO; hence, the use of the exposure model and confounder-specific loss function proposed in this paper improve performance when the primary interest is in estimating an exposure effect. Compared to a purely penalized regression approach, by using the Bayesian framework that motivated this method a practitioner can select priors for the regression coefficients. Even a vague normal prior improves finite sample performance as demonstrated by the empirical Bayes and fully Bayesian penalized credible region methods used here, and there is potential for additional gains when more prior information is available.

4.5.2 Simulation in the ultra high-dimensional setting.

In many cases there is interest in inference for \( p > n \). To test the confounder selection performance in this setting we add additional covariates to the simulation that are independent standard normal uncorrelated with both the outcome and the exposures. Hence, the simulation design remains unchanged with the exception of additional noise variables added to the design. We use the fully Bayesian prior with gamma prior on \( \tau_x \) and \( \tau_y \) and let \( n \) grow at three rates: \( n = 100 \), \( n = 15 + \sqrt{p} \), and \( n = p/2 \). Figure 4.2 shows the resulting AUC. When \( n = p/2 \) the credible regions approach performs well for variable selection at large sample sizes. However, for slower rates of growth the credible regions approach does not adequately select confounding variables. Hence, for \( p > n \) if \( n \) grows at a sufficiently rate the penalized credible region method performs well.

4.5.3 Simulation with binary treatment and logistic confounder model

Our second simulation design assumes two treatment groups with no treatment effect. We assume that there is a set of five confounders that are predictors of \( Y \) and have different distribution for the two treatment groups. We generate \( X_i \) as independent Bernoulli with probability 0.5 and let \( U_{ij} \) be normal with mean \( j^{-1}X_i \) and standard deviation \( j^{-1} \) for \( j = 1, \ldots, 5 \). The
Figure 4.2: AUC for simulation design one with $p > n$ for three different rates of growth for $n$. 

$n=100$

$n=15+\sqrt{p}$

$n=p/2$
remaining variables \((U_{i6}, \ldots, U_{i50})\) are independent standard normal. The regression coefficients are \(\beta_1, \cdots, \beta_5 = 0.1\) and \(\beta_6 = \cdots = \beta_{50} = 0\) and \(Y\) is \(N(W\beta, I)\).

We fit the penalized credible regions using a logistic exposure model: \(X_i\) is normal \(\text{Bern}(p_i)\), \(p_i = \text{logit}(\gamma_0 + u_i^T \gamma_u)\), and the elements of \(\gamma\) have independent standard normal priors. We present the same three priors for the outcome model as used in simulation design one: flat, empirical Bayes, and Gamma. For comparison we show the true and full regression model as well as Bayesian model averaging and adaptive LASSO, but not BAC because the BEAU package does not model a binary treatment effect.

The credible regions methods and BMA had smaller MSE than the true model but were unbiased at the smaller sample sizes. For \(n = 200\) and \(500\) the credible region methods were unbiased and had MSE similar to the true model, whereas BMA and adaptive LASSO were still biased and had larger MSE. In addition, the credible region method had interval coverage near the nominal level throughout, whereas BMA and adaptive LASSO were lower. The AUC with the credible region method was larger than with the adaptive LASSO at all sample sizes and larger than with BMA for all except the smallest sample size.

4.6 Data Analysis

To illustrate the credible region method we estimate the effect of mean fine particulate matter (PM\(2.5\)) in the first trimester of pregnancy on birth weight in Mecklenburg County, North Carolina, while accounting for several potential confounders related to the mother’s socioeconomic status, medical history, seasonality, and other weather variables. PM\(2.5\) air pollution has been associated with various adverse pregnancy outcomes (Bosetti et al., 2010; Šrám et al., 2005). We study the effect of PM\(2.5\) on birth weight using data similar to Chang et al. (2012). PM\(2.5\) levels were obtained from the U.S. Environmental Protection Agency’s Fused Air Quality Predictions Using Downscaling (http://www.epa.gov/esd/land-sci/lcb/lcb_faqsd.html). From these data we computed average PM\(2.5\) over the first trimester for each birth. We used birth and covariate data from the North Carolina Vital Statistics – Births 2003 through 2007 (State Center for
Health Statistics, 2008) from the State Center for Health Statistics (SCHS) and the Howard W. Odum Institute for Research in Social Science at University of North Carolina at Chapel Hill. Temperature and dew point data were obtained from National Oceanic and Atmospheric Administration Climate Data Online (NOAA-CDO; http://www.ncdc.noaa.gov/cdo-web/).

We looked at the group of at risk women age 40 and over and limited the sample to single births that reached at least 37 weeks of gestation who were self-reported non-hispanic white, non-hispanic black, or hispanic. The exposure variable of interest is first trimester mean PM$_{2.5}$. In addition, following Warren et al. (2012), we include as potential confounders other variables observed during pregnancy, including: principal components (PC) of mean daily temperature throughout the pregnancy, PCs of mean daily dew point throughout the pregnancy, PCs for interaction of temperature and dew point, indicators for a birth in spring, summer, or fall. The PCs that explained 0.99 percent of the week-to-week variation in co-exposures were included. These covariates are correlated with PM$_{2.5}$ and potentially impact birth weight. In addition we included variables relating to the birth and mother’s medical history. These variables are listed in Appendix C. All observations with missing data were excluded. The final sample contains 1399 women of which 902 are non-hispanic white, 308 are non-hispanic black, and 171 are hispanic. There are $p = 65$ covariates in total. Within each subgroup the race/ethnicity covariates were removed as well as any other covariates that were not observed in the subgroup, resulting in 58, 61, and 63 covariates in the hispanic, non-hispanic black, and non-hispanic white models, respectively.

We used the credible region method with fully Bayesian priors. The resulting analysis shows that the credible regions method results in a similar point estimate for the PM$_{2.5}$ exposure effect as the full model but has smaller variance. This is appealing because the full model includes all observed confounding variables; hence, a point estimate that substantially differs from the full model implies that important confounding variables in the data are omitted from the model.

Figure 4.3 shows the solution paths for each subgroup analysis. Figure 4.3a shows that for the hispanic subgroup the model with only PM$_{2.5}$ included (step 0) indicates that PM$_{2.5}$ is
associated with increased birth weight. However, as additional confounding variables correlated with both PM$_{2.5}$ and birth weight are added to the model the point estimate becomes negative. Starting at step 20, the exposure effect is unchanged by including additional covariates, indicating that all important confounding variables have been included.

Figure 4.4a shows that the credible region approach results in similar estimates to the full model overall and for each subgroup, but the standard errors of the effect estimate are 10% to 20% smaller with the credible regions approach (see Figure 4.4b). On the other hand, BMA and adaptive LASSO have smaller variances, but very different point estimates compared to the full model. This stems from selecting smaller models and omitting important confounders, particularly in the smaller subgroups. As a result, adaptive LASSO and the exposure only model find a statistically significant positive association between PM$_{2.5}$ and birth weight among hispanics, a result that contradicts previous findings (Savitz et al., 2014; Pearce et al., 2012). This result mirrors that of the simulation study which showed these methods can be biased at small sample sizes.

4.7 Discussion

This paper presents a new method for confounder selection and effect estimation using a decision-theoretic approach. The proposed estimator is the Bayes rule estimate associated with the confounder-specific loss function and allows the practitioner freedom to select an appropriate prior. Given the posterior mean and covariance for the coefficients of any generalized linear model the credible region confounder method can easily be applied and, in most cases, existing software to be used.

The proposed method outperformed alternative Bayesian confounder or variable selection approaches that utilize zero-inflates priors (BMA and BAC) in a simulation study. In addition to good finite sample performance, the credible region confounder method is consistent in variable selection and asymptotically normal under general conditions.

With flat priors on the regression coefficients, the proposed method is an adaptive LASSO-
Figure 4.3: Solution path for three subgroups and overall. The thick black line is the estimated PM effect, \( \hat{\beta}_x \). The other lines are the estimates for the other covariates in the model. The regression coefficients correspond to centered and scaled variables.
Figure 4.4: Estimates of the PM effect (\( \hat{\beta}_x \)). Top: The point estimates with each model and the 95% interval. Bottom: The ratio of the SE or SD of \( \hat{\beta}_x \) with each model compared to the full OLS model.
type estimator with data-driven weights tailored to confounder selection. The simulation study demonstrated that the confounder-specific weights in the penalty improved performance at small sample sizes (i.e. comparing adaptive LASSO to the credible region method with flat priors). While an alternative approach might be to skip the Bayesian motivation for the estimators, adding even a vague normal prior made a noticeable improvement in confounder selection at small sample sizes. In addition, the adaptive LASSO solution is identical to that of Bondell and Reich (2012) with flat priors, thus, demonstrating the advantage of the confounder-specific loss function in this similar Bayesian framework.

4.8 Software

The R package BayesPen (Wilson et al., 2014) implements the confounder selection via penalized credible regions method.
Table 4.1: Simulation results for design 1. Bias, MSE, and coverage (Cover) are for the effect of interest $\hat{\beta}_x$. Coverage is 95% confidence or credible interval coverage. CPU Time is reported in seconds on a MacBook Pro with OS X, 8 GB RAM, and 2 GHz Intel Core i7. SEs for the AUC range from 0.002 to 0.004, and for CPU time from less than 0.001 to 0.498.

