Measurement error in environmental exposures: Statistical implications for spatial air pollution models and gene environment interaction tests

Measurement error in environmental exposures: Statistical implications for spatial air pollution models and gene environment interaction tests

A dissertation presented

by

Stacey Elizabeth Alexeeff

to

The Department of Biostatistics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biostatistics

Harvard University

Cambridge, Massachusetts

June 2013

Dissertation Advisor: Professor Brent Coull Stacey Elizabeth Alexeeff

Measurement error in environmental exposures: Statistical implications for spatial air pollution models and gene environment interaction tests

Abstract

Measurement error is an important issue in studies of environmental epidemiology. We considered the effects of measurement error in environmental covariates in several important settings affecting current public health research. Throughout this dissertation, we investigate the impacts of measurement error and consider statistical methodology to fix that error.

In Chapter 1, we investigate the effects of measurement error in a linear health effects model with a gene-environment interaction term. We examine these effects under gene-environment dependence. We derive closed-form solutions for the bias in naive parameter estimates, and we find that the resulting bias may be toward or away from the null. We also identify specific cases when the bias will be attenuated and when tests will preserve the Type I error rate.

In Chapters 2 and 3, we consider the problem of measurement error in studies of air pollution health effects, considering the case when air pollution exposure is predicted by kriging or land use regression. Chapter 2 approaches this problem from a more theoretical standpoint, and develops the spatial SIMEX methodology to correct for spatially-correlated classical measurement error. Chapter 3 complements the theoretical work in Chapter 2 in a practical assessment of the effects of measurement error on actual air pollution surfaces. This question is addressed by a simulation study using high-resolution satellite data.

Contents

Acknowledgements

I have been so fortunate to have so many people supporting me and encouraging me throughout graduate school.

I would like to thank to my committee members, Brent Coull, Xihong Lin, and Joel Schwartz. Each of them has shared a unique perspective of the pursuit of scientific knowledge which has shaped my research perspective moving forward. I am so grateful to Joel Schwartz for introducing me to the field of environmental epidemiology and sharing his enthusiasm with me. Xihong Lin has been an inspiration in her ability to provide direction and insight into any statistical question. Brent Coull has been the most supportive advisor I could imagine, helping me every step of the way throughout my time here at Harvard.

Thank you to my graduate school classmates, both here at Harvard and at Carnegie Mellon. At Harvard, I am especially thankful to Jennifer Sinnott, Mark Meyer, Lauren Kunz, and Natalie Exner. I also want to thank my graduate school classmates at Carnegie Mellon, especially Sonia Todorova, Chris Neff, James Sharpnack, and Darren Homrighausen.

I would also like to acknowledge my early mentors in statistics, in particular Javier Rojo who first introduced me to statistics research at the Rice University Summer Institute in Statistics (RUSIS) program.

Thank you to my parents Jean Ackerman and George Alexeeff, and to my sister Kathy. My family has encouraged me in all my educational endeavors throughout my life, they have inspired me to work hard, and they have set the bar high. It is hard to imagine any of my achievements without them.

Finally, thank you to my loving husband, Joe, who always believes in me.

List of Figures

List of Tables

Chapter 1

Measurement error in tests for gene-environment interactions: Implications of gene-environment dependence.

Stacey E. Alexeeff[1] and Xihong Lin[1]

[1] Department of Biostatistics, Harvard School of Public Health

1.1 Introduction

Many complex disease processes are thought to be influenced by a number of genetic and environmental factors. Modern genetics studies seek to identify a set of genetic and environmental risk factors that can explain a meaningful proportion of disease risk. Examples of genetic data of scientific interest include single nucleotide polymorphisms (SNPs), gene expression data, proteomics data, and epigenetic measures such as DNA methylation data. Genetic variants that have been identified and validated to explain complex diseases only explain small proportions of the estimated heritability of these diseases; one hypothesis is that gene-environment interactions could help explain the missing heritability of complex diseases.(Manolio et al., 2009)

Studies of gene-environment interaction studies can have several purposes: (i) identify novel genes which act through interactions rather than marginal effects and help explain the "missing heritability", (ii) identify potential causal mechanisms of how the environmental exposure may affect risk and (iii) identify genetically susceptible or resistant subpopulations.(Thomas, 2010) Identifying these more susceptible subpopulations allows us to better stratify estimates of disease risk, and ultimately inform guidelines on how much reduction in environmental exposure could reduce disease risk to protect the most susceptible populations.

An important issue gaining attention in the study of gene-environment interactions is the assumption of gene-environment (G-E) independence. Although the assumption of G-E independence is often reasonable for exogenous exposures, other environmental exposures related to behavioral or anthropometric characteristics may be dependent on genetic traits.(Thomas, 2010)

Previous research on the detection of gene-environment interactions in the presence of mismeasured exposures assumes G-E independence in the underlying population.(Garcia-Closas et al., 1998, 1999; Wong et al., 2004; Zhang et al., 2008) These studies have demonstrated that under G-E independence, non-differential measurement error in binary or continuous environmental exposures leads to the typical attenuation bias and loss of power when testing for the presence of an interaction. However, the allowing for potential correlation between G and E may induce a different more complicated structure of bias than what is observed under G-E independence.

Only one recent study has considered the role of G-E dependence in tests for gene-environment interactions in the presence of mismeasured exposures.(Lindstrom et al., 2009). That study considers testing for gene-environment interactions only in the setting of a logistic regression model when the exposure and gene are both binary. Thus, the issue of how G-E dependence impacts linear models with environmental health effects has not yet been investigated.

In this paper, we investigate the effects of measurement error in tests for gene-environment interactions under gene-environment dependence. We first present a general bias analysis by deriving closed-form solutions for the naive gene-environment interaction model parameters in Section 1.2. We show that the general form of the bias is not attenuation toward the null, rather the bias could be in either direction. In Section 1.3 we study the inference for the naive test by deriving the general form of the variance of the naive MLEs. We find that the Type I error rates are only preserved in certain special cases. In Section 1.4 we study several special cases assuming different models for G-E dependence to illustrate the effects of measurement error on bias and inference. In Section 1.5, we consider two functional-type measurement error correction strategies, regression calibration and SIMEX. We present a simulation study in Section 1.6. We then apply this to a real dataset in Section 1.7. We end with a concluding discussion.

1.2 Bias Analysis for Naive GxE Interaction Model

First, we introduce our model setup for the true model and the naive model when the surrogate of the environmental exposure that is measured with error is used ignoring possible measurement error. Then we derive the naive bias in the parameter estimate for the effect of the GxE interaction.

1.2.1 Model Setup

True Model. For subjects $i = 1, \ldots, n$, let Y_i be a continuous health outcome of interest, let E_i be an environmental exposure, and let G_i be a genetic covariate. We assume a linear health effects model with an interaction between E_i and G_i,

$$Y_i = \alpha_0 + \alpha_G G_i + \alpha_E E_i + \beta E_i G_i + \epsilon_i \tag{1.1}$$

with errors ϵ_i assumed to be independent $N(0, \sigma^2)$.

Then assume that we observe a surrogate of the environmental exposure measured with classical error, $\widetilde{E}_i = E_i + U_i$, where $U_i \overset{iid}{\sim} N(0, \sigma_U^2)$ and $E_i \overset{iid}{\sim} N(0, \sigma_E^2)$. The parameter of interest is β, the parameter relating the interaction effect to the health outcome.

Naive Model. The naive model is the model where the surrogate exposure $\widetilde{E}_i$ is used in place of $\widetilde{E}_i$, and we denote these parameters by superscript N,

$$Y_i = \alpha_{0,Naive} + \alpha_{G,Naive} G_i + \alpha_{E,Naive} \widetilde{E}_i + \beta_{Naive} \widetilde{E}_i G_i + \epsilon_{Naive,i} \tag{1.2}$$

with errors $\epsilon_{N,i}$ assumed to be independent $N(0, \sigma^2_{Naive})$. Denote the vector of parameters for the mean as $\boldsymbol{\theta}_{Naive} = (\alpha_{0,Naive}, \alpha_{G,Naive}, \alpha_{E,Naive}, \beta_{Naive})$. We are interested in how the naive parameters $\boldsymbol{\theta}_{Naive}$, σ_{Naive} are related to the true parameters $\boldsymbol{\theta}$, σ. We are particularly interested in how β_{Naive} is related to β.

In both models, we assume that the covariates are centered to have mean zero.

1.2.2 General form of bias in the naive model

For the naive linear regression model, the maximum likelihood estimates (MLEs) of the model parameters $\boldsymbol{\theta}_{Naive}, \sigma^2_{Naive}$ are the solutions to the score equations $\frac{\partial}{\partial \boldsymbol{\theta}_{Naive}} \log \mathcal{L}_n(\boldsymbol{\theta}_{Naive}, \sigma^2_{Naive}) = \mathbf{0}$ and $\frac{\partial}{\partial \sigma^2_{Naive}} \log \mathcal{L}_n(\boldsymbol{\theta}_{Naive}, \sigma^2_{Naive}) = 0$. Therefore, the probability limits of the naive model MLE's are the solutions to the expected score equations $\mathbb{E}\left\{\frac{\partial}{\partial \boldsymbol{\theta}_{Naive}} \log \mathcal{L}_1(\boldsymbol{\theta}_{Naive}, \sigma^2_{Naive})\right\} = \mathbf{0}$ and

$\mathbb{E}\left\{\frac{\partial}{\partial \sigma^2_{Naive}} \log \mathcal{L}_1(\boldsymbol{\theta}_{Naive}, \sigma^2_{Naive})\right\} = 0$.

Let the subscript 1 denote an arbitrary iid observations from our data. Let $\mathcal{X}_1 = (1, G_1, E_1, E_1 G_1)$, let $\mathcal{W}_1 = (1, G_1, \widetilde{E}_1, \widetilde{E}_1 G_1)$, let $\mathcal{U}_1 = (0, 0, U_1, U_1 G_1)$ so that $\mathcal{W}_1 = \mathcal{X}_1 + \mathcal{U}_1$. Let $\mathcal{X}_1$ and $\mathcal{W}_1$ be centered to have mean zero. Then the solutions to the expected score equations for $\boldsymbol{\theta}_{Naive}$ satisfy

$$\mathbb{E}\{\mathcal{W}_1(Y_1 - \mathcal{W}_1^{\mathsf{T}} \boldsymbol{\theta}_{Naive})\} = \mathbf{0} \tag{1.3}$$

and the solutions to the expected score equations for σ_{Naive} satisfy

$$-\frac{1}{\sigma_{Naive}} + \frac{1}{\sigma^3_{Naive}} \mathbb{E}\left[Y_1 - \mathcal{W}_1^{\mathsf{T}} \boldsymbol{\theta}_{Naive}\right]^2 = 0 \tag{1.4}$$

Starting from equation (1.3), using iterated expectation conditional on (E_1, G_1) and taking expectation under the true model, calculations in the appendix show that

$$\boldsymbol{\theta}_{Naive} = \boldsymbol{\Lambda} \boldsymbol{\theta} \tag{1.5}$$

where

$$\boldsymbol{\Lambda} = \left[Cov(\mathcal{W}_1)\right]^{-1} Cov(\mathcal{X}_1) \tag{1.6}$$

Thus, we have solutions for $\boldsymbol{\theta}_{Naive}$ in terms of the true parameters $\boldsymbol{\theta}$, which depend on these particular covariance matrices. This solution relies on our assumption that all covariates are centered so that they have mean 0. Solutions for the case when covariates are not centered are given in the Appendix Section. To solve equation (1.4) for σ^2_{Naive}, we substitute our solutions for $\boldsymbol{\theta}_{Naive}$ and use properties of vectors and quadratic forms. Calculations in the appendix show that

$$\sigma^2_{Naive} = \sigma^2 + \boldsymbol{\theta}^\mathsf{T} \left[\left(\mathbf{I} - \boldsymbol{\Lambda}^\mathsf{T} \right) Cov(\mathcal{X}_1) \right] \boldsymbol{\theta} \tag{1.7}$$

From equation (1.7), it is clear that σ^2_{Naive} will always be inflated compared to the true σ^2, and the degree of inflation will depend on the particular matrix $\boldsymbol{\Lambda}$ relating $\boldsymbol{\theta}_{Naive}$ to $\boldsymbol{\theta}$.

We next consider how the covariance matrices $Cov(\mathcal{W}_1)$ and $Cov(\mathcal{X}_1)$ affect the relationship between the naive model coefficients and the true model coefficients.

1.2.3 Bias under dependence between E and G

Let E, G be correlated. Let the correlations be parameterized as $\rho_1 \equiv Corr(E_1, G_1)$, $\rho_2 \equiv Corr(G_1, E_1G_1)$, $\rho_3 \equiv Corr(E_1, E_1G_1)$.

Under this general parameterization, not yet assuming any particular model for $E|G$, Calculations in the Appendix show that the solutions to equation (1.5) can be written as

$$\boxed{\begin{aligned} \alpha_{0,Naive} &= \alpha_0 \\ \alpha_{G,Naive} &= \alpha_G + \lambda_5 \alpha_E + \lambda_6 \beta \\ \alpha_{E,Naive} &= \lambda_3 \alpha_E + \lambda_4 \beta \\ \beta_{Naive} &= \lambda_1 \alpha_E + \lambda_2 \beta \end{aligned}}$$

where

$$
\begin{aligned}
\lambda_1 &= \frac{1}{\det[Cov(\mathcal{W}_1)]} \cdot (\rho_3 - \rho_1\rho_2)\,\sigma_E\sigma_{EG}\sigma_G^2\sigma_U^2 \\
\lambda_2 &= \frac{1}{\det[Cov(\mathcal{W}_1)]} \cdot \Big[\left(1 + 2\rho_1\rho_2\rho_3 - \rho_1^2 - \rho_2^2 - \rho_3^2\right)\sigma_E^2 + \left(1 - \rho_2^2\right)\sigma_U^2\Big]\sigma_G^2\sigma_{EG}^2 \\
\lambda_3 &= \frac{1}{\det[Cov(\mathcal{W}_1)]} \cdot \Big[\left(1 + 2\rho_1\rho_2\rho_3 - \rho_1^2 - \rho_2^2 - \rho_3^2\right)\sigma_{EG}^2 + \left(1 - \rho_1^2\right)\sigma_G^2\sigma_U^2\Big]\sigma_E^2\sigma_G^2 \\
\lambda_4 &= \frac{1}{\det[Cov(\mathcal{W}_1)]} \cdot (\rho_3 - \rho_1\rho_2)\,\sigma_E\sigma_{EG}\sigma_G^4\sigma_U^2 \\
\lambda_5 &= \frac{1}{\det[Cov(\mathcal{W}_1)]} \cdot \Big[(\rho_1 - \rho_2\rho_3)\sigma_{EG}^2 + \rho_1\sigma_G^2\sigma_U^2\Big]\sigma_E\sigma_G\sigma_U^2 \\
\lambda_6 &= \frac{1}{\det[Cov(\mathcal{W}_1)]} \cdot \Big[(\rho_2 - \rho_1\rho_3)\sigma_E^2 + \rho_2\sigma_U^2\Big]\sigma_{EG}\sigma_G^3\sigma_U^2
\end{aligned}
$$

and

$$
\det[Cov(\mathcal{W}_1)] = \Big[\sigma_G^2\sigma_U^2 + \left(1 - \rho_2^2\right)\sigma_{EG}^2 + \left(1 - \rho_1^2\right)\sigma_E^2\sigma_G^2\Big]\sigma_G^2\sigma_U^2 + \left(1 + 2\rho_1\rho_2\rho_3 - \rho_1^2 - \rho_2^2 - \rho_3^2\right)\sigma_E^2\sigma_G^2\sigma_{EG}^2.
$$

Result 1. In general under G-E dependence, the direction of bias of β_{Naive} could be toward or away from the null, and the bias depends on the magnitude and direction of the marginal effect of the environmental exposure.

Next, we consider the interpretation of the λ_j's. Since all variances are assumed to be positive, then $\det[Cov(\mathcal{W}_1)] > 0$, and hence $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6 \in (-1, 1)$ where they are well-defined. Based on the expressions for $\alpha_{G,Naive}$, $\alpha_{E,Naive}$, β_{Naive}, the naive coefficient for the marginal genetic effect has an additive bias factor that depends on the magnitudes and directions of both the marginal effect of the environmental exposure and the interaction effect. The naive coefficient for the marginal environmental exposure effect is attenuated toward the null as long as $\lambda_3 > 0$. We are most interested in the direction of bias for β_{Naive}.

Further interpretation requires considering particular distributions for G, E, and the dependence relationship $E|G$. Those distributional assumptions will determine the moments $\mathbb{E}[EG]$, $\mathbb{E}[EG^2]$, $\mathbb{E}[E^2G]$ and the relationships between ρ_1, ρ_2, ρ_3. In Section 1.4 we consider a number of specific models and examine the λ_j's under those models.

1.3 Inference in naive GxE models

Next, we turn our attention to inferences made using the naive model. The first question of interest is whether the naive test preserves the Type I error rate under the null hypothesis.

Under the true model, the test statistic for the t-test of $H_0 : \beta = 0$ using the MLE $\widehat{\beta}$ is given by $\widehat{\beta}/s.e.(\widehat{\beta})$, where $Var(\hat{\boldsymbol{\theta}}|\mathbf{X}) = \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$. The naive t-test for H_0 is constructed using the test statistic $\widehat{\beta}_{Naive}/s.e._{Naive}$ based on the naive variance $\sigma^2_{Naive}(\mathbf{W}^\mathsf{T}\mathbf{W})^{-1}$. However, it is not necessarily true that this naive variance $\sigma^2_{Naive}(\mathbf{W}^\mathsf{T}\mathbf{W})^{-1}$ is equal to $Var(\hat{\boldsymbol{\theta}}_{Naive}|\mathbf{W})$, the true variability of the naive MLEs given the data with measurement error.

1.3.1 Type I error rate

For the size of the naive test to be correct, we require that

(i) $\beta_{Naive} = 0$ under the null, and

(ii) the naive variance, $\sigma^2_{Naive}(\mathbf{W}^\mathsf{T}\mathbf{W})^{-1}$, is equal to the true true variability of the naive MLEs given the data with measurement error, $Var(\hat{\boldsymbol{\theta}}_{Naive}|\mathbf{W})$.

Based on our investigation of the bias in Section 1.2, we know that condition (i) will be violated unless $\lambda_1 = 0$ or $\alpha_E = 0$. We now derive $Var(\hat{\boldsymbol{\theta}}_{Naive}|\mathbf{W})$ to determine whether condition (ii) is violated.

1.3.2 General Form of Variance of MLEs $\hat{\boldsymbol{\theta}}_{Naive}$

Let $\hat{\boldsymbol{\theta}}_{Naive}$ be the MLEs for the naive parameters $\boldsymbol{\theta}_{Naive}$. Then,

$$Var\left\{\hat{\boldsymbol{\theta}}_{Naive}\middle|\mathbf{W}\right\} = (\mathbf{W}^\mathsf{T}\mathbf{W})^{-1}\mathbf{W}^\mathsf{T}Var\left\{\mathbf{Y}\middle|\mathbf{W}\right\}\mathbf{W}(\mathbf{W}^\mathsf{T}\mathbf{W})^{-1} \tag{1.8}$$

Now, each observation $i = 1, \ldots, n$ is i.i.d. so $Var\left\{\mathbf{Y}\middle|\mathbf{W}\right\}$ is a diagonal matrix. The diagonal entries correspond to the conditional variance for each observation, $Var\left\{Y_i|\mathcal{W}_i\right\}$. To simplify this expression we would need constant variance $Var\left\{Y_i|\mathcal{W}_i\right\} = \sigma_*^2$ for all $i = 1, \ldots, n$, for some σ_*^2. Under that condition, equation (1.8) would simplify to $Var\left\{\hat{\boldsymbol{\theta}}_{Naive}\middle|\mathbf{W}\right\} = \sigma_*^2(\mathbf{W}^\mathsf{T}\mathbf{W})^{-1}$. In order to meet the condition that the estimated model variance in the naive model, $\sigma^2_{Naive}(\mathbf{W}^\mathsf{T}\mathbf{W})^{-1}$, is equal to the true $Var\left\{\hat{\boldsymbol{\theta}}_{Naive}|\mathbf{W}\right\}$, that constant σ_*^2 would have to be equal to the naive model variance parameter σ^2_{Naive}.

Calculations in the Appendix show that by using iterated expectation, the conditional variance for any observation i can be expressed as

$$Var\left\{Y_i|\mathcal{W}_i\right\} = \sigma^2 + \left(\alpha_E^2 + \beta^2 G_i^2 + 2\alpha_E\beta G_i\right)Var\left\{E_i\middle|(G_i, \widetilde{E}_i)\right\} \tag{1.9}$$

Under H_0, equation (1.9) simplifies to

$$Var\left\{Y_i|\mathcal{W}_i\right\} = \sigma^2 + \left(\alpha_E^2\right)Var\left\{E_i\middle|(G_i, \widetilde{E}_i)\right\} \tag{1.10}$$

In general, this variance will not be constant for all $i = 1, \ldots, n$ because the conditional variance could depend on G_i. When the variance is not constant, this equation will not simplify and will not be equal to $\sigma^2_{Naive}(\mathbf{W}^\mathsf{T}\mathbf{W})^{-1}$. Thus, in general the Type I error rates will not be preserved.

Result 2. In general under G-E dependence, the naive variance estimator will not reflect the true variability of the naive MLE's. The Type I error of the naive test will be inflated, leading to an increased rate of spurious associations.

There may be some special cases where the inferences will be valid under the null. One such case is the independent case. We will explore potential special cases further in the next section.

1.4 Bias and Inference Derivations for Specific Models

We now consider four special cases of interest for G-E dependence to illustrate the impacts of this measurement error. We derive particular expressions for the biases and illustrate the parameters of the bias expressions in Figure 1.1.

Scenario (a) Independence between E and G

When E and G are independent, equations (1.5) and (1.6) greatly simplify because the covariance matrices $Cov(\mathcal{W}_1)$ and $Cov(\mathcal{X}_1)$ are both diagonal.

Bias. Defining $\lambda' = \sigma_E^2/(\sigma_E^2 + \sigma_U^2)$ and with some algebra, then $\lambda_1 = \lambda_4 = \lambda_5 = 0$ and $\lambda_2 = \lambda_3 = \lambda'$. Then equation (1.5) can be written as

$$\begin{aligned} \alpha_{0,Naive} &= \alpha_0 \\ \alpha_{X,Naive} &= \alpha_X \\ \alpha_{G,Naive} &= \alpha_G \\ \alpha_{E,Naive} &= \lambda'\alpha_E \\ \beta_{Naive} &= \lambda'\beta \end{aligned}$$

The naive coefficients $\alpha_{E,Naive}$ and β_{Naive} are attenuated toward the null by a factor of λ'. We note that $0 < \lambda' < 1$, and λ' corresponds to the usual attenuation factor found in linear models with classical measurement error.

Inference. Calculations in the Appendix show that under the null,

$$Var(\hat{\boldsymbol{\theta}}_{Naive}|\mathbf{W}) = \sigma^2_{Naive}(\mathbf{W}^\top\mathbf{W})^{-1} = \sigma^2 + \lambda'\sigma_U^2\alpha_E^2$$

Hence, for this scenario the conditions needed to preserve Type I error rates are met, so the naive test is a valid test. In our simulation study in Section 1.6, we show that the Type I error rates are preserved.

Scenario (b) Linear dependence between G and E.

Suppose that the dependence between E and G is linear, satisfying $E_i = \gamma_1 G_i + \delta_i$ where $\delta_i \sim Normal(0, \sigma_\delta^2)$ for $i = 1, \ldots, n$. We do not assume any particular distribution for G_i.

Bias. We observed previously that $\beta_{Naive} = \lambda_2 \beta$ only when $\lambda_1 = 0$ or $\alpha_E = 0$. The condition for $\lambda_1 = 0$ is that $(\rho_1\rho_2 - \rho_3) = 0$. Calculations in the Appendix show that under this linear dependence model, $\rho_1\rho_2 = \rho_3$ and thus $\beta_{Naive} = \lambda_2\beta$. These calculations use the linear relationship for $E|G$, but do not make any assumptions about the particular distribution of G. Hence, this result holds for any distribution of G, including binary, SNP coding, or continuous distributions.

Figure 1.1 shows the λ_j's as a function of $Corr(E, G)$ under G binary, Normal, and SNP. We can see that the λ_j's vary by the distribution of G. We can also see the independent case represented at the point of Corr(E,G)=0, so this plot shows how independent and dependent cases relate. When G is binary, $\lambda_2 \leq \lambda$, where the degree of attenuation in the independent case is the minimum attenuation in the dependent case, and the attenuation becomes more severe as the dependence between G and E increases.

Inference. Next we consider inference in the linear G-E dependence model. We first show that Type I error rates are preserved under this dependence model.

Result 3. In the special case when $E|G$ depends linearly on G and $Var[E|G]$ does not depend on G, then the resulting direction of bias of β_{Naive} is attenuation toward the null and Type I error rates are preserved.

Under linear dependence between E and G, we can use the particular model for $E|G$, and the properties of multivariate Normality to derive the particular form of the expression for the variance of the naive MLE's. Calculations in the Appendix show that in this special

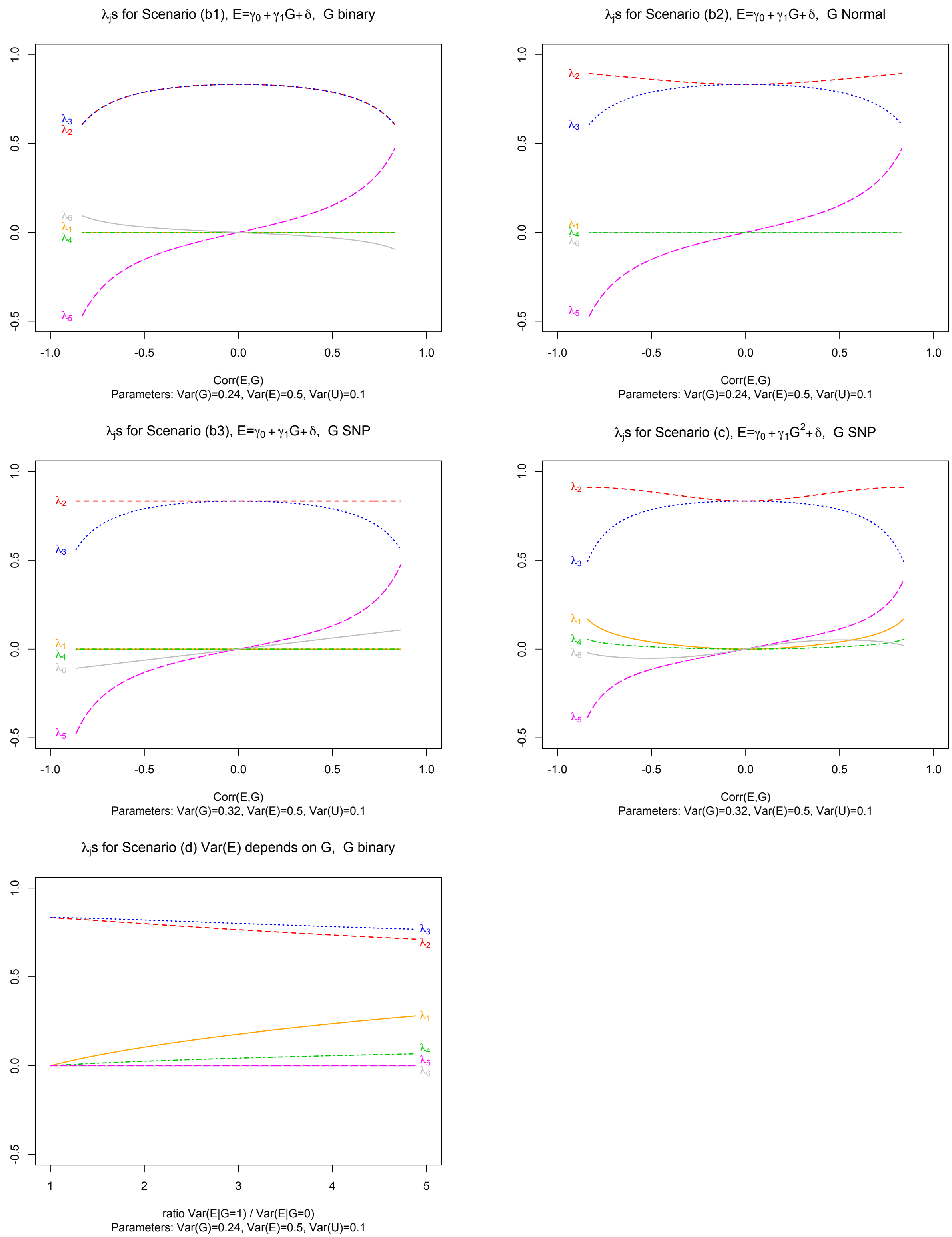

Figure 1.1: Lambdas for each scenario of interest, showing varying degrees of dependence between E and G when variances of G, E, and U are fixed such that E has twice the variance of G and the measurement error variance is fixed to be 20% of the variance of E.

case, equation (1.8) simplifies to

$$Var\left\{\hat{\boldsymbol{\theta}}_{Naive}\middle|\mathbf{W}\right\} = \left(\sigma^2 + \alpha_E^2 \cdot \frac{\sigma_\delta^2\sigma_U^2}{\sigma_\delta^2 + \sigma_U^2}\right)(\mathbf{W}^\mathsf{T}\mathbf{W})^{-1} \tag{1.11}$$

Next, we derive a simplified expression for equation (1.7), the naive model variance, for this scenario to demonstrate that these variances are equal. Under H_0 and linear dependence between E and G, we can simplify this expression, using the properties that $\beta = 0$, $\lambda_1 = 0$, $\lambda_4 = 0$. Then equation (1.7) can be expressed as

$$\sigma_{Naive}^2 = \sigma^2 + \alpha_E^2\Big[-\lambda_5\rho_1\sigma_G\sigma_E + (1-\lambda_3)\sigma_E^2\Big] + \alpha_G\alpha_E\Big[-\lambda_5\sigma_G^2 + (1-\lambda_3)\rho_1\sigma_G\sigma_E\Big] \tag{1.12}$$

Calculations in the Appendix show that the second term can be simplified and the last term is equal to zero. Hence, using these simplified expressions, equation (1.7) simplifies to

$$\sigma_{Naive}^2 = \left(\sigma^2 + \alpha_E^2 \cdot \frac{\sigma_\delta^2\sigma_U^2}{\sigma_\delta^2 + \sigma_U^2}\right) \tag{1.13}$$

Thus, we have our result

$$\sigma_{Naive}^2(\mathbf{W}^\mathsf{T}\mathbf{W})^{-1} = \left(\sigma^2 + \alpha_E^2 \cdot \frac{\sigma_\delta^2\sigma_U^2}{\sigma_\delta^2 + \sigma_U^2}\right)(\mathbf{W}^\mathsf{T}\mathbf{W})^{-1} \tag{1.14}$$

Hence, we have shown that for the case of linear dependence between E and G, we have

$$Var(\hat{\boldsymbol{\theta}}_{Naive}|\mathbf{W}) = \sigma_{Naive}^2(\mathbf{W}^\mathsf{T}\mathbf{W})^{-1}.$$

Scenario (c) G binary, variance of E depends on G.