<table>
<thead>
<tr>
<th>n = 60</th>
<th>Bias</th>
<th>MSE</th>
<th>Cover</th>
<th>AUC</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>0.005 (0.011)</td>
<td>0.060 (0.004)</td>
<td>0.93</td>
<td>NA</td>
<td>0.003</td>
</tr>
<tr>
<td>Full</td>
<td>0.007 (0.073)</td>
<td>2.624 (0.392)</td>
<td>0.95</td>
<td>NA</td>
<td>0.003</td>
</tr>
<tr>
<td>Cred. Reg. (Flat Priors)</td>
<td>0.181 (0.008)</td>
<td>0.065 (0.003)</td>
<td>0.69</td>
<td>0.530</td>
<td>0.084</td>
</tr>
<tr>
<td>Cred. Reg. (Empirical Bayes)</td>
<td>0.145 (0.009)</td>
<td>0.062 (0.003)</td>
<td>0.73</td>
<td>0.547</td>
<td>0.089</td>
</tr>
<tr>
<td>Cred. Reg. (Gamma Priors)</td>
<td>0.060 (0.010)</td>
<td>0.050 (0.003)</td>
<td>0.82</td>
<td>0.593</td>
<td>2.971</td>
</tr>
<tr>
<td>BMA</td>
<td>0.090 (0.009)</td>
<td>0.045 (0.003)</td>
<td>0.86</td>
<td>0.577</td>
<td>1.615</td>
</tr>
<tr>
<td>BAC ($\omega = \infty$)</td>
<td>0.080 (0.014)</td>
<td>0.110 (0.008)</td>
<td>0.76</td>
<td>0.633</td>
<td>6.166</td>
</tr>
<tr>
<td>Adaptive LASSO</td>
<td>0.076 (0.049)</td>
<td>1.214 (0.200)</td>
<td>0.58</td>
<td>0.521</td>
<td>0.049</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n = 100</th>
<th>Bias</th>
<th>MSE</th>
<th>Cover</th>
<th>AUC</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>-0.005 (0.008)</td>
<td>0.032 (0.002)</td>
<td>0.95</td>
<td>NA</td>
<td>0.003</td>
</tr>
<tr>
<td>Full</td>
<td>-0.013 (0.011)</td>
<td>0.065 (0.004)</td>
<td>0.95</td>
<td>NA</td>
<td>0.004</td>
</tr>
<tr>
<td>Cred. Reg. (Flat Priors)</td>
<td>0.051 (0.009)</td>
<td>0.044 (0.003)</td>
<td>0.76</td>
<td>0.624</td>
<td>0.108</td>
</tr>
<tr>
<td>Cred. Reg. (Empirical Bayes)</td>
<td>0.001 (0.007)</td>
<td>0.028 (0.002)</td>
<td>0.93</td>
<td>0.680</td>
<td>0.127</td>
</tr>
<tr>
<td>Cred. Reg. (Gamma Priors)</td>
<td>0.006 (0.007)</td>
<td>0.028 (0.002)</td>
<td>0.91</td>
<td>0.670</td>
<td>3.001</td>
</tr>
<tr>
<td>BMA</td>
<td>0.100 (0.006)</td>
<td>0.030 (0.002)</td>
<td>0.80</td>
<td>0.627</td>
<td>1.671</td>
</tr>
<tr>
<td>BAC ($\omega = \infty$)</td>
<td>0.050 (0.008)</td>
<td>0.033 (0.002)</td>
<td>0.91</td>
<td>0.701</td>
<td>7.874</td>
</tr>
<tr>
<td>Adaptive LASSO</td>
<td>0.071 (0.009)</td>
<td>0.043 (0.003)</td>
<td>0.71</td>
<td>0.571</td>
<td>0.041</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n = 200</th>
<th>Bias</th>
<th>MSE</th>
<th>Cover</th>
<th>AUC</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>0.004 (0.005)</td>
<td>0.015 (0.001)</td>
<td>0.95</td>
<td>NA</td>
<td>0.003</td>
</tr>
<tr>
<td>Full</td>
<td>0.004 (0.006)</td>
<td>0.018 (0.001)</td>
<td>0.95</td>
<td>NA</td>
<td>0.005</td>
</tr>
<tr>
<td>Cred. Reg. (Flat Priors)</td>
<td>0.012 (0.006)</td>
<td>0.017 (0.001)</td>
<td>0.90</td>
<td>0.748</td>
<td>0.165</td>
</tr>
<tr>
<td>Cred. Reg. (Empirical Bayes)</td>
<td>0.003 (0.005)</td>
<td>0.014 (0.001)</td>
<td>0.95</td>
<td>0.787</td>
<td>0.184</td>
</tr>
<tr>
<td>Cred. Reg. (Gamma Priors)</td>
<td>0.008 (0.005)</td>
<td>0.014 (0.001)</td>
<td>0.94</td>
<td>0.776</td>
<td>3.093</td>
</tr>
<tr>
<td>BMA</td>
<td>0.114 (0.005)</td>
<td>0.024 (0.001)</td>
<td>0.73</td>
<td>0.718</td>
<td>1.832</td>
</tr>
<tr>
<td>BAC ($\omega = \infty$)</td>
<td>0.044 (0.006)</td>
<td>0.017 (0.001)</td>
<td>0.93</td>
<td>0.779</td>
<td>40.779</td>
</tr>
<tr>
<td>Adaptive LASSO</td>
<td>0.048 (0.006)</td>
<td>0.020 (0.001)</td>
<td>0.78</td>
<td>0.687</td>
<td>0.051</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n = 500</th>
<th>Bias</th>
<th>MSE</th>
<th>Cover</th>
<th>AUC</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>0.000 (0.003)</td>
<td>0.005 (0.000)</td>
<td>0.95</td>
<td>NA</td>
<td>0.004</td>
</tr>
<tr>
<td>Full</td>
<td>0.000 (0.003)</td>
<td>0.006 (0.000)</td>
<td>0.95</td>
<td>NA</td>
<td>0.010</td>
</tr>
<tr>
<td>Cred. Reg. (Flat Priors)</td>
<td>0.001 (0.003)</td>
<td>0.006 (0.000)</td>
<td>0.94</td>
<td>0.906</td>
<td>0.268</td>
</tr>
<tr>
<td>Cred. Reg. (Empirical Bayes)</td>
<td>0.000 (0.003)</td>
<td>0.006 (0.000)</td>
<td>0.95</td>
<td>0.917</td>
<td>0.367</td>
</tr>
<tr>
<td>Cred. Reg. (Gamma Priors)</td>
<td>0.001 (0.003)</td>
<td>0.006 (0.000)</td>
<td>0.95</td>
<td>0.914</td>
<td>3.313</td>
</tr>
<tr>
<td>BMA</td>
<td>0.101 (0.003)</td>
<td>0.016 (0.001)</td>
<td>0.64</td>
<td>0.868</td>
<td>2.198</td>
</tr>
<tr>
<td>BAC ($\omega = \infty$)</td>
<td>0.024 (0.003)</td>
<td>0.006 (0.000)</td>
<td>0.95</td>
<td>0.896</td>
<td>499.235</td>
</tr>
<tr>
<td>Adaptive LASSO</td>
<td>0.022 (0.004)</td>
<td>0.008 (0.000)</td>
<td>0.80</td>
<td>0.861</td>
<td>0.098</td>
</tr>
</tbody>
</table>
Table 4.2: Simulation results for design 2. Bias, MSE, and coverage (Cover) are for the effect of interest $\hat{\beta}_x$. Coverage is 95% confidence or credible interval coverage. CPU Time is reported in seconds on a MacBook Pro with OS X, 8 GB RAM, and 2 GHz Intel Core i7. SEs for the AUC range from less than 0.001 to 0.006, and for CPU time from less than 0.001 to 0.003.