Suppose that G is binary and that the dependence between E and G is structured so that the variance of E depends on G rather than the mean of E. Specifically, let $E_i|G_i = 0 \sim N\left(0, \sigma_\delta^2\right)$ and $E_i|G_i = 1 \sim N\left(0, \sigma_\delta^2 + \sigma_\tau^2\right)$ for $i = 1, \ldots, n$.

Bias. Under this linear dependence model, $\rho_1\rho_2 \neq \rho_3$ and thus $\lambda_1 \neq 0$ and $\beta_{Naive} = \lambda_1\alpha_E + \lambda_2\beta$ cannot be simplified. Thus, when α_E and β are in the same direction and when α_E is sufficiently large, the direction of bias will be away from the null. Figure 1 shows the λ_j's as a function of the ratio $(\sigma_\delta^2 + \sigma_\tau^2)/\sigma_\delta^2$.

Inference. Based on the results for bias, we know that condition (i) for preserving the Type I error rate is violated.

Scenario (d) G SNP coding, mean of E depends on G^2.

Suppose that G has SNP coding $\{0, 1, 2\}$ and that $E|G$ depends on G^2. Specifically, let $E_i = \gamma_1 G_i^2 + \delta_i$ where $\delta_i \sim Normal(0, \sigma_\delta^2)$ for $i = 1, \ldots, n$.

Bias. Under this linear dependence model, $\rho_1\rho_2 \neq \rho_3$ and thus $\beta_{Naive} = \lambda_1\alpha_E + \lambda_2\beta$ cannot be simplified. Thus, when α_E and β are in the same direction and when α_E is sufficiently large, the direction of bias will be away from the null. Figure 1 shows the λ_j's as a function of $Corr(E, G$.

Inference. Based on the results for bias, we know that condition (i) for preserving the Type I error rate is violated.

1.4.1 Implications of Not Centering Covariates.

A key assumption worth highlighting is the assumption that covariates either had mean zero or were centered to have mean zero, which was assumed from equation (1.6) onward. Derivations for the versions of (1.6) and the particular algebraic expressions for the nonzero terms of the matrix $\mathbf{\Lambda}$ can be found in the Appendix. Overall, the form of the expressions are similar, but the bias coefficients involve more terms. We now briefly explain the implications of not centering covariates based on these differences in the derived expressions of bias.

In the case of G-E independence, the biases of the naive coefficients $\alpha_{E,Naive}$ and β_{Naive} remain the same, but the genetic effect is now biased. Specifically, $\alpha_{G,Naive} = \alpha_G + \mu_E(1 - \lambda')\beta$, where μ_E is the mean of the environmental covariate. Thus, the direction of bias of the genetic effect depends on the signs of μ_E and β. Centering E will eliminate this bias; centering G is not required.

1.5 Corrected Test

1.5.1 Regression Calibration.

Regression calibration is a widely-used method which operated by replacing the mismeasured $\tilde{E}$ by an unbiased estimate of $\mathbb{E}[E|\tilde{E}, G]$.(Carroll et al., 2006) This expectation must correctly specify the relationship between E and G to ensure that the resulting health model coefficient estimates are unbiased. In other words, we require that the true G-E dependence relationship is known and used for the calibration step.

The Appendix outlines the derivations needed for regression calibration to work in our model setup. Regression calibration will work for Scenarios (b) and (d) where the E-G dependence relates the mean of $E|G$ and the variance does not depend on G. Regression calibration will not work for Scenario (c) because regression calibration is a correction of the mean model and in this scenario the mean has no G-E dependence.

1.5.2 SIMEX.

SIMEX is a method for measurement error correction baed on resampling. A description of SIMEX and its asymptotic properties are given in Carroll et al. (2006) and Cook and Stefanski (1994). Briefly, SIMEX has two steps. The first step is a simulation step where simulated measurement error is added to the mismeasured exposures $\tilde{E}$ in increasing amounts. The outcome model is refit for the increasing measurement error to determine the trend in the bias of the naive model parameters. The second step is the extrapolation step where a trend is fit to the distribution of parameter estimates over the increasing error levels and extrapolated back to the case of no error. A typical default for the extrapolation function is quadratic which is what we used throughout our simulations and data example.

1.6 Simulation Study

We conducted a simulation study to assess the finite-sample performance of the naive model and the regression calibration and SIMEX correction methods compared to the true model. We investigated the Type I error rates of the test for $\beta = 0$ and the bias in the parameters $\alpha_G, \alpha_E, \beta$ for each of the scenarios of G-E dependence discussed in Section 1.4.

A total of $n = 400$ iid observations were generated for each simulation. The continuous health outcome Y_i for $i = 1, \ldots, n$ was generated from the true model given in equation (1). We set the total variance of E as $\sigma_E^2 = 0.5$ and the measurement error variance as $\sigma_U = 0.1$ so that the ratio of measurement error variance to variability in the true exposure was 20%. For scenario (a), we modeled G as binary. For scenario (b), E was generated as $E_i = \gamma_0 + \gamma_1 G_i + \delta_i$, $\delta_i \sim N(0, \sigma_\delta^2)$, where we considered several distributions for G: binary, continuous and categorical (0,1,2). For scenario (c), we modeled G as binary and E was generated as $E_i|(G_i = 0) \sim N(0, \sigma_\delta^2)$ and $E_i|(G_i = 1) \sim N(0, \sigma_\delta^2 + \tau^2)$. For scenario (d), E was generated as $E_i = \gamma_0 + \gamma_1 G_i^2 + \delta_i$, $\delta_i \sim N(0, \sigma_\delta^2)$, with G categorical (0,1,2).

For the simulations of bias, we set the parameters to $\alpha_G = 1.0, \alpha_E = 3.5, \beta = 3.0$ and $\sigma^2 = 1.0$ and we ran 5,000 simulations for each setting. For the simulations of Type I error rates, we set the parameters to $\alpha_G = 1, \alpha_E = 2, \beta = 0$ and $\sigma^2 = 1$ and we ran 10,000 simulations for each setting to obtain an accurate estimate of the rate of false positives.

1.6.1 Type I Error Rates.

The results of the simulations for Type I error rates are shown in Tables 1.1-1.3. Table 1.1 shows that the Type I error rates are preserved for scenarios (a) and (b), which was shown theoretically in Section 1.4. Table 1.2 illustrates the inflated Type I error rates in scenario (c), where the degree of inflation increases with the ratio of the $E|G$ variances, illustrating the impact of the degree of G-E dependence. The SIMEX correction improves the Type I error rate, although it is still slightly inflated when the ratio of variances is large. Even in those cases the SIMEX method provides substantial improvement. Regression calibration was not attempted in scenario (c) because regression calibration is a correction of the mean model and in this scenario the mean has no G-E dependence.

Table 1.3 illustrates the inflated Type I error rate in scenario (d), and illustrates corrections by both SIMEX and regression calibration. We considered three models for regression calibration: the correct model regressing on $\tilde{E}$ and G^2, an incorrect model regressing only on $\tilde{E}$, and an incorrect model regressing on $\tilde{E}$ and G. The correct regression model successfully obtains a 0.05 Type I error rate, while incorrectly specifying the regression calibration model yields approximately the same degree of inflation in Type I error rate as the naive model. This underscores the importance of knowing the correct regression calibration model when using this approach. The SIMEX correction of scenario (d) improves the Type I error rate but it is still slightly inflated.

Table 1.1: Type I error rates for test of $\beta = 0$ in true model and naive model are preserved under Scenario (a) G-E independence and under Scenario (b) G-E dependence where the mean of $E|G$ depends linearly on G but the variance does not depend on G.

Scenario (a)					
Model	G	Theoretical value	Mean estimate β	empirical SE	size
True model	Independent	0.0	0.002	0.212	0.0474
Naive model	Independent	0.0	0.003	0.208	0.0478
Scenario (b)					
Model	G	Theoretical value	Mean estimate β	empirical SE	size
True model	Binary	0.000	0.002	0.155	0.0496
Naive model	Binary	0.000	0.004	0.161	0.0489
True model	Normal	0.000	-0.000	0.066	0.0508
Naive model	Normal	0.000	0.000	0.071	0.0492
True model	SNP	0.000	-0.001	0.099	0.0518
Naive model	SNP	-0.000	-0.000	0.103	0.0477

Table 1.2: Type I error rates for test of $\beta = 0$ in true model and naive model and correction by SIMEX under Scenario (c) G-E dependence where only the variance of $E|G$ depends on G but the mean of $E|G$ does not depend on G.

Variance Ratio	Model	Theoretical	Mean β	empirical SE	size
1.5	True model	0.0	0.000	0.154	0.0478
	Naive model	0.117	0.119	0.160	0.1124
	SIMEX		0.033	0.215	0.0504
2.0	True model	0.0	-0.000	0.164	0.0468
	Naive model	0.21	0.210	0.167	0.2326
	SIMEX		0.065	0.215	0.0666
3.0	True model	0.0	-0.000	0.184	0.0467
	Naive model	0.35	0.353	0.179	0.4943
	SIMEX		0.131	0.215	0.1067

Table 1.3: Type I error rates for test of $\beta = 0$ in true model and naive model and correction by SIMEX and Regression Calibration under Scenario (d) G-E dependence where G has SNP coding and the mean of $E|G$ depends on G^2 but the variance does not depend on G.

Model	Theoretical	Mean	model SE	empirical SE	size
True model	0	-0.001	0.096	0.097	0.0480
Naive model	0.106	0.107	0.102	0.102	0.1742
RC correct E, G^2		-0.002	0.118	0.118	0.0479
RC only E		0.128	0.122	0.123	0.1833
RC only E, G		0.139	0.120	0.121	0.2100
SIMEX		0.034	0.116	0.121	0.0572

1.6.2 Bias of naive model parameters and correction by Regression Calibration and SIMEX.

The results of the simulations for parameter bias are shown in Tables 1.4 and 1.5. An independent validation dataset with 200 observations was generated in each simulation, and the parameters for the regression calibration and the SIMEX corrections were estimated using that validation set. Table 1.4 shows the bias in the parameter estimates for scenario (b). The parameter for the main effect of the gene is biased upward, the parameter for the main effect of the environmental exposure is biased toward the null, and the parameter for the gene-environment interaction effect is also biased toward the null. The averages of the naive parameter estimates over the simulations closely match the theoretical asymptotic values. SIMEX and regression calibration using the correct model both worked well in correcting bias, while regression calibration on $\tilde{E}$ only corrected some of the bias. In particular, the upward bias in main effect of the gene remained when regression calibration on $\tilde{E}$ was performed.

Table 1.5 shows the bias in the parameter estimates for scenarios (c) and (d). Again, the averages of the naive parameter estimates over the simulations closely match the theoretical asymptotic values. In scenario (c), SIMEX works well in correcting the bias of the parameter estimates, while regression calibration was ineffective in correcting any parameter bias, as expected. In scenario (d), SIMEX and regression calibration using the correct model both worked well in correcting bias, while the misspecified regression calibration models only corrected some of the bias.

1.7 Data Example

The apolipoprotein E (APOE) gene makes the apolipoprotein E which controls lipid protein metabolism and transport. In humans, the two SNPs rs429358 and rs7412 define the APOE gene epsilon alleles which are functional polymorphisms. The most common is the APOE-ϵ3 allele, rs429358(T) + rs7412(C), found in approximately 78% of the general population.(Farrer et al., 1997) The APOE-ϵ2 allele is defined by the rs7412(T) mutation while the APOE-ϵ4 allele is defined by the rs429358(C) mutation, with frequencies of 8% and 14%, respectively in the general population.(Farrer et al., 1997)

Epidemiologic studies and meta-analyses have demonstrated that the presence of at least one APOE-ϵ4 allele is associated with a greater risk of coronary artery disease.(Wilson et al., 1996; Bennet et al., 2007) This increased risk is largely attributed to elevated cholesterol, a well-established risk factor for coronary disease. Associations have been reported between

Table 1.4: Bias of naive model parameters in linear health effect model for Scenario (b) and correction by Regression Calibration and SIMEX.

		Scenario (b1)			Scenario (b2)			Scenario (b3)		
Parameter	Method	Theoretical	Mean	SE	Theoretical	Mean	SE	Theoretical	Mean	SE
α_G	True	1.0	1.001	0.155	1.0	1.000	0.121	1.0	1.000	0.143
	Naive	1.27	1.271	0.180	1.21	1.206	0.148	1.29	1.291	0.173
	RC on E,G		0.996	0.224		0.997	0.169		0.994	0.210
	RC on E only		1.268	0.175		1.204	0.150		1.290	0.170
	SIMEX		1.029	0.188		1.025	0.155		1.028	0.183
α_E	True	3.5	3.499	0.107	3.5	3.500	0.170	3.5	3.501	0.128
	Naive	2.85	2.851	0.119	2.81	2.815	0.169	2.92	2.918	0.136
	RC on E,G		3.505	0.189		3.500	0.244		3.504	0.199
	RC on E only		3.426	0.180		3.376	0.230		3.386	0.186
	SIMEX		3.426	0.162		3.414	0.222		3.445	0.174
β	True	3.0	3.000	0.154	3.0	2.999	0.067	3.0	2.999	0.091
	Naive	2.44	2.446	0.191	2.57	2.564	0.132	2.59	2.587	0.140
	RC on E,G		3.006	0.260		2.994	0.180		3.001	0.182
	RC on E only		2.938	0.252		3.075	0.185		3.005	0.185
	SIMEX		2.936	0.250		2.969	0.173		2.972	0.176

Table 1.5: Bias of naive model parameters in linear health effect model for Scenarios (c), (d) and correction by Regression Calibration and SIMEX.

		Scenario (c)			Scenario (d)		
Parameter	Method	Theoretical	Mean	SE	Theoretical	Mean	SE
α_G	True	1.0	0.998	0.144	1.0	1.007	0.144
	Naive	1.00	0.999	0.172	1.62	1.616	0.192
	RC on E, G^2		0.997	0.208		1.007	0.221
	RC on E only		0.999	0.173		1.615	0.191
	SIMEX		0.998	0.179		1.088	0.210
α_E	True	3.5	3.501	0.104	3.5	3.501	0.134
	Naive	2.94	2.942	0.115	2.70	2.697	0.140
	RC on E, G^2		3.540	0.184		3.509	0.221
	RC on E only		3.540	0.183		3.246	0.199
	SIMEX		3.451	0.151		3.361	0.194
β	True	3.0	3.001	0.154	3.0	3.000	0.097
	Naive	2.66	2.656	0.183	2.78	2.789	0.157
	RC on E, G^2		3.189	0.247		3.003	0.212
	RC on E only		3.189	0.247		3.354	0.218
	SIMEX		2.990	0.234		3.029	0.201

APOE polymorphisms and both total plasma cholesterol and low-density lipoprotein (LDL) cholesterol, where APOE-ϵ2 is associated with lower cholesterol levels and APOE-ϵ4 is associated with higher levels of cholesterol compared to APOE-ϵ3.(Bennet et al., 2007) The APOE genotypes are reported to have an approximately linear relationship (ordered ϵ2, ϵ3, ϵ4) with LDL cholesterol.(Bennet et al., 2007)

In addition, several studies have reported a gene-environment interaction between APOE and BMI on total cholesterol and LDL cholesterol.(Boer et al., 1997; Marques-Vidal et al., 2003; Srinivasan et al., 2001; Pardo Silva et al., 2008) These studies show that those with both the ϵ2 allele and low BMI had the lowest total cholesterol and LDL cholesterol levels, while the highest total cholesterol and LDL cholesterol levels were seen in those with both the ϵ4 allele and high BMI.

Despite APOE-ϵ4 carriers having the highest levels of cholesterol and the highest risk of coronary disease, APOE-ϵ4 is also associated with a lower body mass index compared with other APOE alleles; specifically, apoE isoforms are associated with increasing body mass index in the order: APOE-$\epsilon4$ (lowest), APOE-$\epsilon3$, APOE-$\epsilon2$ (highest).(Volcik et al., 2006; Pardo Silva et al., 2008)

BMI is a widely-used surrogate for the measurement of body fat storage because of its low cost and convenience. However, BMI is considered an error-prone surrogate of body composition because it can not distinguish fat mass from lean mass.(Allison and Saunders, 2000). Several papers have examined the relationship between body fat and BMI, illustrating that for a given body fat percentage, BMI is still quite variable. (Jackson et al., 2002; Shah and Braverman, 2012)

We examined the associations between APOE, BMI, and their interactions on plasma cholesterol in a cross-sectional subset of 812 men who participated in the Normative Aging Study during 1992 to 1995. Details on the characteristics of the study participants in the Normative Aging Study have been previously published.(Kawachi et al., 1994) Three APOE genotype groups were defined: APOE-$\epsilon2$ (including $\epsilon2/\epsilon2$, $\epsilon3/\epsilon2$ genotypes), APOE-$\epsilon3$ ($\epsilon3/\epsilon3$), and APOE-$\epsilon4$ (including $\epsilon4/\epsilon3$, $\epsilon4/\epsilon4$ genotypes), and we modeled the APOE associations linearly, as in other published studies. We implemented SIMEX because of its good performance theoretically and in simulations, and because there is no need to specify the true model underlying G-E dependence. In the absence of validation data, we considered a range of reasonable measurement error variances (equal to 10%, 20%, and 30% of the total variance of BMI) to illustrate the potential sensitivity of this association to measurement error.

The median age of study subjects was 65 years, and the mean BMI of study subjects was 27.9, which is considered overweight. To assess G-E dependence in our data, we examined the linear association between BMI and APOE. Covariates were centered for the analyses, so the effects estimates represent deviations from a subject with APOE-$\epsilon3$ genotype and average BMI. The direction of a estimated association was consistent with published studies, but the association was not statistically significant. There was also no evidence of heterogeneity of variance of BMI by APOE group. The naive estimates for the effects of and the SIMEX corrections are given in Table 1.6. We see that all naive coefficients are corrected upward by SIMEX, with the most notable increases in the effect estimate of the interaction.

Table 1.6: Naive associations and confidence intervals for data example. Associations between BMI, APOE gene, and interaction on outcomes total cholesterol and LDL cholesterol.

		Total Cholesterol			LDL Cholesterol		
Parameter	Method	Mean	95% C.I.		Mean	95% C.I.	
α_G	Naive model	6.93	2.58,	11.27	7.04	2.65,	11.43
	SIMEX, 10%	6.94	5.65,	8.23	7.15	5.79,	8.51
	SIMEX, 20%	7.00	5.63,	8.36	7.19	5.75,	8.64
	SIMEX, 30%	6.96	5.50,	8.43	7.33	5.69,	8.96
α_E	Naive model	0.06	-0.66,	0.77	0.81	0.08,	1.53
	SIMEX, 10%	0.07	-1.22,	1.36	0.94	-0.42,	2.29
	SIMEX, 20%	0.13	-1.24,	1.50	0.94	-0.50,	2.39
	SIMEX, 30%	0.11	-1.35,	1.57	1.11	-0.52,	2.75
β	Naive model	1.17	0.02,	2.33	1.69	0.47,	2.91
	SIMEX, 10%	1.23	-0.06,	2.52	1.87	0.51,	3.23
	SIMEX, 20%	1.46	0.09,	2.82	2.06	0.62,	3.51
	SIMEX, 30%	1.60	0.13,	3.06	2.19	0.56,	3.83

1.8 Discussion

In this paper, we examined how tests for gene-environment interactions can be affected by measurement error in the environmental exposure and gene-environment dependence. We showed that in general under G-E dependence, the direction of bias of β_{Naive} could be toward or away from the null. We provided closed-form expressions for bias of coefficients in the linear model. Specific examples where this bias may be away from the null include the scenario where G is binary and the variance of E depends G and the scenario when G has SNP coding and the mean of E depends on G^2. We identified a special case of G-E dependence where the direction of bias in the coefficient for interaction is attenuation, which is the case when the mean of E depends linearly on G.

We also showed that in the naive model ignoring measurement error, G-E dependence may lead to inflated Type I error rates of the test for gene-environment interaction. This problem arises from both the bias of the parameter estimates and the non-constant variance which is not reflected in the naive standard error estimate. In the special case of G-E dependence where the mean of the environmental exposure depends linearly on the gene, we showed that the Type I error rate is preserved.

There is much discussion of the impact of measurement error on detecting gene-environment interactions when G and E are independent. It is well-known that in that case, non-differential measurement error will cause attenuation toward the null in the coefficient of the interaction, and there is much concern for the loss of power.(Thomas, 2010; Bookman et al., 2011)

Only one recent study has investigated the role of G-E dependence in tests for gene-environment interactions.(Lindstrom et al., 2009) In that study, the only setting considered is a logistic regression model when the exposure and gene are both binary. Tests for gene-environment interactions under exposure measurement error are investigated via simulation. The authors find that the test for gene-environment interaction does not have inflated Type I error rates, but tests for the effect of gene and joint tests for gene and interaction have highly inflated Type I error rates. Although that paper considers a logistic health model while our paper considers a linear health model, their results are intuitively consistent with our findings in the special case of linear dependence between G and E because of the upward bias we observed in the estimate for G. Our work considers more general G-E dependence models and provides the theoretical justification for how and when the parameter biases and Type I error violations will occur.

The issue of G-E dependence and its impact on the modeling of health effects has received increasing attention. The question of what types of models may best reflect true G-E dependence is still largely unknown, especially because it is not clear how many studies have tested for non-linear effects of genes on environmental factors. One example of the variance of E depending on G is seen in a recent study in Nature, where the FTO SNP rs7202116 was associated with the variability of BMI.(Yang et al., 2012)

In this paper, we consider two functional measurement error correction methods: regression calibration and SIMEX. These methods are called functional because they do not assume any particular distribution for the covariate measured with error. However, for regression calibration the correct regression calibration model must be used, including the correct G-E dependence structure. Our simulation results demonstrate that regression calibration is very sensitive to using the correct model, making this a key assumption. Since the appeal of functional measurement error models is making fewer distributional assumptions, SIMEX is the preferred method here. SIMEX is also preferred here because there are no gains in efficiency when using regression calibration, despite the added assumptions. Many other measurement error correction techniques could also be applied to this situation; Carroll et al. (2006) discusses a number of different measurement error correction methodologies and their assumptions.

1.9 References

D. Allison and S. Saunders. Obesity in north america. an overview. *Med Clin North Am*, 84:305–332, 2000.

A. Bennet, E. DiAngelantonio, Z. Ye, F. Wensley, A. Dahlin, A. Ahlbom, and et al. Association of apolipoprotein e genotypes with lipid levels and coronary risk. *JAMA*, 298: 1300–1311, 2007.

J. Boer, C. Ehnholm, H. Menzel, L. Havekes, M. Rosseneu, D. O'Reilly, and L. Tiret. Interactions between lifestyle-related factors and the apoe polymorphism on plasma lipids and apolipoproteins. the european atherosclerosis research study. *Arterioscler Thromb Vasc Biol*, 17:1675–1681, 1997.

E. B. Bookman, K. McAllister, E. Gillanders, K. Wanke, D. Balshaw, J. Rutter, J. Reedy, D. Shaughnessy, T. Agurs-Collins, D. Paltoo, A. Atienza, L. Bierut, P. Kraft, M. D. Fallin, F. Perera, E. Turkheimer, J. Boardman, M. L. Marazita, S. M. Rappaport, E. Boerwinkle, S. J. Suomi, N. E. Caporaso, I. Hertz-Picciotto, K. C. Jacobson, W. L. Lowe, L. R. Goldman, P. Duggal, M. R. Gunnar, T. A. Manolio, E. D. Green, D. H. Olster, and L. S. Birnbaum. Gene-environment interplay in common complex diseases: forging an integrative modelrecommendations from an nih workshop. *Genetic Epidemiology*, 35:217225, 2011.

R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective.* Chapman & Hall, New York, New York, 2nd edition, 2006.

J. R. Cook and L. A. Stefanski. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89:13141328, 1994.

L. Farrer, L. Cupples, J. Haines, B. Hyman, W. Kukull, R. Mayeux, R. Myers, M. Pericak-Vance, N. Risch, and C. vanDuijn. Effects of age, sex, and ethnicity on the association between apolipoprotein e genotype and alzheimer disease: a meta-analysis. *JAMA*, 278: 1349–1356, 1997.

M. Garcia-Closas, W. Thompson, and J. Robins. Differential misclassification and the assessment of gene-environment interactions in case-control studies. *Am J Epidemiol*, 147 (5):426–433, 1998.

M. Garcia-Closas, N. Rothman, and J. Lubin. Misclassification in case-control studies of gene-environment interactions: assessment of bias and sample size. *Cancer Epidemiology Biomarkers & Prevention*, 8:1043–1050, 1999.

A. Jackson, P. Stanforth, J. Gagnon, T. Rankinen, A. Leon, D. Rao, J. Skinner, C. Bouchard, and J. Wilmore. The effect of sex, age and race on estimating percentage body fat from body mass index: The heritage family study. *International Journal of Obesity*, 26:789–796, 2002.

I. Kawachi, D. Sparrow, P. S. Vokonas, and S. Weiss. Symptoms of anxiety and risk of coronary heart disease. the normative aging study. *Circulation*, 90:2225–2229, 1994.

S. Lindstrom, Y. Yen, D. Spiegelman, and P. Kraft. The impact of gene-environment dependence and misclassification in genetic association studies incorporating gene-environment interactions. *Human heredity*, 68:171–181, 2009.

T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore,

M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. Mackay, S. A. McCarroll, and P. M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

P. Marques-Vidal, V. Bongard, J. Ruidavets, J. Fauvel, H. Hanaire-Broutin, B. Perret, and J. Ferrieres. Obesity and alcohol modulate the effect of apolipoprotein e polymorphism on lipids and insulin. *Obes Res*, 11:1200–1206, 2003.

M. Pardo Silva, A. Janssens, A. Hofman, J. Witteman, and C. vanDuijn. Apolipoprotein e gene is related to mortality only in normal weight individuals: the rotterdam study. *Eur J Epidemiol*, 23:135–142, 2008.

N. R. Shah and E. R. Braverman. Measuring adiposity in patients: The utility of body mass index (bmi), percent body fat, and leptin. *PLoS ONE*, 4:e33308, 2012.

S. Srinivasan, C. Ehnholm, A. Elkasabany, and G. Berenson. Apolipoprotein e polymorphism modulates the association between obesity and dyslipidemias during young adulthood: the bogalusa heart study. *Metabolism*, 50:696–702, 2001.

D. Thomas. Geneenvironment-wide association studies: emerging approaches. *Nature Reviews Genetics*, 11:259–272, 2010.

K. Volcik, R. Barkley, R. Hutchinson, T. Mosley, G. Heiss, A. Sharrett, and et al. Apolipoprotein e polymorphisms predict low density lipoprotein cholesterol levels and carotid artery wall thickness but not incident coronary heart disease in 12,491 aric study participants. *Am J Epidemiol*, 164:342–348, 2006.

P. Wilson, S. E.J., M. Larson, and J. Ordovas. Apolipoprotein e alleles and risk of coronary disease: a meta-analysis. *Arterioscler Thromb Vasc Biol*, 16:1250–1255, 1996.

M. Wong, N. Day, J. Luan, and N. Wareham. Estimation of magnitude in gene-environment interactions in the presence of measurement error. *Statistics in Medicine*, 23:987–998, 2004.

J. Yang, R. J. F. Loos, J. E. Powell, S. E. Medland, E. K. Speliotes, D. I. Chasman, L. M. Rose, G. Thorleifsson, V. Steinthorsdottir, R. Mgi, L. Waite, A. V. Smith, L. M. Yerges-Armstrong, K. L. Monda, D. Hadley, A. Mahajan, G. Li, K. Kapur, V. Vitart, J. E. Huffman, S. R. Wang, C. Palmer, T. Esko, K. Fischer, J. H. Zhao, A. Demirkan, A. Isaacs, M. F. Feitosa, J. Luan, N. L. Heard-Costa, C. White, A. U. Jackson, M. Preuss, A. Ziegler, J. Eriksson, Z. Kutalik, F. Frau, I. M. Nolte, J. V. V. Vliet-Ostaptchouk, J.-J. Hottenga, K. B. Jacobs, N. Verweij, A. Goel, C. Medina-Gomez, K. Estrada, J. L. Bragg-Gresham, S. Sanna, C. Sidore, J. Tyrer, A. Teumer, I. Prokopenko, M. Mangino, C. M. Lindgren, T. L. Assimes, A. R. Shuldiner, J. Hui, J. P. Beilby, W. L. McArdle, P. Hall, T. Haritunians, L. Zgaga, I. Kolcic, O. Polasek, T. Zemunik, B. A. Oostra, M. J. Junttila, H. Grnberg, S. Schreiber, A. Peters, A. A. Hicks, J. Stephens, N. S. Foad, J. Laitinen, A. Pouta, M. Kaakinen, G. Willemsen, J. M. Vink, S. H. Wild, G. Navis, F. W. Asselbergs, G. Homuth, U. John, C. Iribarren, T. Harris, L. Launer, V. Gudnason, J. R. OConnell, E. Boerwinkle, G. Cadby, L. J. Palmer, A. L. James, A. W. Musk, E. Ingelsson, B. M. Psaty, J. S. Beckmann, G. Waeber, P. Vollenweider, C. Hayward, A. F. Wright, I. Rudan, L. C. Groop, A. Metspalu, K.-T. Khaw, C. M. van Duijn, I. B. Borecki, M. A. Province, N. J. Wareham, J.-C. Tardif, H. V. Huikuri, L. A. Cupples, L. D. Atwood, C. S. Fox, M. Boehnke, F. S. Collins, K. L. Mohlke, J. Erdmann, H. Schunkert, C. Hengstenberg, K. Stark, M. Lorentzon, C. Ohlsson, D. Cusi, J. A. Staessen, M. M. V. der Klauw, P. P. Pramstaller, S. Kathiresan, J. D. Jolley, S. Ripatti, M.-R. Jarvelin, E. J. C. de Geus, D. I. Boomsma, B. Penninx, J. F. Wilson, H. Campbell, S. J. Chanock, P. van der Harst, A. Hamsten, H. Watkins, A. Hofman, J. C. Witteman, M. C. Zillikens, A. G. Uitterlinden, F. Rivadeneira, M. C. Zillikens, L. A. Kiemeney, S. H. Vermeulen, G. R. Abecasis, D. Schlessinger, S. Schipf, M. Stumvoll, A. Tnjes, T. D. Spector, K. E. North, G. Lettre, M. I. McCarthy, S. I. Berndt, A. C. Heath, P. A. F. Madden, D. R. Nyholt, G. W. Montgomery, N. G. Martin, B. McKnight, D. P. Strachan, W. G. Hill, H. Snieder, P. M. Ridker, U. Thorsteinsdottir, K. Stefansson, T. M. Frayling, J. N. Hirschhorn, M. E. Goddard, and P. M. Visscher. Fto genotype is associated with phenotypic variability of body mass index. *Nature*, 490:267–272, 2012.