<table>
<thead>
<tr>
<th>$n$</th>
<th>Bias</th>
<th>MSE</th>
<th>Cover</th>
<th>AUC</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True</td>
<td>-0.029</td>
<td>0.119 (0.008)</td>
<td>0.96</td>
<td>NA</td>
<td>0.003</td>
</tr>
<tr>
<td>Full</td>
<td>0.020</td>
<td>0.815 (0.062)</td>
<td>0.97</td>
<td>NA</td>
<td>0.003</td>
</tr>
<tr>
<td>Cred. Reg. (Flat Priors)</td>
<td>0.113</td>
<td>0.095 (0.006)</td>
<td>0.86</td>
<td>0.713</td>
<td>0.842</td>
</tr>
<tr>
<td>Cred. Reg. (Empirical Bayes)</td>
<td>0.114</td>
<td>0.093 (0.005)</td>
<td>0.87</td>
<td>0.718</td>
<td>0.836</td>
</tr>
<tr>
<td>Cred. Reg. (Gamma Priors)</td>
<td>0.132</td>
<td>0.095 (0.006)</td>
<td>0.88</td>
<td>0.718</td>
<td>1.855</td>
</tr>
<tr>
<td>BMA</td>
<td>0.115</td>
<td>0.080 (0.005)</td>
<td>0.91</td>
<td>0.736</td>
<td>1.143</td>
</tr>
<tr>
<td>Adaptive LASSO</td>
<td>0.151</td>
<td>0.260 (0.022)</td>
<td>0.61</td>
<td>0.532</td>
<td>0.035</td>
</tr>
<tr>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True</td>
<td>-0.008</td>
<td>0.069 (0.004)</td>
<td>0.95</td>
<td>NA</td>
<td>0.003</td>
</tr>
<tr>
<td>Full</td>
<td>-0.005</td>
<td>0.124 (0.008)</td>
<td>0.97</td>
<td>NA</td>
<td>0.004</td>
</tr>
<tr>
<td>Cred. Reg. (Flat Priors)</td>
<td>0.029</td>
<td>0.059 (0.004)</td>
<td>0.93</td>
<td>0.883</td>
<td>1.342</td>
</tr>
<tr>
<td>Cred. Reg. (Empirical Bayes)</td>
<td>0.025</td>
<td>0.060 (0.004)</td>
<td>0.94</td>
<td>0.896</td>
<td>1.346</td>
</tr>
<tr>
<td>Cred. Reg. (Gamma Priors)</td>
<td>0.034</td>
<td>0.059 (0.004)</td>
<td>0.94</td>
<td>0.887</td>
<td>2.363</td>
</tr>
<tr>
<td>BMA</td>
<td>0.155</td>
<td>0.062 (0.003)</td>
<td>0.87</td>
<td>0.760</td>
<td>1.186</td>
</tr>
<tr>
<td>Adaptive LASSO</td>
<td>0.133</td>
<td>0.091 (0.005)</td>
<td>0.65</td>
<td>0.545</td>
<td>0.036</td>
</tr>
<tr>
<td>200</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True</td>
<td>0.003</td>
<td>0.038 (0.003)</td>
<td>0.94</td>
<td>NA</td>
<td>0.003</td>
</tr>
<tr>
<td>Full</td>
<td>0.007</td>
<td>0.049 (0.003)</td>
<td>0.93</td>
<td>NA</td>
<td>0.005</td>
</tr>
<tr>
<td>Cred. Reg. (Flat Priors)</td>
<td>0.003</td>
<td>0.037 (0.002)</td>
<td>0.92</td>
<td>0.965</td>
<td>2.599</td>
</tr>
<tr>
<td>Cred. Reg. (Empirical Bayes)</td>
<td>0.001</td>
<td>0.037 (0.002)</td>
<td>0.94</td>
<td>0.975</td>
<td>2.603</td>
</tr>
<tr>
<td>Cred. Reg. (Gamma Priors)</td>
<td>0.002</td>
<td>0.036 (0.002)</td>
<td>0.94</td>
<td>0.971</td>
<td>3.637</td>
</tr>
<tr>
<td>BMA</td>
<td>0.174</td>
<td>0.055 (0.003)</td>
<td>0.75</td>
<td>0.812</td>
<td>1.267</td>
</tr>
<tr>
<td>Adaptive LASSO</td>
<td>0.115</td>
<td>0.058 (0.003)</td>
<td>0.66</td>
<td>0.621</td>
<td>0.043</td>
</tr>
<tr>
<td>500</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True</td>
<td>-0.002</td>
<td>0.013 (0.001)</td>
<td>0.95</td>
<td>NA</td>
<td>0.003</td>
</tr>
<tr>
<td>Full</td>
<td>-0.001</td>
<td>0.014 (0.001)</td>
<td>0.96</td>
<td>NA</td>
<td>0.011</td>
</tr>
<tr>
<td>Cred. Reg. (Flat Priors)</td>
<td>-0.001</td>
<td>0.013 (0.001)</td>
<td>0.95</td>
<td>0.995</td>
<td>6.379</td>
</tr>
<tr>
<td>Cred. Reg. (Empirical Bayes)</td>
<td>0.000</td>
<td>0.013 (0.001)</td>
<td>0.96</td>
<td>0.998</td>
<td>6.433</td>
</tr>
<tr>
<td>Cred. Reg. (Gamma Priors)</td>
<td>0.000</td>
<td>0.013 (0.001)</td>
<td>0.96</td>
<td>0.996</td>
<td>7.460</td>
</tr>
<tr>
<td>BMA</td>
<td>0.164</td>
<td>0.039 (0.002)</td>
<td>0.64</td>
<td>0.878</td>
<td>1.635</td>
</tr>
<tr>
<td>Adaptive LASSO</td>
<td>0.065</td>
<td>0.023 (0.001)</td>
<td>0.74</td>
<td>0.731</td>
<td>0.077</td>
</tr>
</tbody>
</table>
Chapter 5

Future Directions

In the preceding chapters we present methodology for hierarchical and spatial monotone regression and confounder selection with environmental applications. In this chapter we present ideas on future directions for the work presented here.

5.1 Extensions for Hierarchical Dose-Response Modeling

Most predictive models used to link HTS data and biological outcomes such as reproductive and developmental outcomes focus on relating one summary statistic from a dose-response relationship with an in vivo outcome estimated in rodent or other animal studies. A single summary statistic, such as potency measured by AC50, can potentially give an incomplete view of a chemicals bioactivity on a given assay. For example, two chemicals may have the same potency but one may have high efficacy while the other may have low efficacy. In the context of chemical prioritization, the chemical with low efficacy may be lower priority, a result not captured by potency alone. A potentially fruitful approach to use more information could be to extend beyond a single summary statistic to extract more information from the dose-response relationship.

One approach to include more information is to build predictive models that relate the potency, efficacy, and probability of response triplet to in vivo outcomes. Due to the large
numbers of chemicals and assays this will require some method of feature selection. A potential approach for group variable selection is discussed in Section 5.3.

An alternative approach is clustering chemicals based on the functional form of the dose-response relationships. This approach relies on the logic that chemicals with similar in vitro response across a battery of HTS assays may have similar in vivo effects. Hence, the subset of untested chemicals that clusters with chemicals that have been tested and are known to cause in vivo reactions may be more likely to have similar biological effects. Hence, their prioritization for additional testing could be proportional to the known effects of chemicals in the cluster.

A third approach is to use a functional model to relate the functional form of chemicals to known biological endpoints. This method relies on training a functional model using a subset of chemicals that have been tested to identify the bioassays and features that are associated with increased risk of in vivo activity. Then predicting the potential for in vivo effects for the set of untested chemicals.

5.2 Extensions for Multi-Pollutant Modeling

IPCC (2007) raised concerns whether current air quality management practices can adequately protect public health under future climate regimes. The synergistic ozone-temperature effect estimated in the Chapter 3 implies that ozone-related mortality may increase due to 1) increases in ambient ozone exposures and 2) increases in temperatures in combination with the synergistic ozone effect. A direct extension of this work is to estimate the risk of mortality under current and future climate scenarios to gain an understanding of ozone-related risk of mortality under these climate scenarios.

Another extension of the methodology developed in Chapter 3 is to estimate the joint effects of other co-exposures. However, there is great interest in estimating the effect of mixtures of more than two exposures. A direct expansion of the risk surface model using the outer product of basis expansions will be computationally intensive. Hence, new methodology is needed to estimate the non-linear and non-additive effects of complex multi-pollutant mixtures. Al-
ternatively, the inference made from the bivariate risk surface model may inform a simpler modeling approach for multi-pollutant modeling. For example, it may be determined from pairwise estimates of the effects of co-exposures that a quadratic model with first-order interaction, or otherwise reduced parametric model, may be sufficient for estimating the health effects of complex multi-pollutant mixtures.

5.3 Extensions for Confounder and Variable Selection

There are two important follow-up studies that will allow for this confounder selection method to be applied more broadly, including multi-city, time-series studies of air pollution: spatial confounder selection and grouped confounder selection. For example, it may be appropriate for a set of race-ethnicity indicator variables to be treated as a group rather than selected individually. Because of the spatial aspect of multi-city, time-series studies it is important to account for spatial dependence in confounder selection. In addition, many potential confounders are naturally grouped and should be selected as groups. Hence, another extension is to methodology for grouped selection. These future papers will build on the important field of fast Bayesian model selection method for big data.

In addition to confounder selection, the methodology of Bondell and Reich (2012) and that of Chapter 4 can also be extended to grouped variable selection. Grouped penalized credible region variable selection method results in a piecewise linear penalty and will result in a similar piecewise linear solution path (Rosset and Zhu, 2007). A third extension to the this work is to develop an efficient algorithm for the grouped confounder selection.
REFERENCES


over the eastern United States due to changes in global and regional climate and emissions. 


82
APPENDICES
Appendix A

Supplemental Material for Hierarchical Dose-Response Modeling for High-Throughput Toxicity Screening of Environmental Chemicals

A.1 Computation and Full Conditionals

A.1.1 Conditional Posterior Distributions

The full conditionals for ZIPPLL model are listed below. For ZILL, the updates follow along the same form except the log-linear spline is replaced with the log-linear form and adjusting their priors.
The full conditional for \( Z_{ij} \) is \( Z_{ij} | \text{rest} \sim \text{Binomial}(\hat{\psi}_{ij}) \) where

\[
\hat{\psi}_{ij} = \frac{\psi}{\psi + (1 - \psi) \exp \left\{ -\frac{1}{2} \sigma^{-2} \left[ \sum_{k=1}^{n_{ia}} (y_{ijk} - b_{ij})^2 - \sum_{k=1}^{n_{ia}} (y_{ijk} - f_2(x_{ijk}; \theta_{ij}, w_{ij}))^2 \right] \right\}}.
\]

The full conditional for \( \psi | \text{rest} \) is \( \text{Beta}(a\psi + \sum_{ij} Z_{ij}, b\psi + N - \sum_{ij} Z_{ij}) \).