L. Zhang, B. Mukherjee, M. Ghosh, S. Gruber, and V. Moreno. Accounting for error due to misclassification of exposures in case-control studies of gene-environment interaction. *Statistics in Medicine*, 27:2756–2783, 2008.

1.A Derivations

1.A.1 Derivation of equation (1.5)

Starting from equation the score equations, we have

$$
\begin{aligned}
(1.3) \quad \Leftrightarrow \quad & \mathbb{E}\left\{\mathcal{W}_1\mathcal{W}_1^{\mathsf{T}}\right\}\boldsymbol{\theta}_{Naive} = \mathbb{E}\Big\{\mathcal{W}_1 Y_1\Big\} \\
\Leftrightarrow \quad & \mathbb{E}\left\{\mathcal{W}_1\mathcal{W}_1^{\mathsf{T}}\right\}\boldsymbol{\theta}_{Naive} = \mathbb{E}\Big\{\mathbb{E}\left[\mathcal{W}_1 Y_1 | G_1, E_1\right]\Big\} \\
\Leftrightarrow \quad & \mathbb{E}\left\{\mathcal{W}_1\mathcal{W}_1^{\mathsf{T}}\right\}\boldsymbol{\theta}_{Naive} = \mathbb{E}\Big\{\mathcal{W}_1 \mathbb{E}\left[Y_1 | G_1, E_1\right]\Big\} \\
\Leftrightarrow \quad & \mathbb{E}\left\{\mathcal{W}_1\mathcal{W}_1^{\mathsf{T}}\right\}\boldsymbol{\theta}_{Naive} = \mathbb{E}\Big\{\mathcal{W}_1 \mathcal{X}_1^{\mathsf{T}}\boldsymbol{\theta}\Big\} \\
\Leftrightarrow \quad & \mathbb{E}\left\{\mathcal{W}_1\mathcal{W}_1^{\mathsf{T}}\right\}\boldsymbol{\theta}_{Naive} = \mathbb{E}\Big\{(\mathcal{X}_1 + \mathcal{U}_1)\mathcal{X}_1^{\mathsf{T}}\Big\}\boldsymbol{\theta} \\
\Leftrightarrow \quad & \mathbb{E}\left\{\mathcal{W}_1\mathcal{W}_1^{\mathsf{T}}\right\}\boldsymbol{\theta}_{Naive} = \mathbb{E}\Big\{\mathcal{X}_1\mathcal{X}_1^{\mathsf{T}}\Big\}\boldsymbol{\theta} \qquad (1.15)
\end{aligned}
$$

Then by properties of random vectors we can rewrite each expecation as

$$\mathbb{E}\left\{\mathcal{X}_1\mathcal{X}_1^{\mathsf{T}}\right\} = \mathrm{Var}(\mathcal{X}_1) + \mu_{\mathcal{X}}\mu_{\mathcal{X}}^{\mathsf{T}} \quad \text{and} \quad \mathbb{E}\left\{\mathcal{W}_1\mathcal{W}_1^{\mathsf{T}}\right\} = \mathrm{Var}(\mathcal{W}_1) + \mu_{\mathcal{W}}\mu_{\mathcal{W}}^{\mathsf{T}}$$

Substituting these expressions gives us a system of equations that we can solve algebraically for $\boldsymbol{\theta}_{Naive}$. Note that the solution to the system of equations depends on the covariance matrix of the covariates, $\mathrm{Cov}(\mathcal{X}_1)$ and $\mathrm{Cov}(\mathcal{W}_1)$. Thus,

$$\left[\mathrm{Cov}(\mathcal{W}_1)\right]\boldsymbol{\theta}_{Naive} = \left[\mathrm{Cov}(\mathcal{X}_1)\right]\boldsymbol{\theta}$$

1.A.2 Derivation of equation (1.7)

$$
\begin{aligned}
\sigma^2_{Naive} &= \mathbb{E}\left[Y_1 - \mathcal{W}_1^{\mathsf{T}}\boldsymbol{\theta}_{Naive}\right]^2 \\
&= \mathbb{E}\left[\mathcal{X}_1^{\mathsf{T}}\boldsymbol{\theta} + \epsilon_1 - \mathcal{W}_1^{\mathsf{T}}\boldsymbol{\theta}_{Naive}\right]^2 \\
&= \sigma^2 + \mathbb{E}\left[\mathcal{X}_1^{\mathsf{T}}\boldsymbol{\theta} - \mathcal{W}_1^{\mathsf{T}}\boldsymbol{\theta}_{Naive}\right]^2 \\
&= \sigma^2 + \mathbb{E}\Big[\mathcal{X}_1^{\mathsf{T}}\boldsymbol{\theta} - \mathcal{W}_1^{\mathsf{T}}\boldsymbol{\theta}_{Naive}\Big]^2 \\
&= \sigma^2 + \mathbb{E}\Big[\mathcal{X}_1^{\mathsf{T}}\boldsymbol{\theta} - \mathcal{W}_1^{\mathsf{T}}\boldsymbol{\Lambda}\boldsymbol{\theta}\Big]^2 \\
&= \sigma^2 + \mathbb{E}\Big[\left(\mathcal{X}_1^{\mathsf{T}} - \mathcal{W}_1^{\mathsf{T}}\boldsymbol{\Lambda}\right)\boldsymbol{\theta}\Big]^2 \\
&= \sigma^2 + \boldsymbol{\theta}^{\mathsf{T}}\mathbb{E}\Big[\left(\mathcal{X}_1^{\mathsf{T}} - \mathcal{W}_1^{\mathsf{T}}\boldsymbol{\Lambda}\right)^{\mathsf{T}}\left(\mathcal{X}_1^{\mathsf{T}} - \mathcal{W}_1^{\mathsf{T}}\boldsymbol{\Lambda}\right)\Big]\boldsymbol{\theta} \\
&= \sigma^2 + \boldsymbol{\theta}^{\mathsf{T}}\Big[\Sigma_X - \Sigma_X\boldsymbol{\Lambda} - \Sigma_X\boldsymbol{\Lambda}^{\mathsf{T}} + \boldsymbol{\Lambda}\Sigma_W\boldsymbol{\Lambda}^{\mathsf{T}}\Big]\boldsymbol{\theta} \\
&= \sigma^2 + \boldsymbol{\theta}^{\mathsf{T}}\Big[\left(\mathbf{I} - \boldsymbol{\Lambda}^{\mathsf{T}}\right)Cov(\mathcal{X}_1)\Big]\boldsymbol{\theta}
\end{aligned}
$$

All the cross-terms involving U_1 have expectation 0.

1.A.3 Derivation of entires of $\boldsymbol{\Lambda}$ under G-E dependence

The covariance matrix for the true covariates is given by

$$
\mathrm{Var}(\mathcal{X}_1) = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 \\
0 & \sigma_X^2 & 0 & 0 & 0 \\
0 & 0 & \sigma_G^2 & \rho_1\sigma_E\sigma_G & \rho_2\sigma_G\sigma_{EG} \\
0 & 0 & \rho_1\sigma_E\sigma_G & \sigma_E^2 & \rho_3\sigma_E\sigma_{EG} \\
0 & 0 & \rho_2\sigma_G\sigma_{EG} & \rho_3\sigma_E\sigma_{EG} & \sigma_{EG}^2
\end{pmatrix}
$$

And the covariance matrix for the covariates measured with error $\mathcal{W}_1 = (1, \tilde{E}_1, G_1, \tilde{E}_1 G_1)$, is given by,

$$\text{Var}(\mathcal{W}_1) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_X^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_G^2 & \rho_1\sigma_E\sigma_G & \rho_2\sigma_G\sigma_{EG} \\ 0 & 0 & \rho_1\sigma_E\sigma_G & \sigma_E^2 + \sigma_U^2 & \rho_3\sigma_E\sigma_{EG} \\ 0 & 0 & \rho_2\sigma_G\sigma_{EG} & \rho_3\sigma_E\sigma_{EG} & \sigma_{EG}^2 + \sigma_U^2\sigma_G^2 \end{pmatrix}$$

1.A.4 Derivation of equation (1.9)

To calculate the conditional variance for any observation i, we use the property of iterated expectations,

$$\begin{aligned} Var\left\{Y_i|\mathcal{W}_i\right\} =& \mathbb{E}\Big\{Var[Y_i|\mathcal{X}_i, \mathcal{W}_i]\Big|\mathcal{W}_i\Big\} + Var\Big\{\mathbb{E}[Y_i|\mathcal{X}_i, \mathcal{W}_i]\Big|W_i\Big\} \\ =& \mathbb{E}\Big\{\sigma^2\Big|W_i\Big\} + Var\Big\{\theta X_i\Big|W_i\Big\} \\ =& \sigma^2 + Var\Big\{\alpha_0 + \alpha_G G_i + \alpha_E E_i + \beta E_i G_i\Big|(G_i, \widetilde{E}_i)\Big\} \\ =& \sigma^2 + Var\Big\{\alpha_E E_i\Big|(G_i, \widetilde{E}_i)\Big\} + Var\Big\{\beta E_i G_i\Big|(G_i, \widetilde{E}_i)\Big\} \\ &+ 2Cov\Big\{\alpha_E E_i, \beta E_i G_i\Big|(G_i, \widetilde{E}_i)\Big\} \\ =& \sigma^2 + \alpha_E^2 Var\Big\{E_i\Big|(G_i, \widetilde{E}_i)\Big\} + \beta^2 G_i^2 Var\Big\{E_i\Big|(G_i, \widetilde{E}_i)\Big\} \\ &+ 2\alpha_E\beta G_i Cov\Big\{E_i, E_i\Big|(G_i, \widetilde{E}_i)\Big\} \\ =& \sigma^2 + \Big(\alpha_E^2 + \beta^2 G_i^2 + 2\alpha_E\beta G_i\Big)Var\Big\{E_i\Big|(G_i, \widetilde{E}_i)\Big\} \end{aligned}$$

1.A.5 Derivation of variance of naive model MLE's in Scenario (a)

To show that

$$Var(\hat{\boldsymbol{\theta}}_{Naive}|\mathbf{W}) = \sigma^2_{Naive}(\mathbf{W}^\intercal\mathbf{W})^{-1} = \sigma^2 + \lambda'\sigma_U^2\alpha_E^2,$$

we derive $Var\left\{E_i\middle|(G_i, \widetilde{E}_i)\right\}$ and substitute into equation (1.10). Using independence and multivariate Normality and $\lambda' = \sigma_E^2(\sigma_E^2 + \sigma_U^2)^{-1}$,

$$
\begin{aligned}
Var\left\{E_i\middle|(G_i, \widetilde{E}_i)\right\} &= Var\left\{E_i\middle|\widetilde{E}_i\right\} \\
&= \sigma_E^2 - \sigma_E^2(\sigma_E^2 + \sigma_U^2)^{-1}\sigma_\delta^2 \\
&= \sigma_E^2\left(1 - \sigma_E^2(\sigma_E^2 + \sigma_U^2)^{-1}\right) \\
&= \sigma_E^2\sigma_U^2(\sigma_E^2 + \sigma_U^2)^{-1} \\
&= \lambda'\sigma_U^2.
\end{aligned}
$$

1.A.6 Derivation of variance of naive model MLE's in Scenario (b)

In Scenario (b), the G-E dependence relationship is modeled as $E_i = \gamma_0 + \gamma_1 G_i + \delta_i$, and $\delta_i \sim Normal(0, \sigma_\delta^2)$. Then,

$$
\begin{bmatrix} E_i|G_i \\ \widetilde{E}_i|G_i \end{bmatrix} \sim MVN\left(\begin{bmatrix} \gamma_0 + \gamma_1 G_i \\ \gamma_0 + \gamma_1 G_i \end{bmatrix}, \begin{bmatrix} \sigma_\delta^2 & \sigma_\delta^2 \\ \sigma_\delta^2 & \sigma_\delta^2 + \sigma_U^2 \end{bmatrix} \right)
$$

By the properties of the multivariate Normal distribution, the conditional distribution $E_i|(G_i, \widetilde{E}_i)$ is Normal with variance

$$
\begin{aligned}
Var\left\{E_i\middle|(G_i, \widetilde{E}_i)\right\} &= \sigma_\delta^2 - \sigma_\delta^2(\sigma_\delta^2 + \sigma_U^2)^{-1}\sigma_\delta^2 \\
&= \sigma_\delta^2\left(1 - \sigma_\delta^2(\sigma_\delta^2 + \sigma_U^2)^{-1}\right) \\
&= \sigma_\delta^2\sigma_U^2(\sigma_\delta^2 + \sigma_U^2)^{-1}
\end{aligned}
$$

which simplifies the expression for equation (1.10) in this case to

$$
Var\left\{Y_i|\mathcal{W}_i\right\} = \sigma^2 + \alpha_E^2 \cdot \frac{\sigma_\delta^2\sigma_U^2}{\sigma_\delta^2 + \sigma_U^2} \tag{1.16}
$$

1.A.7 Derivation of naive model variance in Scenario (b)

Here we derive our simplified expression for naive model variance, σ^2_{Naive}, in Scenario (b). Starting from equation (1.12), we will show that the second term can be simplified and the last term is equal to zero.

First consider the expression for $\lambda_5\sigma^2_G$. Using algebra and using the result that $\rho_3 = \rho_1\rho_2$ which we know holds when $E_i = \gamma_0 + \gamma_1 G_i + \delta_i$, we have

$$\begin{aligned}
\lambda_5\sigma^2_G &= \frac{1}{\det[Cov(\mathcal{W}_1)]} \cdot \Big[(\rho_1 - \rho_2\rho_3)\sigma^2_{EG} + \rho_1\sigma^2_G\sigma^2_U\Big]\sigma_E\sigma_G\sigma^2_U\sigma^2_G \\
&= \frac{1}{\det[Cov(\mathcal{W}_1)]} \cdot \Big[(\rho_1 - \rho_2\rho_1\rho_2)\sigma^2_{EG} + \rho_1\sigma^2_G\sigma^2_U\Big]\sigma_E\sigma_G\sigma^2_U\sigma^2_G \\
&= \frac{1}{\det[Cov(\mathcal{W}_1)]} \cdot \Big[\rho_1(1 - \rho^2_2)\sigma^2_{EG} + \rho_1\sigma^2_G\sigma^2_U\Big]\sigma_E\sigma_G\sigma^2_U\sigma^2_G \\
&= \rho_1\sigma_G\sigma_E \cdot \frac{1}{\det[Cov(\mathcal{W}_1)]} \cdot \Big[(1 - \rho^2_2)\sigma^2_{EG} + \sigma^2_G\sigma^2_U\Big]\sigma^2_G\sigma^2_U \\
&= (1 - \lambda_3)\rho_1\sigma_G\sigma_E
\end{aligned}$$

Thus, the last term in equation (1.12) is equal to zero,

$$\alpha_G\alpha_E\Big[-\lambda_5\sigma^2_G + (1 - \lambda_3)\rho_1\sigma_G\sigma_E\Big] = 0.$$

Using this result and the results from the linear dependence that $\rho_1 = \gamma_1\sigma_G/\sigma_E$ and $\sigma^2_E =$

$\gamma_1^2\sigma_G^2 + \sigma_\delta^2$ to simplify the expression for the second term of equation (1.12) ,

$$\begin{aligned}
\alpha_E^2\Big[-\lambda_5\rho_1\sigma_G\sigma_E + (1-\lambda_3)\sigma_E^2\Big] &= \alpha_E^2\left[-\lambda_5\rho_1\sigma_G\sigma_E + \left(\frac{\lambda_5\sigma_G^2}{\rho_1\sigma_G\sigma_E}\right)\sigma_E^2\right] \\
&= \alpha_E^2\left[-\lambda_5\left(\gamma_1\frac{\sigma_G}{\sigma_E}\right)\sigma_G\sigma_E + \left(\frac{\sigma_E}{\gamma_1\sigma_G}\cdot\frac{\lambda_5\sigma_G^2}{\sigma_G\sigma_E}\right)\sigma_E^2\right] \\
&= \alpha_E^2\lambda_5\left[-\gamma_1\sigma_G^2 + \left(\frac{1}{\gamma_1}\right)\sigma_E^2\right] \\
&= \alpha_E^2\lambda_5\left[-\gamma_1\sigma_G^2 + \frac{1}{\gamma_1}\Big(\gamma_1^2\sigma_G^2 + \sigma_\delta^2\Big)\right] \\
&= \alpha_E^2\lambda_5\left[\frac{1}{\gamma_1}\sigma_\delta^2\right]
\end{aligned}$$

Then, using the result that $\sigma_E^2(1-\rho_1^2) = \sigma_\delta^2$, we can further simplify this expression,

$$\begin{aligned}
\lambda_5\left[\frac{1}{\gamma_1}\sigma_\delta^2\right] &= \frac{1}{\det[Cov(\mathcal{W}_1)]}\cdot\Big[(\rho_1-\rho_2\rho_3)\sigma_{EG}^2 + \rho_1\sigma_G^2\sigma_U^2\Big]\sigma_E\sigma_G\sigma_U^2\left[\frac{1}{\gamma_1}\sigma_\delta^2\right] \\
&= \sigma_U^2\sigma_\delta^2\cdot\frac{1}{\gamma_1}\cdot\frac{1}{\det[Cov(\mathcal{W}_1)]}\cdot\rho_1\Big[(1-\rho_2^2)\sigma_{EG}^2 + \sigma_G^2\sigma_U^2\Big]\sigma_E\sigma_G \\
&= \sigma_U^2\sigma_\delta^2\cdot\frac{[(1-\rho_2^2)\sigma_{EG}^2 + \sigma_G^2\sigma_U^2]}{\sigma_U^2\,[\sigma_G^2\sigma_U^2 + (1-\rho_2^2)\sigma_{EG}^2] + (1-\rho_1^2)\sigma_E^2\,[\sigma_G^2\sigma_U^2 + (1-\rho_2^2)\sigma_{EG}^2]} \\
&= \sigma_U^2\sigma_\delta^2\cdot\frac{1}{\sigma_U^2 + (1-\rho_1^2)\sigma_E^2} \\
&= \frac{\sigma_\delta^2\sigma_U^2}{\sigma_\delta^2 + \sigma_U^2}
\end{aligned}$$

Finally,

$$\begin{aligned}
\sigma_{Naive}^2 &= \sigma^2 + \alpha_E^2\Big[-\lambda_5\rho_1\sigma_G\sigma_E + (1-\lambda_3)\sigma_E^2\Big] + \alpha_G\alpha_E\Big[-\lambda_5\sigma_G^2 + (1-\lambda_3)\rho_1\sigma_G\sigma_E\Big] \\
&= \sigma^2 + \alpha_E^2\cdot\frac{\sigma_\delta^2\sigma_U^2}{\sigma_\delta^2 + \sigma_U^2}
\end{aligned}$$

1.A.8 Regression calibration derivations in Scenarios (b) and (d)

First consider Scenario (d) where the true G-E dependence relationship is modeled as $E_i = \gamma_0 + \gamma_1 G_i^2 + \delta_i$, and $\delta_i \sim Normal(0, \sigma_\delta^2)$. Then,

$$\begin{bmatrix} E_i|G_i \\ \widetilde{E}_i|G_i \end{bmatrix} \sim MVN\left(\begin{bmatrix} \gamma_0 + \gamma_1 G_i^2 \\ \gamma_0 + \gamma_1 G_i^2 \end{bmatrix}, \begin{bmatrix} \sigma_\delta^2 & \sigma_\delta^2 \\ \sigma_\delta^2 & \sigma_\delta^2 + \sigma_U^2 \end{bmatrix} \right)$$

By the properties of the multivariate Normal distribution, the conditional distribution $E_i|(G_i, \widetilde{E}_i)$ is Normal with expectation

$$\begin{aligned} \mathbb{E}\Big\{E_i\Big|(G_i, \widetilde{E}_i)\Big\} &= (\gamma_0 + \gamma_1 G_i^2) + \sigma_\delta^2(\sigma_\delta^2 + \sigma_U^2)^{-1}\left[\tilde{E}_i - (\gamma_0 + \gamma_1 G_i^2)\right] \\ &= \underbrace{\gamma_0\left[1 - \sigma_\delta^2(\sigma_\delta^2 + \sigma_U^2)^{-1}\right]}_{\eta_0} + \underbrace{\left[\sigma_\delta^2(\sigma_\delta^2 + \sigma_U^2)^{-1}\right]}_{\eta_E}\tilde{E}_i + \underbrace{\gamma_1\left[1 - \sigma_\delta^2(\sigma_\delta^2 + \sigma_U^2)^{-1}\right]}_{\eta_G} G_i^2 \end{aligned}$$

and variance

$$\begin{aligned} Var\Big\{E_i\Big|(G_i, \widetilde{E}_i)\Big\} &= \sigma_\delta^2 - \sigma_\delta^2(\sigma_\delta^2 + \sigma_U^2)^{-1}\sigma_\delta^2 \\ &= \sigma_\delta^2\left(1 - \sigma_\delta^2(\sigma_\delta^2 + \sigma_U^2)^{-1}\right) \\ &= \sigma_\delta^2\sigma_U^2(\sigma_\delta^2 + \sigma_U^2)^{-1} \end{aligned}$$

Thus, η_0, η_G, η_E define the true regression calibration parameters for Scendrio (d).

Following the derivations used to compute the biases in the naive model, we start with the score equations and fine that the regression model coefficients $\boldsymbol{\theta}_{RC}$ satisfy

$$\mathbb{E}\left\{\mathcal{W}_1\mathcal{W}_1^\mathsf{T}\right\}\boldsymbol{\theta}_{RC} = \mathbb{E}\Big\{\mathcal{W}_1\mathcal{X}_1^\mathsf{T}\Big\}\boldsymbol{\theta} \tag{1.17}$$

Comparison of terms show that each entry j, l of these matrices are equal, $\mathbb{E}\left\{\mathcal{W}_1\mathcal{W}_1^\mathsf{T}\right\}_{j,l} =$

$\mathbb{E}\left\{\mathcal{W}_1\mathcal{X}_1^\mathsf{T}\right\}_{j,l}$. Thus, $\boldsymbol{\theta}_{RC} = \boldsymbol{\theta}$.

The derivations for Scenario (b) follow the same steps with the slight change to the G-E dependence model.

1.A.9 Bias Derivations for Non Centered Covariates in Section 1.4

Starting from the score equations and following the derivation of equation (1.5), equation (1.3) can be written as

$$\mathbb{E}\left\{\mathcal{W}_1\mathcal{W}_1^\mathsf{T}\right\}\boldsymbol{\theta}_{Naive} = \mathbb{E}\Big\{\mathcal{X}_1\mathcal{X}_1^\mathsf{T}\Big\}\boldsymbol{\theta} \tag{1.18}$$

By properties of random vectors, each expecation can be rewritten as

$$\mathbb{E}\left\{\mathcal{X}_1\mathcal{X}_1^\mathsf{T}\right\} = \mathrm{Var}(\mathcal{X}_1) + \mu_{\mathcal{X}}\mu_{\mathcal{X}}^\mathsf{T} \quad \text{and} \quad \mathbb{E}\left\{\mathcal{W}_1\mathcal{W}_1^\mathsf{T}\right\} = \mathrm{Var}(\mathcal{W}_1) + \mu_{\mathcal{W}}\mu_{\mathcal{W}}^\mathsf{T}$$

Then, allowing $\mu_{\mathcal{X}}$ and $\mu_{\mathcal{W}}$ to have nonzero means, this can be solved for to get a more general version of equations (1.5) and (1.6).

Suppose that E and G are correlated and have nonzero means μ_E and μ_G respectively. Let the correlations be parameterized as $\rho_1 \equiv Corr(E_1, G_1)$, $\rho_2 \equiv Corr(G_1, E_1G_1)$, $\rho_3 \equiv Corr(E_1, E_1G_1)$.

Under this general parameterization, allowing E and G to have nonzero means and not yet assuming any particular model for $E|G$, the solutions to equation (1.5) can be written as

$$\boxed{\begin{aligned}
\alpha_{0,Naive} &= \alpha_0 + [-\mu_G\lambda_5 + \mu_E(1-\lambda_3) - \mu_{EG}\lambda_1]\alpha_E + [-\mu_G\lambda_6 - \mu_E\lambda_4 + \mu_{EG}(1-\lambda_2)]\beta \\
\alpha_{G,Naive} &= \alpha_G + \lambda_5\alpha_E + \lambda_6\beta \\
\alpha_{E,Naive} &= \lambda_3\alpha_E + \lambda_4\beta \\
\beta_{Naive} &= \lambda_1\alpha_E + \lambda_2\beta
\end{aligned}}$$

where

$$\lambda_1 = \frac{1}{\det[Cov(\mathcal{W}_1)]} \cdot \Big[(\rho_3 - \rho_1\rho_2)\, \sigma_E \sigma_{EG} - \big(1-\rho_1^2\big)\, \sigma_E^2 \mu_G \Big] \sigma_G^2 \sigma_U^2$$

$$\lambda_2 = \frac{1}{\det[Cov(\mathcal{W}_1)]} \cdot \Big[\big(1 + 2\rho_1\rho_2\rho_3 - \rho_1^2 - \rho_2^2 - \rho_3^2\big)\, \sigma_E^2 \sigma_{EG}^2 + \big(1-\rho_2^2\big)\, \sigma_{EG}^2 \sigma_U^2$$

$$+ (\rho_1\rho_2 - \rho_3)\, \sigma_E \sigma_{EG} \sigma_U^2 \mu_G \Big] \sigma_G^2$$

$$\lambda_3 = \frac{1}{\det[Cov(\mathcal{W}_1)]} \cdot \Big[\big(1 + 2\rho_1\rho_2\rho_3 - \rho_1^2 - \rho_2^2 - \rho_3^2\big)\, \sigma_E^2 \sigma_{EG}^2 + \big(1-\rho_1^2\big)\, (\sigma_G^2 + \mu_G^2) \sigma_E^2 \sigma_U^2$$

$$+ (\rho_1\rho_2 - \rho_3)\, \sigma_E \sigma_{EG} \sigma_U^2 \mu_G \Big] \sigma_G^2$$

$$\lambda_4 = \frac{1}{\det[Cov(\mathcal{W}_1)]} \cdot \Big[(\rho_3 - \rho_1\rho_2)\, \sigma_E \sigma_{EG} (\sigma_G^2 + \mu_G^2) - (1-\rho_2^2) \sigma_{EG}^2 \mu_G \Big] \sigma_G^2 \sigma_U^2$$

$$\lambda_5 = \frac{1}{\det[Cov(\mathcal{W}_1)]} \cdot \Big[(\rho_1 - \rho_2\rho_3) \sigma_{EG}^2 + \rho_1 \sigma_G^2 \sigma_U^2 + (\rho_2 - \rho_1\rho_3) \sigma_E \sigma_{EG} \mu_G \Big] \sigma_E \sigma_G \sigma_U^2$$

$$\lambda_6 = \frac{1}{\det[Cov(\mathcal{W}_1)]} \cdot \Big[(\rho_2 - \rho_1\rho_3) \sigma_E^2 (\sigma_G^2 + \mu_G^2) + \rho_2 \sigma_G^2 \sigma_U^2 + (\rho_1 - \rho_2\rho_3) \sigma_E \sigma_{EG} \mu_G \Big] \sigma_{EG} \sigma_G \sigma_U^2$$

and

$$\det[Cov(\mathcal{W}_1)] = \Big[\sigma_G^2 \sigma_U^2 + \big(1-\rho_2^2\big)\, \sigma_{EG}^2 + \big(1-\rho_1^2\big)\, \sigma_E^2 (\sigma_G^2 + \mu_G^2) \Big] \sigma_G^2 \sigma_U^2$$

$$+ \big(1 + 2\rho_1\rho_2\rho_3 - \rho_1^2 - \rho_2^2 - \rho_3^2\big)\, \sigma_E^2 \sigma_G^2 \sigma_{EG}^2 + 2\, (\rho_1\rho_2 - \rho_3)\, \sigma_E \sigma_{EG} \sigma_U^2 \sigma_G^2 \mu_G .$$

Chapter 2

Effects of spatial measurement error in health effect analyses using predicted air pollution exposures and correction by spatial SIMEX.