The full conditional for \( b_{ij}^* | \text{rest} \) is

\[
b_{ij}^* | \text{rest} \sim \begin{cases} 
N(\bar{E}_{b,ij}, \bar{V}_{b,ij}) & \text{if } Z_{ij} = 0 \\
\pi_{b,ij} \text{TN}(-\infty, t_{ij}^*) (\bar{E}_{b,ij}, \bar{V}_{b,ij}) + (1 - \pi_{b,ij}) \text{TN}(t_{ij}^*, \infty) (\hat{E}_{b,ij}, \hat{V}_{b,ij}) & \text{if } Z_{ij} = 1
\end{cases}
\]

where

\[
\bar{V}_{b,ij} = \left[ n_{ij} \sigma^{-2} + (\Sigma^{-1})_{bb} \right]^{-1} \\
\bar{E}_{b,ij} = \bar{V}_{b,ij} \left[ \sigma^{-2} \sum_{k=1}^{n_{ij}} y_{ijk} + \mu_b (\Sigma^{-1})_{bb} - (\Sigma^{-1})_{-b,b} (\theta_{ij,-b} - \mu_b) \right] \\
\hat{V}_{b,ij} = \left[ (\Sigma^{-1})_{bb} + \sigma^{-2} \sum_{k=1}^{n_{ij}} \left( 1 + (x_{ijk}/a_{ij})^{w_{ij}} \right)^2 \right]^{-1} \\
\hat{E}_{b,ij} = \hat{V}_{b,ij} \left\{ \mu_b (\Sigma^{-1})_{bb} - (\Sigma^{-1})_{-b,b} (\theta_{ij,-b} - \mu_b) \right. \\
\left. + \sigma^{-2} \sum_{k=1}^{n_{ij}} \frac{1}{1 + \exp \left\{ [\Psi (x_{ijk}, a_{ij})]^T w_{ij} \} \right\}} \times \left[ y_{ijk} - t_{ij} \left( 1 - \frac{1}{1 + \exp \left\{ [\Psi (x_{ijk}, a_{ij})]^T w_{ij} \} \right\} \right) \right] \left. \right\} \\
\hat{E}_{b,ijk} = t_{ij} \left[ 1 - (x_{ijk}/a_{ij})^{w_{ij}} \right]^{-1} \\
\pi_{b,ij} = \frac{F(t_{ij}^*; \bar{E}_{b,ij}, \bar{V}_{b,ij})}{N(0; \bar{E}_{b,ij}, \bar{V}_{b,ij})} + \frac{1 - F(t_{ij}^*; \bar{E}_{b,ij}, \bar{V}_{b,ij})}{N(0; \bar{E}_{b,ij}, \bar{V}_{b,ij})} \prod_{k=1}^{n_{ij}} \frac{N(0; \hat{E}_{b,ijk}^*, \sigma^2)}{N(0; \bar{E}_{b,ijk}, \bar{V}_{b,ij})}
\]

and \( F(\cdot; \mu, \sigma^2) \) and \( N(\cdot; \mu, \sigma^2) \) are the normal distribution and density functions with mean \( \mu \) and variance \( \sigma^2 \).
The full conditional for $t^*_{ij}$ is

$$t^*_{ij} | \text{rest} \sim \begin{cases} 
N(\tilde{E}_{t,ij}, \tilde{V}_{t,ij}) & \text{if } Z_{ij} = 0 \\
\pi_{t,ij} \text{TN}(-\infty, b^*_{ij}) (\tilde{E}_{t,ij}, \tilde{V}_{t,ij}) + \left(1 - \pi_{t,ij}\right) \text{TN}(b^*_{ij}, \infty)(\hat{E}_{t,ij}, \hat{V}_{t,ij}) & \text{if } Z_{ij} = 1
\end{cases}$$

where

$$\tilde{V}_{t,ij} = [(\Sigma^{-1})_{tt}]^{-1}$$

$$\tilde{E}_{t,ij} = \tilde{V}_{t,ij} \left[ \mu_t(\Sigma^{-1})_{tt} - (\Sigma^{-1})_{t,t} (\theta^*_{ij,-t} - \mu_t) \right]$$

$$\hat{V}_{t,ij} = \left[ (\Sigma^{-1})_{tt} + \sigma^{-2} \sum_{k=1}^{n_{ij}} \left( 1 - \frac{1}{1 + \exp \{ [\Psi(x_{ijk}, a_{ij})]^T w_{ij} ] \} } \right) \right]^{-1}$$

$$\hat{E}_{t,ij} = \hat{V}_{t,ij} \left[ \mu_t(\Sigma^{-1})_{tt} - (\Sigma^{-1})_{t,t} (\theta^*_{ij,-t} - \mu_t) \right.$$  
$$\left. + \sigma^{-2} \sum_{k=1}^{n_{ij}} \left( 1 + \exp \{ [\Psi(x_{ijk}, a_{ij})]^T w_{ij} ] \} \right) \right]$$

$$\times \left( y_{ijk} - \frac{b_{ij}}{1 + \exp \{ [\Psi(x_{ijk}, a_{ij})]^T w_{ij} ] \} } \right)$$

$$\hat{E}^*_{t,ijk} = b_{ijk} (1 + (x_{ijk}/a_{ijk})^{w_{ijk}})^{-1}$$

$$\pi_{t,ij} = \frac{F(b_{ij}; \hat{E}_{t,ij}, \hat{V}_{t,ij})}{N(0; \hat{E}_{t,ij}, \hat{V}_{t,ij})} \prod_{k=1}^{n_{ij}} N(y_{ijk}; b_{ij}, \sigma^2)$$

$$\frac{1 - F(b_{ij}; E^*_{t,ijk}, V^*_{t,ijk})}{N(0; E^*_{t,ijk}, V^*_{t,ijk})} \prod_{k=1}^{n_{ij}} N(y_{ijk}; \hat{E}^*_{t,ijk}, \sigma^2)$$
The remaining full conditionals for the conjugate parameters are:

\[
\begin{align*}
\sigma^{-2}|_{\text{rest}} & \sim \text{Ga} \left( d + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{A} n_{ij}, d_2 + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{A} \sum_{k=1}^{n_{ij}} \left( y_{ijk} - f(x_{ijk}; \theta_{ij}, w_{ij}) \right)^2 \right) \\
\mu|_{\text{rest}} & \sim \text{N}_3 \left( \left[ \Sigma_0^{-1} + NA \Sigma^{-1} \right]^{-1} \left[ \Sigma_0^{-1} \mu_0 + \Sigma^{-1} \sum_{i=1}^{N} \sum_{j=1}^{A} \theta_{ij}^r \right], \left[ \Sigma_0^{-1} + NA \Sigma^{-1} \right]^{-1} \right) \\
\Sigma|_{\text{rest}} & \sim \text{InvWish} \left( NA + \nu_2, V_2 + \sum_{i=1}^{N} \sum_{j=1}^{A} \left( \theta_{ij}^r - \mu \right) \left( \theta_{ij}^r - \mu \right)^T \right) \\
\lambda|_{\text{rest}} & \sim \text{Gamma} \left( g_1 + \frac{K}{2}, g_2 + \frac{1}{2} \sum_{k=1}^{K} (S_k - S_{k-1})^2 \right) \\
S_k|_{\text{rest}} & \sim \text{N} \left( \left[ NA \Sigma^{-1}_S + \Delta \lambda \right]^{-1} \left[ \Sigma^{-1}_S \sum_{i=1}^{N} \sum_{j=1}^{A} \log(w_{ij}) + \Delta \lambda \xi \right], \left[ NA \Sigma^{-1}_S + \Delta \lambda \right]^{-1} \right) \\
\Sigma_S|_{\text{rest}} & \sim \text{InvWish} \left( NA + \nu_s, V_s + \sum_{i=1}^{N} \sum_{j=1}^{A} \log(w_{ij}) (\bar{w}) (\log(w_{ij}) - \bar{w})^T \right)
\end{align*}
\]

where \( \bar{w} = (NA)^{-1} \sum_{i=1}^{N} \sum_{j=1}^{A} \log(w_{ij}) \), \( \Delta \lambda = \text{diag}(2\lambda, \ldots, 2\lambda, \lambda) \) is \( K \times K \) and \( \xi = ((S_0 + S_2)/2, (S_1 + S_3)/2, \ldots, (S_{K-2} + S_K)/2, S_{K-1})^T \) is a \( K \times 1 \) vector of the means of the previous and next slope.