Stacey E. Alexeeff[1], Raymond J. Carroll[2], and Brent A. Coull[1]

[1] Department of Biostatistics, Harvard School of Public Health

[2] Department of Statistics, Texas A & M University

2.1 Introduction

To improve exposure assessment in air pollution epidemiology research, spatial modeling of air-pollution levels is now commonplace. However, the implications of including these predicted exposures on the estimates of the health effects of air pollution is still not well-understood. Levels of $PM_{2.5}$ are typically measured only at a small number of stationary monitoring sites which may only capture a small part of the overall regional heterogeneity. Ambient levels of $PM_{2.5}$ often vary within a given city region, in part due to traffic sources, (Brauer et al., 2003; Clougherty et al., 2008)

It is not usually feasible to obtain exposure recordings at each study subject's residence over an entire study period, researchers often set up pollution monitoring networks to gather data on the variability in traffic pollution levels over space, and then build prediction models based on these data that can be used to estimate location-specific exposures throughout the study region. Oftentimes, we are interested in the level of air pollution at the homes of each subject in a study, but instead of directly monitoring every home we have measurements collected from a set of monitors throughout the region. This setting where the set of exposure locations of interest in the health model does not match the set of measured exposure locations is called *spatial misalignment*. Because of this spatial misalignment, exposure prediction models are used to predict the exposure level at the location of interest.

The most common use of predicted exposures in a health effects analysis is the direct plug-in of the individual-specific exposure estimates. This approach treats the exposures as known, without acknowledgement of the uncertainty in the prediction process. Ignoring this measurement error can lead to biased health effect estimates and overstated confidence in the resulting risk assessments.(Carroll et al., 2006; Gryparis et al., 2009) In the current literature, authors have found in simulation studies that using the plug-in estimator often induces little to no bias.(Szpiro et al., 2011; Madsen et al., 2008; Lopiano et al., 2011; Gryparis et al., 2009) However, those papers investigate bias in simulation studies only by fitting the correct exposure model used to generate the data. Madsen et al. (2008) and Szpiro et al. (2011) assume smooth exposure surfaces that can be fit well using kriging methods, finding no need for bias correction, thus the papers focus on ways to estimate standard errors instead. However, in real data scenarios, the actual performance of the naive plug-in estimator, the degree to which bias and variance adjustments need to be made, and the performance of the current adjustment methods is largely unknown. It is important to identify situations where the plug-in estimator may be unbiased versus situations where both a bias and variance adjustment may be necessary.

This article investigates two factors of exposure estimation that may affect resulting health effect estimates: *estimation error* and *model misspecification*. In practice, the underlying exposure model that generates air pollution levels in any given region is not known exactly and is fit with sparse monitoring data. Since air pollution monitors are often somewhat sparse throughout any given region, we wanted to examine the effect of estimation error given this sparse data setup. In addition, since the true underlying exposure surfaces are not exactly known, model misspecification is important to investigate as well. The impact on biases in the health effect model due to model misspecification in the exposure model also has not been investigated previously.

The simulation extrapolation method (SIMEX) has been developed as a flexible method to correct for bias in the case of classical measurement error.(Cook and Stefanski, 1994) SIMEX is a functional method which uses resampling techniques and has several attractive properties, including placing minimal assumptions on the underlying distribution of the exposures. In addition, it has been suggested that SIMEX may be suitable for several exposures with correlated classical errors.(Carroll et al., 2006) Thus, we were interested in adapting SIMEX procedure for use in correcting air pollution exposure prediction, where we expect that the classical error variance would be spatially correlated across all exposure estimates in the health effect study.

The remainder of this paper is arranged as follows. In Section 2 we introduce our modeling framework and the specific exposure models of interest under universal kriging: (i) constant mean, (ii) land-use covariates, and (iii) land-use covariates with model mis-specification. In Section 3 we examine the bias analytically for each of the exposure models of interest and we derive the probability limits of the misspecified parameters and the resulting classical error variance. In Section 4 we propose a new correlated-error SIMEX correction method for Berkson and classical error mixtures where correlated classical error is added to the exposure predictions to correct for bias. In Section 5 we present a simulation study to investigate the degree of bias induced by each of the exposure models of interest and to demonstrate the performance of our spatial SIMEX correction method. We then illustrate the a SIMEX correction in a study of air pollution and birthweight in the greater Boston area with a kriging model including for air pollution exposure levels in Section 6. We end with a concluding discussion in Section 7.

2.2 Exposure Models in Air Pollution and Health Studies

2.2.1 Model Framework

Consider the health effects model of interest to be a linear regression model for each subject $i = 1, \ldots, n$,

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_{2,i} + \cdots + \beta_j Z_{j,i} + \epsilon_i$$

where Y_i is a continuous health outcome, X_i is the exposure of interest, $Z_{2,i}, \ldots, Z_{j,i}$ are other covariates related to the health outcome, and ϵ_i are independent and identically distributed (i.i.d.) with mean 0 and variance σ_ϵ^2. The goal of the analysis is to estimate β_1, the parameter measuring the association between the health outcome and the exposure of interest.

Let $\mathbf{X}^*$ denote the measured air pollution levels at monitor sites and let $\mathbf{X}$ denote the true unmeasured air pollution exposures corresponding to the health locations. We assume that each subject's exposure X_i is not measured directly at the subject's address location, but exposures $\mathbf{X}^*$ are measured at m monitor locations spread throughout the same geographic region. We assume the setting of *spatial misalignment*, where the set of n exposure locations needed for the health model does not match the set of $m << n$ measured exposure locations where monitors are stationed. Because of this spatial misalignment, exposure prediction models are used to predict the exposure level at the location of interest.

Generally, we consider a universal kriging model for the exposure which may include land use covariates as part of the mean model. Suppose the true pollution process in a given region is generated by a Gaussian Process, which generates $\mathcal{X} = (\mathbf{X}, \mathbf{X}^*)$, all the pollution levels of interest. Suppose that the realizations of this process take a parametric form. Specifically, let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ be the $k \times 1$ vector of exposure model parameters where $\boldsymbol{\theta}_1 = (\theta_{11}, \ldots, \theta_{1k_1})$ are the parameters for the mean model and $\boldsymbol{\theta}_2 = (\theta_{21}, \ldots, \theta_{2k_2})$ are the parameters for the variance model. Then a realization of one surface follows a Multivariate Normal distribution,

$$\begin{bmatrix} \mathbf{X} \\ \mathbf{X}^* \end{bmatrix} \sim MVN\left(\begin{bmatrix} \boldsymbol{\mu}_X(\boldsymbol{\theta}_1) \\ \boldsymbol{\mu}_{X^*}(\boldsymbol{\theta}_1) \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{XX}(\boldsymbol{\theta}_2) & \boldsymbol{\Sigma}_{XX^*}(\boldsymbol{\theta}_2) \\ \boldsymbol{\Sigma}_{X^*X}(\boldsymbol{\theta}_2) & \boldsymbol{\Sigma}_{X^*X^*}(\boldsymbol{\theta}_2) \end{bmatrix} \right) \tag{2.1}$$

By the properties of the Multivariate Normal distribution,

$$E[\mathbf{X}|\mathbf{X}^*] = \boldsymbol{\mu}_X(\boldsymbol{\theta}_1) + \boldsymbol{\Sigma}_{XX^*}(\boldsymbol{\theta}_2)\boldsymbol{\Sigma}^{-1}_{X^*X^*}(\boldsymbol{\theta}_2)\Big\{\mathbf{X}^* - \boldsymbol{\mu}_{X^*}(\boldsymbol{\theta}_1)\Big\} \tag{2.2}$$

Then the function that we are interested in estimating by Kriging to model our exposure is given by

$$g(\boldsymbol{\theta}; \mathbf{X}^*) = \boldsymbol{\mu}_X(\boldsymbol{\theta}_1) + \boldsymbol{\Sigma}_{XX^*}(\boldsymbol{\theta}_2)\boldsymbol{\Sigma}^{-1}_{X^*X^*}(\boldsymbol{\theta}_2)\Big\{\mathbf{X}^* - \boldsymbol{\mu}_{X^*}(\boldsymbol{\theta}_1)\Big\} \tag{2.3}$$

We now consider the specific exposure models of interest.

2.2.2 Specific exposure models of interest

We consider three common exposure model scenarios for the single-pollutant, spatial case to be of interest in this study.

Scenario I: Constant mean (ordinary kriging)

Universal kriging is simplified to ordinary kriging when the mean model is assumed to be constant. Specifically, assume that the mean model is $\boldsymbol{\mu}_{\mathcal{X}}(\boldsymbol{\theta}_1) = \alpha\mathbf{1}$ and we assume that $\boldsymbol{\Sigma}_{\mathcal{X}}(\boldsymbol{\theta}_2)$ follows a Matern family with $\boldsymbol{\theta}_2 = (\phi, \sigma^2)$ where ϕ is the range parameter and σ^2 is the variance parameter. Then we can write equation (3) for this scenario as

$$g(\alpha, \phi, \sigma^2; \mathbf{X}^*) = \alpha\mathbf{1} + \boldsymbol{\Sigma}_{XX^*}(\phi, \sigma^2)\boldsymbol{\Sigma}^{-1}_{X^*X^*}(\phi, \sigma^2)\Big(\mathbf{X}^* - \alpha\mathbf{1}\Big).$$

The Matern family also has a third parameter, κ, which controls the level of smoothness of the Matern function. This parameter is typically chosen and fixed rather than estimated via maximum likelihood.(Moller, 2003) We assume that κ is fixed and known.

Scenario II: Land-use regression with universal kriging

Universal kriging allows the mean model to contain land-use regression covariates, and allows the residuals to have a spatial correlation structure. The mean and variance parameters can be estimated jointly via maximum likelihood. For this scenario, we assume that the mean model depends linearly on an intercept and two spatially-varying covariates, $\mathbf{S_1}$, $\mathbf{S_2}$ via the parameters $\boldsymbol{\theta}_1 = (\alpha_0, \alpha_1, \alpha_2)$. Again we assume that $\boldsymbol{\Sigma}_{\mathcal{X}}(\boldsymbol{\theta}_2)$ follows a Matern family with $\boldsymbol{\theta}_2 = (\phi, \sigma^2)$ where ϕ is the range parameter and σ^2 is the variance parameter. Then we can write equation (3) for this scenario as $g(\alpha, \phi, \sigma^2; \mathbf{X}^*) = \alpha_0\mathbf{1} + \alpha_1\mathbf{S_1} + \alpha_2\mathbf{S_2} + \boldsymbol{\Sigma}_{XX^*}(\phi, \sigma^2)\boldsymbol{\Sigma}^{-1}_{X^*X^*}(\phi, \sigma^2)\Big(\mathbf{X}^* - \big(\alpha_0\mathbf{1} + \alpha_1\mathbf{S_1} + \alpha_2\mathbf{S_2}\big)\Big)$.

We assume that κ, the level of smoothness, is fixed and known, as described above.

Scenario III: Land-use regression with universal kriging, misspecified

Here we assume that the true model is the same as Scenario II, but we consider a potentially common form of model misspecification where the model is fit without $\mathbf{S_2}$. This could happen if data were not collected on all key predictors, and hence the prediction model that is fit is an imperfect model. We assume that $\mathbf{S_2}$ and $\mathbf{S_1}$ are each spatially correlated, but that their distributions are independent of one another. In other words, we assume that $\mathbf{S_2}$ is a missing covariate but not a *confounder* of $\mathbf{S_1}$. Thus, the misspecified model that we fit is

$$\begin{aligned} g(\boldsymbol{\theta}^N; \mathbf{X}^*) &= \boldsymbol{\mu}_X\left(\boldsymbol{\theta}_1^N\right) + \boldsymbol{\Sigma}_{XX^*}\left(\boldsymbol{\theta}_2^N\right)\boldsymbol{\Sigma}_{X^*X^*}^{-1}\left(\boldsymbol{\theta}_2^N\right)\Big(\mathbf{X}^* - \boldsymbol{\mu}_{X^*}\left(\boldsymbol{\theta}_1^N\right)\Big) \\ &= \alpha_0^N\mathbf{1} + \alpha_1^N\mathbf{S_1} + \boldsymbol{\Sigma}_{XX^*}\left(\phi, \sigma^2\right)\boldsymbol{\Sigma}_{X^*X^*}^{-1}\left(\phi, \sigma^2\right)\Big(\mathbf{X}^* - \left(\alpha_0^N\mathbf{1} + \alpha_1^N\mathbf{S_1}\right)\Big) \end{aligned}$$

We now show how the error can be decomposed into Berkson and Classical components for each of these three scenarios.

2.2.3 Decomposition into Berkson and Classical error components

Let $\widehat{\boldsymbol{\theta}}$ be the vector of maximum likelihood estimates of $\boldsymbol{\theta}$. We are interested in the difference between the true unknown exposures $\mathbf{X}$ and the predicted exposures $\widehat{\mathbf{X}} = g(\widehat{\boldsymbol{\theta}}; \mathbf{X}^*)$. Since the predicted exposures are *surrogates* for the unknown true exposures, we can view this difference under a general measurement error framework. This decomposition of errors into Berkson and classical type measurement error components follows in the same spirit as Gryparis et al. (2009) and Szpiro et al. (2011), where Gryparis et al. (2009) considers a Bayesian Gaussian Process model with constant mean and Szpiro et al. (2011) considers a linear regression model. We now extend this viewpoint to our particular models of interest defined in Section 2.2.

Berkson and Classical error components for Scenarios I and II

For scenarios I and II, the predicted exposures are generated by fitting the true model given the collected monitoring data $\mathbf{X}^*$, so the predictions have the form

$$\widehat{\mathbf{X}} = g(\widehat{\boldsymbol{\theta}}; \mathbf{X}^*) = \boldsymbol{\mu}_X\big(\widehat{\boldsymbol{\theta}}_1\big) + \boldsymbol{\Sigma}_{XX^*}\big(\widehat{\boldsymbol{\theta}}_2\big)\boldsymbol{\Sigma}^{-1}_{X^*X^*}\big(\widehat{\boldsymbol{\theta}}_2\big)\Big(\mathbf{X}^* - \boldsymbol{\mu}_{X^*}(\widehat{\boldsymbol{\theta}}_1)\Big)$$

By adding and subtracting the underlying true model $g\big(\boldsymbol{\theta}; \mathbf{X}^*\big)$, we can decompose the error into Berkson and classical components,

$$\mathbf{X} - g\big(\widehat{\boldsymbol{\theta}}; \mathbf{X}^*\big) = \underbrace{\mathbf{X} - g\big(\boldsymbol{\theta}; \mathbf{X}^*\big)}_{U_b} + \underbrace{g\big(\boldsymbol{\theta}; \mathbf{X}^*\big) - g\big(\widehat{\boldsymbol{\theta}}; \mathbf{X}^*\big)}_{U_c}$$

The first term, U_b, is the Berkson error, which represents the difference between the true measurements and the true model. The second term, U_c, is the classical error, which represents the difference between the true model and the estimated model. Thus, we can call this classical error *estimation error*.