Finally, \( a_{ij}^r \) and \( \log(w_{ij}) \) do not have closed form full conditionals but are updated with a resolvent random-walk Metropolis-Hastings transition kernel with a normal proposal distribution. The acceptance rate for \( \log(w_{ij}) \) is

\[
R_w = \frac{\mathcal{N}(\log(\hat{w}_{ij}); \mathbf{S}, \Sigma_S) \prod_{k=1}^{n_{ij}} \mathcal{N} \left( y_{ijk}; f(x_{ijk}; \theta_{ij}, \hat{w}_{ij}), \sigma^2 \right)}{\mathcal{N}(\log(w_{ij}); \mathbf{S}, \Sigma_S) \prod_{k=1}^{n_{ij}} \mathcal{N} \left( y_{ijk}; f(x_{ijk}; \theta_{ij}, w_{ij}), \sigma^2 \right)}
\]

and the acceptance rate for \( a_{ij}^r \) is

\[
R_a = \frac{\mathcal{N} \left( \hat{\theta}_{ij}; \mu, \Sigma \right) \prod_{k=1}^{n_{ij}} \mathcal{N} \left( y_{ijk}; f(x_{ijk}; \hat{\theta}_{ij}, \hat{w}_{ij}), \sigma^2 \right)}{\mathcal{N} \left( \theta_{ij}; \mu, \Sigma \right) \prod_{k=1}^{n_{ij}} \mathcal{N} \left( y_{ijk}; f(x_{ijk}; \theta_{ij}, w_{ij}), \sigma^2 \right)},
\]

where \( \mathcal{N}(\cdot; \mu, \sigma^2) \) denotes the normal pdf with mean \( \mu \) and variance \( \sigma^2 \).
The resolvant transition Robert and Casella (2004) kernel does multiple Metropolis-Hastings updated on a parameter. The number of updates performed each iteration is stochastic. This an effective method to reduce autocorrelation. The updates work as follows.

\[ u \sim \text{Geometric}\{\text{mean}=1/(1-\epsilon)\} \]

for(i in 1 to u) do:

Propose a new value, in our case with a random walk

Calculate the Metropolis-Hastings the acceptance ratio \( R \)

Accept or reject the proposed value with acceptance probability \( R \)
end loop

A.1.2 Priors for ZILL

In the simulation and cross validation study we compare ZIPLL to the simplified version ZILL, the zero-inflated log-logistic model. ZILL replaces the piecewise linear spline with a log-linear form and active responses follow FPPLL,

\[
f(x_{ijk}; \theta_{ij}, w_{ij}, Z_{ij}) = \begin{cases} t_{ij} - (t_{ij} - b_{ij}) \times \text{Logit} \left[-w \{\log(x) - \log(a)\}\right] & \text{if } Z_{ij} = 1 \\ b_{ij} & \text{if } Z_{ij} = 0. \end{cases} \tag{A.1}
\]

Because there is only one slope \( w_{ij} \) for each curve, we simplify the model by including \( w_{ij} \) in \( \theta_{ij} \). Hence, for ZILL, \( \theta_{ij} = (t_{ij}, b_{ij}, a_{ij}, w_{ij}) \) and \( \theta_{ij}^* = (t_{ij}^*, b_{ij}, a_{ij}^*, w_{ij}^*) \). Computation is the same as ZIPLL.

A.2 Simulation

In order to evaluate ZIPLL, we conducted a simulation study. We generated 100 data sets from \( y_{ijk} \sim N(h_m(x_{ijk}), \sigma^2) \) for chemical \( i = 1, \ldots, 25 \), assay \( j = 1, \ldots, 4 \), two sample sizes of \( n_{ij} \in \{8, 24\} \). We use eight concentrations characteristic of the ToxCast data currently being analyzed in house from the Attagene platform, \( x_{ijk} \in \{0.091, 0.270, 0.820, 2.500, \ldots\} \).
7.400, 22.000, 67.000, 200.000}, generating one or three responses at each concentration for the two values of \( n_{ij} \), respectively. The two mean functions are:

\[
   h_{LL}(x_{ijk}; t_{ij}, b_{ij}, a_{ij}, w_{ij}, Z_{ij}) = b_{ij}^{(1-Z_{ij})} \left( t_{ij} - \frac{t_{ij} - b_{ij}}{1 + \exp \left[ w_{ij} \{ \log(x_{ijk}) - \log(a_{ij}) \} \right]} \right)^{Z_{ij}} \tag{A.2}
\]

and

\[
   h_{Mix}(x_{ijk}; t_{ij}, b_{ij}, a_{ij}, w_{ij}, Z_{ij}) = b_{ij}^{(1-Z_{ij})} \left[ t_{ij} - (t_{ij} - b_{ij}) \left\{ 1 - p_{ij} F \left( -\log x_{ijk}; c_{ij}, 1.5^2 \right) \right. \right. \\
   \left. \left. - (1 - p_{ij}) F \left( -\log x_{ijk}; d_{ij}, 0.5^2 \right) \right\} \right]^{Z_{ij}} \tag{A.3}
\]

where \( F(\cdot; \mu, \sigma^2) \) is the normal cdf with mean \( \mu \) and variance \( \sigma^2 \). Equation (A.2) is the log-logistic ("LL") mean response assumed by ZILL, while the mixture of normal CDFs ("Mix") in equation (A.3) is not a log-logistic model, and has a slower increase at low concentrations followed by a steeper increase at later concentrations than the log-logistic model. The top and bottom parameters for both mean functions are generated as \( t^{*}_{ij} \sim N(4,4^2) \) and \( b_{ij} \sim N(1,0.1^2) \).

For \( h_{LL} \), \( a_{ij} \sim TN_{[0,\infty]}(10,30^2) \) and \( w_{ij} \sim TN_{[1,8]}(2,1^2) \). For \( h_{Mix} \), \( p_{ij} \sim Unif(0.1,0.4) \), \( c_{ij} \sim TN_{[-2,0]}(-1,1^2) \), and \( d_{ij} \sim TN_{[1,5]}(3,1^2) \). Figure A.1b shows the 100 dose-response curves in one of the simulated datasets from each data model.

The first simulation design uses mean function (A.2) with \( Z_{ij} \sim Bern(0.9) \) and is denoted by \( h_{LL}^{0.9} \), with the mean function indicated by the subscript and \( \Pr(Z_{ij} = 1) \) indicated in superscript. The second design (\( h_{Mix}^{0.9} \)) uses mean function (A.3) and \( Z_{ij} \sim Bern(0.9) \). Under these models, the probability of an active response for a curve is 0.67. We use a third design with mean function (A.2) and \( Z_{ij} \sim Ber(0.5) \) (\( h_{LL}^{0.5} \)). In this model the proportion of curves that are active is 0.37. In both cases, we consider a curve active if the maximum response is at least one fold greater than a minimum response. This is the assumption used for classifying curves in the ToxCast project.

The fitted models for the simulation were run without chemical and assay effects. Instead, we modeled \( Z_{ij} | \psi \sim Bern(\psi) \) and \( \psi \sim Beta(1,1) \). The priors for \( \mu, \Sigma, S, \lambda, \) and \( \sigma^2 \) are as specified
Figure A.1: Example of simulated regression functions under using ZILL A.1a and mixture of normals A.1b.
in Section 3 of Chapter 2. Using these priors, we ran our algorithm for 20,000 iterations and discarded the first 5,000 for burnin. For the resolvant kernel we use $\epsilon = 0.8$.

For comparison, we used a least squares fitting of the FPLL model using \texttt{nlm} in R (LS). This method is similar to the current methodology used by ToxCast. For a Bayesian, nonhierarchical comparison we used the monotone linear splines (MLS) model proposed by Neelon and Dunson (2004). This model assumes linear basis functions similar to the positive part of our basis functions but with knots at each unique data point. The slopes are modeled $\beta_k = \max(\beta_k^*, 0)$ and $\beta_0^* \sim N(1, 5^2)$, $\beta_k^* \sim N(\beta_{k-1}^*, \lambda)$, and $\lambda \sim \text{Gamma}(0.5, 1)$. The inverse error variance is modeled $\sigma^{-2} \sim \text{Gamma}(1, .5)$ and random intercept is $\alpha \sim N(1, 0.05^2)$.

After fitting the four models on each simulated data set, we compared the methods’ abilities to estimate the mean function, AC50, Emax, and whether or not an active response occurred. We estimated the pointwise root mean square error (RMSE) for the overall fit at 50 gridpoints evenly spaced on the log base 3 scale, the same spacing used for the true concentration. We also compare the RMSE for the AC50 and top of the curve. Table A.1 shows the simulation results.

When ZILL is true (design $h_{LL}^{0.9}$ and $h_{LL}^{0.5}$) both ZILL and ZIPLL have lower overall RMSE than the other models, demonstrating the benefit of the hierarchical approach. ZILL, ZIPLL, and MLS all maintain the nominal credible interval coverage probability. For the mixture of normals link data ($h_{Mix}^{0.9}$), ZIPLL has lowest RMSE suggesting it is more robust to misspecification of shape of the dose-response curve.

For inference on the top of the curve the four methods performed similarly when ZILL was the true model, but ZIPLL had the lowest RMSE for the top for $h_{Mix}^{0.9}$ at the smaller sample size. Estimation of the AC50 follows a similar story. The most notable difference in the methods is for the mixture of normals model where ZIPLL has lower RMSE for the AC50 than the other methods and tighter credible intervals. Both ZILL and ZIPLL have credible interval coverage above 0.95 for both sample sizes under $h_{Mix}^{0.9}$, but ZIPLL has smaller mean interval width (47.6 versus 63.5 for the smaller sample size and 29.9 versus 44.3 at the larger sample size). MLS has smaller intervals but had coverage of 0.82 and 0.79 at the two sample sizes, respectively. Again,
this shows the robustness of the ZIPLL model.

We compared the ability to estimate the probability of active response with the area under the ROC curve (AUC). In all cases, ZILL, ZIPLL, and MLS performed excellently, while LS, the only model not to estimate a probability of an active response, lagged slightly behind. This result highlights the information added by estimating the probability of an active response instead of a binary classification.

Table A.1: Simulation results for mean response and probability of active response. The displayed values are the mean (standard error) across simulated data sets. RMSE is the pointwise root mean square error taken over 50 evenly spaced points on log base 3, the spacing scale for concentrations in the real data. RMSE Top and RMSE AC50 are only calculated for dose-response curves that are active. AUC is the area under the ROC curve.

<table>
<thead>
<tr>
<th>Design</th>
<th>n_{ij}</th>
<th>Model</th>
<th>RMSE</th>
<th>RMSE Top</th>
<th>RMSE AC50</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>h_{0.9}</td>
<td>8</td>
<td>ZILL</td>
<td>0.257 (0.001)</td>
<td>0.565 (0.007)</td>
<td>14.787 (0.383)</td>
<td>0.996 (0.000)</td>
</tr>
<tr>
<td>h_{0.9}</td>
<td>8</td>
<td>ZIPLL</td>
<td>0.251 (0.001)</td>
<td>0.569 (0.011)</td>
<td>13.242 (0.311)</td>
<td>0.996 (0.000)</td>
</tr>
<tr>
<td>h_{0.9}</td>
<td>8</td>
<td>MLS</td>
<td>0.367 (0.003)</td>
<td>0.875 (0.008)</td>
<td>17.099 (0.267)</td>
<td>0.995 (0.000)</td>
</tr>
<tr>
<td>h_{0.9}</td>
<td>8</td>
<td>LS</td>
<td>0.443 (0.007)</td>
<td>0.530 (0.005)</td>
<td>15.066 (0.296)</td>
<td>0.929 (0.003)</td>
</tr>
<tr>
<td>h_{0.9}</td>
<td>24</td>
<td>ZILL</td>
<td>0.166 (0.001)</td>
<td>0.381 (0.006)</td>
<td>11.055 (0.266)</td>
<td>0.999 (0.000)</td>
</tr>
<tr>
<td>h_{0.9}</td>
<td>24</td>
<td>ZIPLL</td>
<td>0.163 (0.001)</td>
<td>0.432 (0.021)</td>
<td>10.024 (0.343)</td>
<td>0.998 (0.000)</td>
</tr>
<tr>
<td>h_{0.9}</td>
<td>24</td>
<td>MLS</td>
<td>0.238 (0.002)</td>
<td>0.500 (0.006)</td>
<td>11.465 (0.244)</td>
<td>0.998 (0.000)</td>
</tr>
<tr>
<td>h_{0.9}</td>
<td>24</td>
<td>LS</td>
<td>0.209 (0.001)</td>
<td>0.357 (0.004)</td>
<td>10.900 (0.323)</td>
<td>0.972 (0.002)</td>
</tr>
<tr>
<td>h_{0.9}</td>
<td>8</td>
<td>ZILL</td>
<td>0.212 (0.002)</td>
<td>0.598 (0.009)</td>
<td>15.944 (0.432)</td>
<td>0.998 (0.000)</td>
</tr>
<tr>
<td>h_{0.9}</td>
<td>8</td>
<td>ZIPLL</td>
<td>0.207 (0.001)</td>
<td>0.618 (0.014)</td>
<td>14.699 (0.399)</td>
<td>0.998 (0.000)</td>
</tr>
<tr>
<td>h_{0.9}</td>
<td>8</td>
<td>MLS</td>
<td>0.286 (0.003)</td>
<td>0.873 (0.011)</td>
<td>17.454 (0.383)</td>
<td>0.997 (0.000)</td>
</tr>
<tr>
<td>h_{0.9}</td>
<td>8</td>
<td>LS</td>
<td>0.396 (0.006)</td>
<td>0.539 (0.007)</td>
<td>15.707 (0.402)</td>
<td>0.945 (0.003)</td>
</tr>
<tr>
<td>h_{0.9}</td>
<td>24</td>
<td>ZILL</td>
<td>0.139 (0.001)</td>
<td>0.400 (0.008)</td>
<td>12.455 (0.385)</td>
<td>1.000 (0.000)</td>
</tr>
<tr>
<td>h_{0.9}</td>
<td>24</td>
<td>ZIPLL</td>
<td>0.136 (0.001)</td>
<td>0.458 (0.017)</td>
<td>11.030 (0.337)</td>
<td>0.999 (0.000)</td>
</tr>
<tr>
<td>h_{0.9}</td>
<td>24</td>
<td>MLS</td>
<td>0.190 (0.002)</td>
<td>0.495 (0.008)</td>
<td>10.996 (0.306)</td>
<td>0.999 (0.000)</td>
</tr>
<tr>
<td>h_{0.9}</td>
<td>24</td>
<td>LS</td>
<td>0.194 (0.001)</td>
<td>0.357 (0.006)</td>
<td>10.920 (0.438)</td>
<td>0.983 (0.002)</td>
</tr>
<tr>
<td>h_{0.9}</td>
<td>8</td>
<td>ZILL</td>
<td>0.337 (0.002)</td>
<td>0.748 (0.013)</td>
<td>15.642 (0.396)</td>
<td>0.995 (0.000)</td>
</tr>
<tr>
<td>h_{0.9}</td>
<td>8</td>
<td>ZIPLL</td>
<td>0.273 (0.001)</td>
<td>0.446 (0.010)</td>
<td>13.656 (0.320)</td>
<td>0.995 (0.000)</td>
</tr>
<tr>
<td>h_{0.9}</td>
<td>8</td>
<td>MLS</td>
<td>0.379 (0.003)</td>
<td>0.694 (0.009)</td>
<td>18.314 (0.407)</td>
<td>0.995 (0.000)</td>
</tr>
<tr>
<td>h_{0.9}</td>
<td>8</td>
<td>LS</td>
<td>0.462 (0.003)</td>
<td>0.525 (0.006)</td>
<td>17.082 (0.432)</td>
<td>0.924 (0.003)</td>
</tr>
<tr>
<td>h_{0.9}</td>
<td>24</td>
<td>ZILL</td>
<td>0.251 (0.001)</td>
<td>0.665 (0.013)</td>
<td>13.387 (0.303)</td>
<td>0.998 (0.000)</td>
</tr>
<tr>
<td>h_{0.9}</td>
<td>24</td>
<td>ZIPLL</td>
<td>0.180 (0.001)</td>
<td>0.351 (0.022)</td>
<td>10.263 (0.289)</td>
<td>0.999 (0.000)</td>
</tr>
<tr>
<td>h_{0.9}</td>
<td>24</td>
<td>MLS</td>
<td>0.242 (0.001)</td>
<td>0.323 (0.004)</td>
<td>12.171 (0.305)</td>
<td>0.998 (0.000)</td>
</tr>
<tr>
<td>h_{0.9}</td>
<td>24</td>
<td>LS</td>
<td>0.281 (0.001)</td>
<td>0.423 (0.005)</td>
<td>13.733 (0.417)</td>
<td>0.969 (0.002)</td>
</tr>
</tbody>
</table>
A.3 Additional Figure

Figure A.2: Comparison of the posterior probability of an active response estimated with ZIPLL (x-axis) and the binary indicator of active response from ToxCast (y-axis) for all 309 chemicals on three assays. The majority of chemicals are considered not active (clustered in bottom left) or active (upper right) with both methods. However, using ZIPLL we estimated that several chemicals have posterior probabilities of response between 0.1 and 0.9, suggesting that there is not conclusive evidence that these chemicals responded or not, but are forced to be classified as either active or not in the ToxCast data.
Appendix B

Supplemental Material for Modeling the Effect of Temperature on Ozone Related Mortality

B.1 Additional Figures

This section contains the additional figures referred to in the introduction and in Section 2 of Chapter 3.

B.2 Cross-Validation Results

This section contains the additional figures referred to in Section 3.1 of Chapter 3.
(a) Distribution of ozone by city.

(b) Distribution of temperature by city.

(c) Distribution of ozone on days above and below the city-specific median temperature.

(d) National ozone-temperature distribution.

Figure B.1: Summary of the ozone and temperature distribution.
Figure B.2: Demonstration of the different basis expansions used in the first and second stages. The bottom left shows the observed ozone-temperature distribution in each city. To the right are the basis functions in the temperature direction and to the top are the basis functions in the ozone direction for each stage. The first stage basis functions are different in each city.
Figure B.3: Difference in deviance from the linear, additive model (NMMAPS model) for the spatial monotone model with different numbers of basis functions in the ozone (M1) and temperature (M2) directions.
Figure B.4: Average rank of CV results across Figures B.3a through B.3d, with one being the best performing model. The best model on average was $(M_1, M_2) = (7, 9)$ with an average rank of 3.75; this is the model used for analysis. The next best models were $(9, 10), (10, 9), \text{ and } (9, 9)$ with average ranks of 5.25, 5.50, and 6.75, respectively.

B.3 Full Conditional Distribution

Let $p_1 = M_1 + 1, \ p_2 = M_2 + 1, \ p = p_1 \ast p_2, \ \mathbf{S} = \mathbf{S}_2 \otimes \mathbf{S}_1$. The posterior takes the form

\[
L(\mathbf{\theta}, \mathbf{\mu}, \rho, \mathbf{S}_2, \mathbf{S}_1 | \mathbf{\beta}_{c_1}, \ldots, \mathbf{\beta}_{c_N}, \mathbf{V}_1^{-1}, \ldots, \mathbf{V}_N^{-1}) \propto \left| \mathbf{S}_2 \right|^{-(N+1)p_1 + p_2 + 1/2} \left| \mathbf{S}_1 \right|^{-(N+1)p_2 + p_1 + 1/2} |\mathbf{R}|^{-p/2} \tau^{-p/2 + a_\tau - 1} \\
\times \exp\left\{-\frac{1}{2} \text{tr}\left\{ \mathbf{S}_2^{-1} \right\} - \frac{1}{2} \text{tr}\left\{ \mathbf{S}_1^{-1} \right\} - \frac{1}{2} \mathbf{\mu}^T \left( \mathbf{1}_{p \times 1} \right)^T \mathbf{R}^{-1} \left( \mathbf{\mu} - \mathbf{1}_{p \times 1} \right) \right. \\
- \frac{1}{2} \sum_c \left( \mathbf{\beta}_c - \mathbf{A}_c \mathbf{\theta}_c \right)^T \mathbf{V}_c^{-1} \left( \mathbf{\beta}_c - \mathbf{A}_c \mathbf{\theta}_c \right) - \frac{1}{2} \mathbf{\theta}_*^T \mathbf{R}^{-1} \mathbf{S}^{-1} \mathbf{\theta}_* - \mathbf{\mu}_*^T \mathbf{R}^{-1} \mathbf{S}^{-1} \mathbf{\mu}_* \\
- \frac{(\log \rho - \mu_\rho)^2}{2\sigma_\rho^2} - \frac{\mu_0^2}{2\tau_0} - b_\tau \tau^{-1} \right\}.
\]

Where $\mathbf{\theta}_* = (\mathbf{\theta}_*(c_1)^T, \ldots, \mathbf{\theta}_*(c_N)^T)^T$ and $\mathbf{\mu} = (\mathbf{\mu}(c_1)^T, \ldots, \mathbf{\mu}(c_N)^T)^T$ are $Np \times 1$ vectors.
### B.3.1 Full Conditionals

\[ \mu|\text{rest} \sim \mathcal{N}\left( \frac{1}{\tau} + \frac{1}{N} \mathbf{R}^{-1} \mathbf{1}_N \right)^{-1} \left( \frac{1}{\tau} \mu_0 + \frac{1}{N} \sum_{c=1}^{N} \theta_c \mathbf{R}^{-1}_c \mathbf{1}_N \right), \]

\[ \left( \frac{1}{\tau} + \frac{1}{N} \mathbf{R}^{-1} \mathbf{1}_N \right)^{-1} \mathbf{S} \] (B.2)

\[ \mu_0|\text{rest} \sim \mathcal{N}\left( \tau^{-1} \mathbf{1}_p \mathbf{S}^{-1} \mathbf{1}_p + \tau_0^{-1} \right)^{-1} \left( \tau^{-1} \mathbf{1}_p \mathbf{S}^{-1} \mu_0 + \tau_0^{-1} \right)^{-1} \] (B.3)

\[ \tau^{-1}|\text{rest} \sim \text{Ga}\{a_\tau + p/2, b_\tau + \left( \mu - \mathbf{1}_p \mu_0 \right)^T \mathbf{S}^{-1} \left( \mu - \mathbf{1}_p \mu_0 \right) / 2 \} \] (B.4)

\[ \mathbf{S}_2|\text{rest} \sim \text{IW}\{(N + 1) p_1 + p_2 + 1, \]

\[ \mathbf{I} + (\hat{\theta} - \bar{\mu}) (\mathbf{R} \otimes \mathbf{S}_1)^{-1} (\hat{\theta} - \bar{\mu})^T + \tau^{-1} (\bar{\mu} - \bar{\mu}_0) \mathbf{S}_1^{-1} (\bar{\mu} - \bar{\mu}_0)^T \} \] (B.5)

\[ \mathbf{S}_1|\text{rest} \sim \text{IW}\{(N + 1) p_2 + p_1 + 1, \]

\[ \mathbf{I} + (\hat{\theta} - \bar{\mu}) (\mathbf{R} \otimes \mathbf{S}_2)^{-1} (\hat{\theta} - \bar{\mu})^T + \tau^{-1} (\bar{\mu} - \bar{\mu}_0) \mathbf{S}_2^{-1} (\bar{\mu} - \bar{\mu}_0)^T \} \] (B.6)

\[ \theta_{jk,c}|\text{rest} \sim (1 - \delta) \text{TN}_{(-\infty, 0)} \left( \mathbf{0}; \mathbf{E}_b, \mathbf{V}_b \right) + \delta \text{TN}_{(0, -\infty)} \left( \mathbf{0}; \mathbf{E}_a, \mathbf{V}_a \right) \] (B.7)

where

\[ \mathbf{V}_a = \left[ \mathbf{S}_{1,jj}^{-1} \mathbf{S}_{2,kk}^{-1} \mathbf{R}_{cc}^{-1} + [\mathbf{A}_c^T \mathbf{V}^{-1}_c \mathbf{A}_c]_{(jk), (jk)} \right]^{-1} \] (B.8)

\[ \mathbf{E}_a = \mathbf{V}_a \left[ \mathbf{S}_{1,jj}^{-1} \mathbf{S}_{2,kk}^{-1} \mathbf{R}_{cc}^{-1} \mu_{jk} + \sum_{k'} \mathbf{S}_{1,jj'}^{-1} \mathbf{S}_{2,kk'}^{-1} \mathbf{R}_{cc}^{-1} \left( \mu_{j'k'} - \theta_{j'k'}^*(c') \right) \right] \]

\[ + [\mathbf{A}_c^T \mathbf{V}^{-1}_c]_{c'(jk)[jk]} \left( \hat{\theta}_{c'(jk)} - (\mathbf{A}_c \theta(c'))_{(jk)} \right) + \mathbf{A}_c^T \mathbf{V}^{-1}_c \] (B.9)

\[ \mathbf{V}_b = \left[ \mathbf{S}_{1,jj}^{-1} \mathbf{S}_{2,kk}^{-1} \mathbf{R}_{cc}^{-1} \right]^{-1} \] (B.10)

\[ \mathbf{E}_b = \mathbf{V}_b \left[ \mathbf{S}_{1,jj}^{-1} \mathbf{S}_{2,kk}^{-1} \mathbf{R}_{cc}^{-1} \mu_{jk} + \sum_{k'} \mathbf{S}_{1,jj'}^{-1} \mathbf{S}_{2,kk'}^{-1} \mathbf{R}_{cc}^{-1} \left( \mu_{j'k'} - \theta_{j'k'}^*(c') \right) \right] \] (B.11)
and $\mathcal{K} = \{(j', k', c') : \{j' \neq j\} \cup \{k' \neq k\} \cup \{c' \neq c\}\}$. For the truncated parameters

$$
\delta = \frac{1 - F(x; E_a, V_a)}{f(x; E_a, V_a)} \frac{1 - F(x; E_b, V_b)}{f(x; E_b, V_b)},
$$

(B.12)

and $\delta = 1$ for unconstrained intercept, $j = 0$.

The acceptance probability for the proposed range parameter $\rho'$ proposed from $\rho' \sim N(\rho, s^2)$ centered on the current value $\rho$ with proposal variance $s^2$ is

$$
r_{\rho} = \frac{|R'|^{-p/2} \exp \left\{ -\frac{1}{2} (\theta^* - \mu)^T (R' \otimes S)^{-1} (\theta^* - \mu) - \frac{(\log \rho' - \mu^2 \rho)^2}{2\sigma_{\rho}^2} \right\}}{|R|^{-p/2} \exp \left\{ -\frac{1}{2} (\theta^* - \mu)^T (R \otimes S)^{-1} (\theta^* - \mu) - \frac{(\log \rho - \mu^2 \rho)^2}{2\sigma_{\rho}^2} \right\}},
$$

(B.13)

where $R$ is the current correlation matrix and $R'$ is the proposed correlation matrix.
B.4 MCMC algorithm

The MCMC algorithm proceeds as follows.

**Step 1:** Generate a random integer \( r \) from a geometric distribution with mean \((1 - \epsilon_\rho)/\epsilon_\rho\).

Propose a new \( \log \rho \) from a normal and accept the proposed values with probability (B.13) \( r \) times.

**Step 2:** For each city in \( c = 1, \ldots, n \) do the following:

  **Step2a:** Generate a random integer \( r \) from a geometric distribution with mean \((1 - \epsilon_\theta)/\epsilon_\theta\).

  **Step2b:** Do the following \( r \) times.

    for(\( j \) in 0 to \( M_1 \)){
      for(\( k \) in 0 to \( M_2 \)){
        Calculate \( E_a, V_a, E_b, V_b \), and \( \delta \).
        Draw a \( \theta_{j,k,c} \) from (B.7).
      }
    }

**Step 3:** Update \( S_1 \) and \( S_2 \) from (B.6) and (B.7).

**Step 4:** Update \( \mu \) from (B.2).

**Step 5:** Update \( \mu_0 \) from (B.3).

**Step 6:** Update \( \mu \) from (B.4).

In steps 1 and 2, we choose a random number of updates for the range and for each city’s basis coefficients. This approach substantially reduced autocorrelation and improves performance. We used \( \epsilon_\rho = 0.8 \) and \( \epsilon_\theta = 0.9 \).
B.5 Trace Plots

This section contains a random sample of trace plots as well as the trace plot for the range.

Figure B.5: Trace plots of the point wise estimates of the risk surface.
Figure B.6: Trace of the posterior sample of log $\rho$ (log range in log kilometers). The trace plot only includes the posterior sample and does not show the portion of the burn in portion of the chain. The posterior mean is 922.9, the median is 659.4, and the 95% symmetric interval is (291.8, 3134.5).
Appendix C

Supplemental Material for
Confounder Selection via Penalized
Credible Regions

C.1 Tuning

We first compare various tuning criteria for selecting $\lambda$. Figure C.1 compares the model selection performance using the empirical Bayes prior with different selection criteria: forward selection, CV, $C_p$, AIC, and BIC. Even as the area under the ROC curve (AUC) approaches one (Figure C.1a), the criteria give different sensitivity, false selection rate, and model size (Figures C.1b through C.1d). At the smallest sample sizes forward selection with $\alpha_{fs} = 0.25$ resulted in the lowest MSE and bias for the effect of interest $\hat{\beta}_x$, but had one of the largest average model size. For more moderate sample sizes all methods except BIC, which selected too few covariates, had similar MSE and bias.

In general, a conservative choice for model selection is preferred because the errors resulting from including too many covariates are less severe than from too few. Confounding variables that are highly correlated with $X$ may not reduce the sum of squared errors after $X$ is included
in the model. Methods such as AIC, BIC, and CV that focus on minimizing sum of squared errors may not properly reward models that include these confounding variables. By performing forward selection on both the outcome and exposure models such variables are rewarded with forward selection because of their value to the exposure model, regardless of their effect on the outcome model fit. This makes forward selection appealing in this setting with correlated covariates. We proceed using $\alpha^{ts} = 0.25$.

### C.2 Proof of Theorem 1

The proof of Theorem 1 follows closely to that of Theorem 2 of Zou (2006). We start by proving asymptotic normality. We now let the true coefficients be denoted by $\beta^*$ and $\gamma^*$. Let

$$
\beta = \beta^* + \nu / \sqrt{n}
$$

and

$$
\Psi_n(\nu) = \left\{ \hat{\beta} - (\beta^* + \nu / \sqrt{n}) \right\}^T \Sigma_y^{-1} \left\{ \hat{\beta} - (\beta^* + \nu / \sqrt{n}) \right\} + \lambda_n \sum_{j=1}^{p} \hat{w}_j \left| \beta^*_j + \nu_j \right|.
$$

Let $\bar{v}^{(n)} = \arg \min \Psi_n(\nu)$ and $\bar{\beta}^{(n)} = \beta^* + \bar{v}^{(n)} / \sqrt{n}$ or $\bar{\nu}^{(n)} = \sqrt{n}(\bar{\beta}^{(n)} - \beta^*)$. Then, $V^{(n)}(\nu) = \Psi_n(\nu) - \Psi_n(0)$ where

$$
V^{(n)}(\nu) = \frac{1}{n} \nu^T \Sigma_y^{-1} \nu - 2 \left( \beta^* \right)^T \Sigma_y^{-1} \nu + \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^{p} \hat{w}_j \sqrt{n} \left( \left| \beta^*_j + \nu_j \right| - \left| \beta^*_j \right| \right).
$$

(C.1)

By assumption in the first term $\Sigma_y^{-1} / n \to C$ and in the second term $n^{-1/2} \Sigma_y^{-1}(\bar{\beta} - \beta^*) = (\Sigma_y^{-1} / n) \sqrt{n}(\bar{\beta} - \beta^*) \xrightarrow{d} J$, where $J \sim N(0, C)$.

For the third term, $\delta_j = (\lambda_n / \sqrt{n}) \hat{w}_j \sqrt{n} \left( \left| \beta^*_j \right| + v_j / \sqrt{n} - \left| \beta^*_j \right| \right)$, we consider two cases.

**Case 1.** $\beta^*_j \neq 0$ or $\gamma^*_j \neq 0$: Then $\hat{w}_j \xrightarrow{p} (\left| \beta^*_j \right| + \left| \gamma^*_j \right|)^{-2}$. If $\beta^*_j \neq 0$ then $\sqrt{n}(\left| \beta^*_j \right| + v_j / \sqrt{n} - \left| \beta^*_j \right|) \to v_j \text{sign}(\beta^*_j)$ and if $\beta^*_j = 0$ then $\sqrt{n}(\left| \beta^*_j \right| + v_j / \sqrt{n} - \left| \beta^*_j \right|) = |v_j|$. By Slutsky’s theorem $\delta_j = (\lambda_n / \sqrt{n}) \hat{w}_j \sqrt{n} \left( \left| \beta^*_j \right| + v_j / \sqrt{n} - \left| \beta^*_j \right| \right) \xrightarrow{p} 0$.

**Case 2.** $\beta^*_j = 0$ and $\gamma^*_j = 0$: Then $\sqrt{n}(\left| \beta^*_j \right| + v_j / \sqrt{n} - \left| \beta^*_j \right|) = v_j$ and $(\lambda_n / \sqrt{n}) \hat{w}_j = (\lambda_n / \sqrt{n}) \{ \sqrt{n}(\left| \beta^*_j \right| + |\gamma^*_j|) \}^{-2}$ where $\sqrt{n}(\left| \beta^*_j \right| + |\gamma^*_j|) = O_p(1)$ and $\lambda_n \sqrt{n} \to \infty$. 

105
Figure C.1: Comparison of model selection methods for simulation design 1. Subfigure C.1a shows the AUC while the remaining subfigures compare model size and performance with different selection criteria. The criteria are forward selection with $\alpha_f^s = 0.25$ (solid; ——), forward selection with $\alpha_f^s = 0.15$ (short dash gray; - - - -), $C_p$ (dotted; ·····), CV (short dash blue; - - - -), AIC (dash-dot-dash; -----), and BIC (long dash; — — — —). Model size is the number of confounders selected and does not count the outcome of interest.
Thus $\delta_j = 0$ if $v_j = 0$ and $\delta_j \to \infty$ otherwise. For $\beta_0$ and $\beta_x$ the third term is 0, $\delta_0 = \delta_x = 0$.

By Slutsky’s theorem $V^{(n)}(v) \xrightarrow{d} V(v)$ for every $v$ where

$$V(v) = \begin{cases} v_A^T C_{11} v_A - 2v_A^T J_A & \text{if } v_j = 0 \forall j \notin A \\ \infty & \text{otherwise.} \end{cases}$$

The unique minimizer of $V(v)$ is $(C_{11}^{-1} J_A, 0)^T$. From the epi-convergence results of Geyer (1994) and Knight and Fu (2000),

$$\hat{v}^{(n)}_A \xrightarrow{d} C_{11}^{-1} J_A \quad \text{and} \quad \hat{v}^{(n)}_{A^c} \xrightarrow{d} 0. \tag{C.2}$$

Therefore $\hat{v}^{(n)}_A \xrightarrow{d} N(0, C_{11})$ and $\sqrt{n}(\beta_A - \beta_A^*) \xrightarrow{d} N(0, C_{11})$.

We now proceed with the proof of selection consistency. From asymptotic normality we get that $\hat{\beta}_j \xrightarrow{p} \beta_j^* \forall j \in A$; thus $\Pr(j \in A_n) \to 1$. For each $j \notin A$ we will show $\Pr(j \in A_n) \to 0$.

Assume $j \in A_n$. By the Karush-Kuhn-Tucker (KKT) optimality conditions, $2a_j^T \Sigma^{-1}(\hat{\beta} - \hat{\beta}^{(n)}) = \lambda_n \hat{w}_j$, where $a_j$ is a $p+2$-vector of zeros with a one in the corresponding location for covariate $U_j$. For $j \notin A$, $\lambda_n \hat{w}_j / \sqrt{n} = (\lambda_n \sqrt{n})(\sqrt{n}|\hat{\beta}_j| + \sqrt{n}|\hat{\gamma}_j|)^{-2} \xrightarrow{p} \infty$ and

$$\frac{a_j^T \Sigma^{-1}(\hat{\beta} - \hat{\beta}^{(n)})}{\sqrt{n}} = a_j^T \left( \frac{1}{n} \Sigma^{-1} \right) \sqrt{n}(\beta^* - \hat{\beta}^{(n)}) + a_j^T \left( \frac{1}{n} \Sigma^{-1} \right) \sqrt{n}(\hat{\beta} - \beta^*).$$

From Slutsky’s theorem and (C.2) we get that the first term on the right hand side converges in distribution to some normal distribution. By assumption, the second term on the right hand side converges in distribution to $N(0, C_{jj})$. Thus $\Pr(j \in A_n) \leq \Pr\{2a_j^T \Sigma^{-1}_y(\hat{\beta} - \hat{\beta}^{(n)}) = \lambda_n \hat{w}_j\} \to 0$.

### C.3 Covariates included in data analysis

Table C.1 shows the additional indicator variables included in the data analysis.
Table C.1: Additional covariates included in the data analysis

<table>
<thead>
<tr>
<th>Indicators related to pregnancy and birth</th>
<th>Indicators for family medical history</th>
</tr>
</thead>
<tbody>
<tr>
<td>mother smoked during pregnancy</td>
<td>anemia</td>
</tr>
<tr>
<td>mother drank during pregnancy</td>
<td>cardiac disease</td>
</tr>
<tr>
<td>mother married</td>
<td>acute or chronic lung disease</td>
</tr>
<tr>
<td>previous pregnancy</td>
<td>diabetes</td>
</tr>
<tr>
<td>previous termination</td>
<td>genital herpes</td>
</tr>
<tr>
<td>receiving prenatal care</td>
<td>hydramnios/oligohydramnios</td>
</tr>
<tr>
<td>gender of child</td>
<td>chronic hypertension</td>
</tr>
<tr>
<td></td>
<td>eclampsia</td>
</tr>
<tr>
<td></td>
<td>incompetent cervix</td>
</tr>
<tr>
<td></td>
<td>previous infant 4000 grams or more</td>
</tr>
<tr>
<td></td>
<td>previous preterm birth or small infant</td>
</tr>
<tr>
<td></td>
<td>renal disease</td>
</tr>
<tr>
<td></td>
<td>Rh sensitization</td>
</tr>
<tr>
<td></td>
<td>uterine bleeding</td>
</tr>
<tr>
<td></td>
<td>other medical conditions</td>
</tr>
</tbody>
</table>