Berkson and Classical error components for Scenario III

For scenario III, the predicted exposures are generated from a misspecified model which left out a key predictor, so

$$\widehat{\mathbf{X}} = g\big(\widehat{\boldsymbol{\theta}}^N; \mathbf{X}^*\big) = \boldsymbol{\mu}_X\big(\widehat{\boldsymbol{\theta}}^N_1\big) + \boldsymbol{\Sigma}_{XX^*}\big(\widehat{\boldsymbol{\theta}}^N_2\big)\boldsymbol{\Sigma}^{-1}_{X^*X^*}\big(\widehat{\boldsymbol{\theta}}^N_2\big)\Big(\mathbf{X}^* - \boldsymbol{\mu}_{X^*}(\widehat{\boldsymbol{\theta}}^N_1)\Big)$$

By adding and subtracting the underlying true model $g\big(\boldsymbol{\theta}; \mathbf{X}^*\big)$, we can decompose the error into Berkson and classical components,

$$\mathbf{X} - g\big(\widehat{\boldsymbol{\theta}}^N; \mathbf{X}^*\big) = \underbrace{\mathbf{X} - g\big(\boldsymbol{\theta}; \mathbf{X}^*\big)}_{U_b} + \underbrace{g\big(\boldsymbol{\theta}; \mathbf{X}^*\big) - g\big(\widehat{\boldsymbol{\theta}}^N; \mathbf{X}^*\big)}_{U_c}$$

Conceptually, we can think of further decomposing this classical error as

$$\mathcal{U}_c = \underbrace{g\big(\boldsymbol{\theta};\mathbf{X}^*\big) - g\big(\boldsymbol{\theta}^N;\mathbf{X}^*\big)}_{\text{model misspecification}} + \underbrace{g\big(\boldsymbol{\theta}^N;\mathbf{X}^*\big) - g\big(\widehat{\boldsymbol{\theta}}^N;\mathbf{X}^*\big)}_{\text{estimation error}}$$

where one component of error is due purely to estimation error of the parameters $\widehat{\boldsymbol{\theta}}^N$ and the other component of error is attributed to choosing the incorrect model $g\big(\boldsymbol{\theta}^N; X^*\big)$.

2.3 Analysis of bias in health effect estimates induced by exposure models

Now that we have established the connection to the measurement error framework, it is of interest to study the impact of measurement error in the predicted exposures on bias of the coefficient β_x representing the association between air pollution exposure and the health outcome. First we investigate bias in the case of estimation error only in Scenarios I and II by analytically studying the form of the bias by using Taylor expansions. Next, we study the bias in the case of misspecification error in Scenario III by deriving the probability limits of the MLEs in the misspecified model and deriving the particular form of the classical error variance. Later, in Section 5, we conduct simulations to investigate the degree of bias in small samples.

2.3.1 Bias Analysis for Scenarios I and II

First, we note that the Berkson error component does not induce any bias in the estimate of β_x. To see this, let $\widehat{\beta}_{x,ideal}$ denote the least squares estimate of β_x if we were able to use predicted exposures from the true model $g(\boldsymbol{\theta};\mathbf{X}^*)$ where we know the true parameters exactly. Without loss of generality we assume centered variables. Specifically, let $g_i(\boldsymbol{\theta};\mathbf{X}^*)$

be the exposure for subject i and define

$$\widehat{\beta}_{X,ideal} \equiv \frac{n^{-1}\sum_{i=1}^{n} Y_i g_i(\boldsymbol{\theta};\mathbf{X}^*)}{n^{-1}\sum_{i=1}^{n}\left[g_i(\boldsymbol{\theta};\mathbf{X}^*)\right]^2}.$$

Define $\widehat{\beta}_{x,actual}$ to be the estimator using the actual exposure predictions, which are estimated by $g(\widehat{\boldsymbol{\theta}};\mathbf{X}^*)$. Specifically, define

$$\widehat{\beta}_{x,actual} \equiv \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i g_i(S_i,\widehat{\boldsymbol{\theta}};\mathbf{X}^*)}{\frac{1}{n}\sum_{i=1}^{n}\left[g_i(S_i,\widehat{\boldsymbol{\theta}};\mathbf{X}^*)\right]^2}.$$

Then, since the true model has the property $E[\mathbf{X}|\mathbf{X}^*] = g(\boldsymbol{\theta};\mathbf{X}^*)$, taking iterated expectation of $\widehat{\beta}_{X,ideal}$ conditional on $\mathbf{X}^*$ shows that the bias of $\widehat{\beta}_{X,ideal}$ is 0. Hence, any bias in our actual estimator comes from the classical error component.

Now consider the bias of our actual estimator. We have already established that

$$E\left[\widehat{\beta}_{x,actual} - \beta_x\right] = E\left[\widehat{\beta}_{x,actual} - \widehat{\beta}_{X,ideal},\right]$$

since

$$E\left[\widehat{\beta}_{x,actual} - \beta_x\right] = E\left[\widehat{\beta}_{x,actual} - \widehat{\beta}_{x,ideal} + \widehat{\beta}_{x,ideal} - \beta_x\right] = E\left[\widehat{\beta}_{x,ideal} - \beta_x\right] + E\left[\widehat{\beta}_{x,ideal} - \beta_x\right]$$

and $E\left[\widehat{\beta}_{x,ideal} - \beta_x\right] = 0$. Thus, to compute the bias we need to study the term $E\left[\widehat{\beta}_{x,actual} - \widehat{\beta}_{x,ideal}\right]$.

To derive the bias of $\widehat{\beta}_{x,actual}$, we use a Taylor expansion. We write $\widehat{\beta}_{x,ideal}$ as a function of $\mathbf{Y},\mathbf{S},\boldsymbol{\theta}$, and $\mathbf{X}^*$, which we denote by $M(\mathbf{Y},\mathbf{S},\boldsymbol{\theta};\mathbf{X}^*)$, and $\widehat{\beta}_{x,ideal} = M(\mathbf{Y},\mathbf{S},\widehat{\boldsymbol{\theta}};\mathbf{X}^*)$.

Now, a second order Taylor expansion of $M(\mathbf{Y},\mathbf{S},\widehat{\boldsymbol{\theta}};\mathbf{X}^*)$ around $M(\mathbf{Y},\mathbf{S},\boldsymbol{\theta};\mathbf{X}^*)$ is,

$$\begin{aligned}
&M(\mathbf{Y},\mathbf{S},\boldsymbol{\theta};\mathbf{X}^*) - M(\mathbf{Y},\mathbf{S},\widehat{\boldsymbol{\theta}};\mathbf{X}^*) \\
&\approx \left[\frac{\partial}{\partial\theta}M(\mathbf{Y},\mathbf{S},\boldsymbol{\theta};\mathbf{X}^*)\right]^{\mathsf{T}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + \frac{1}{2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^{\mathsf{T}}\left[\frac{\partial^2}{\partial\theta\partial\theta^{\mathsf{T}}}M(\mathbf{Y},\mathbf{S},\boldsymbol{\theta};\mathbf{X}^*)\right](\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})
\end{aligned}$$

Then the approximate bias of $\widehat{\beta}_{x,actual}$ is

$$
\begin{aligned}
&E\left\{\widehat{\beta}_{x,ideal} - \widehat{\beta}_{x,actual}\right\} \\
&= E\Big\{M(\mathbf{Y}, \mathbf{S}, \boldsymbol{\theta}; \mathbf{X}^*) - M(\mathbf{Y}, \mathbf{S}, \widehat{\boldsymbol{\theta}}; \mathbf{X}^*)\Big\} \\
&\approx E\left\{\left[\frac{\partial}{\partial\theta} M(\mathbf{Y}, \mathbf{S}, \boldsymbol{\theta}; \mathbf{X}^*)\right]^{\mathsf{T}} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + \frac{1}{2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^{\mathsf{T}} \left[\frac{\partial^2}{\partial\theta\partial\theta^{\mathsf{T}}} M(\mathbf{Y}, \mathbf{S}, \boldsymbol{\theta}; \mathbf{X}^*)\right] (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})\right\} \\
&\approx \boldsymbol{\gamma} E\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big) + \frac{1}{2}\text{trace}\Big\{\boldsymbol{\Lambda}\mathsf{Var}\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big)\Big\} + \frac{1}{2}E\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big)^{\mathsf{T}}\boldsymbol{\Lambda}E\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big)
\end{aligned}
\tag{2.4}
$$

where

$$
\boldsymbol{\gamma} = E\left[\frac{\partial}{\partial\theta} M(\mathbf{Y}, \mathbf{S}, \boldsymbol{\theta}; \mathbf{X}^*)\right]^{\mathsf{T}} \text{ and } \boldsymbol{\Lambda} = E\left[\frac{\partial^2}{\partial\theta\partial\theta^{\mathsf{T}}} M(\mathbf{Y}, \mathbf{S}, \boldsymbol{\theta}; \mathbf{X}^*)\right].
$$

Examining equation (4), we see that there are terms involving $E\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big)$ and $\mathsf{Var}\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big)$. We note that the mean model parameters will be unbiased, $E[\widehat{\boldsymbol{\theta}}_1] = \boldsymbol{\theta}_1$, however, the variance parameters will have a small-sample bias factor when fit under maximum likelihood, $E[\widehat{\boldsymbol{\theta}}_2] \neq \boldsymbol{\theta}_2$. We also note that $\boldsymbol{\gamma}$ and $\boldsymbol{\Lambda}$ are both functions of several random variables, $\mathbf{Y}, \mathbf{S}, \mathbf{X}^*$, and thus are complicated functions of vectors and matrices involving the particular distributions of those random variables. In particular, the magnitude of this bias will depend on the relative magnitudes of $E\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big)$ and $\mathsf{Var}\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big)$, as well as the distributions of the other random variables.

2.3.2 Bias Analysis for Scenario III

Derivation of the naive model coefficients, $\boldsymbol{\theta}^N$

First, we study the model misspecification component, $g\big(\boldsymbol{\theta}; X^*\big) - g^N\big(\boldsymbol{\theta}^N; X^*\big)$. We need to understand what the naive model coefficients converge to, $\boldsymbol{\theta}^N$, since they will converge to something different from the true model coefficients $\boldsymbol{\theta}$.

First, we derive the probability limits of the naive model coefficients by using the same general techniques as in Wang et al. (1998).

The naive model score equations for the exposure model fit using $\mathbf{X}^*$ have probability limits

$$\mathbb{E}\big[\mathbf{S}^\mathsf{T}\mathbf{V}_N^{-1}(\mathbf{X}^* - \mathbf{S}\boldsymbol{\theta}_{1,N})\big] = \mathbf{0} \tag{2.5}$$

$$\frac{1}{2}\left(\mathbb{E}\Big\{\big[\mathbf{X}^* - \mathbf{S}\boldsymbol{\theta}_N\big]^\mathsf{T}\mathbf{V}_N^{-1}\frac{\partial \mathbf{V}_N}{\partial \theta_{2,l,N}}\mathbf{V}_N^{-1}\big[\mathbf{X}^* - \mathbf{S}\boldsymbol{\theta}_N\big]^\mathsf{T}\Big\} - \operatorname{tr}\left\{\mathbf{V}_N^{-1}\frac{\partial \mathbf{V}_N}{\partial \theta_{2,l,N}}\right\}\right) = 0 \tag{2.6}$$

where $\mathbf{V}_N^{-1} = \Sigma_{X^*X^*}^{-1}(\boldsymbol{\theta}_{2,N})$ and $l = 1, 2$ indexes the two variance parameters.

First we derive the naive model mean parameters $\boldsymbol{\theta}_{1,N} = (\alpha_{1,N}, \alpha_{2,N}, \alpha_{3,N})$. To do that, we can solve eqn (5) for $\alpha_{1,N}, \alpha_{2,N}, \alpha_{3,N}$. Details for solving this equation are listed in the Appendix. For the mean model parameters, it is possible to derive closed-form solutions. Following the calculations given in the Appendix, the naive model mean parameters are

$$\alpha_{0,N} = \alpha_0 + \mu_{S_2}\alpha_2$$
$$\alpha_{1,N} = \alpha_1$$

Next we derive the naive model variance parameters $\boldsymbol{\theta}_{2,N} = (\phi_N, \sigma_N^2)$ by solving equation (2) for ϕ_N, σ_N^2. Here, we cannot derive closed-form solutions, but we can derive equations which can be solved numerically (see Appendix).

Now, we compare our derived naive model coefficients to the observed naive model coefficients to make sure that matches up.

2.3.3 Error due to model misspecification, $\mathcal{U}_{c,\textsf{model mis}}$

Then, once we know the form of $\boldsymbol{\theta}^N$, we can study the model misspecification difference

$$\mathcal{U}_{c,\textsf{model mis}} = g\big(\boldsymbol{\theta}; X^*\big) - g^N\big(\boldsymbol{\theta}^N; X^*\big),$$

which we now show follows a particular form

$$\mathcal{U}_{c,\textsf{model mis}} \sim MVN(\boldsymbol{\mu}_{Uc}, \boldsymbol{\Sigma}_{Uc}),$$

which we can then use for SIMEX.

To derive $\boldsymbol{\mu}_{Uc}$ and $\boldsymbol{\Sigma}_{Uc}$, the mean and variance of the classical error due to model misspecification, we first derive a simpler expression for $\mathcal{U}_{c,\textsf{model mis}}$ as follows.

$$\begin{aligned}
\mathcal{U}_{c,\textsf{model mis}} &= g\big(\boldsymbol{\theta}; X^*\big) - g^N\big(\boldsymbol{\theta}^N; X^*\big) \\
&= \alpha_0\mathbf{1} + \alpha_1\mathbf{S_1} + \alpha_2\mathbf{S_2} + \boldsymbol{\Sigma}_{XX^*}(\phi,\sigma^2)\boldsymbol{\Sigma}^{-1}_{X^*X^*}(\phi,\sigma^2)\Big(\mathbf{X}^* - \big(\alpha_0\mathbf{1} + \alpha_1\mathbf{S_1} + \alpha_2\mathbf{S_2}\big)\Big) \\
&\quad - \Big[\alpha_0^N\mathbf{1} + \alpha_1^N\mathbf{S_1} + \boldsymbol{\Sigma}_{XX^*}\left(\phi^N,\sigma^{2,N}\right)\boldsymbol{\Sigma}^{-1}_{X^*X^*}\left(\phi^N,\sigma^{2,N}\right)\Big(\mathbf{X}^* - \big(\alpha_0^N\mathbf{1} + \alpha_1^N\mathbf{S_1}\big)\Big)\Big] \\
&= \alpha_0\mathbf{1} + \alpha_1\mathbf{S_1} + \alpha_2\mathbf{S_2} - \left[\alpha_0^N\mathbf{1} + \alpha_1^N\mathbf{S_1}\right] \\
&\quad + \boldsymbol{\Sigma}_{XX^*}(\phi,\sigma^2)\boldsymbol{\Sigma}^{-1}_{X^*X^*}(\phi,\sigma^2)\Big(\mathbf{X}^* - \big(\alpha_0\mathbf{1} + \alpha_1\mathbf{S_1} + \alpha_2\mathbf{S_2}\big)\Big) \\
&\quad - \boldsymbol{\Sigma}_{XX^*}\left(\phi^N,\sigma^{2,N}\right)\boldsymbol{\Sigma}^{-1}_{X^*X^*}\left(\phi^N,\sigma^{2,N}\right)\Big(\mathbf{X}^* - \big(\alpha_0^N\mathbf{1} + \alpha_1^N\mathbf{S_1}\big)\Big) \\
&= \alpha_2\left(\mathbf{S_2} - \mu_{S_2}\mathbf{1}\right) + \boldsymbol{\Sigma}_{XX^*}(\phi,\sigma^2)\boldsymbol{\Sigma}^{-1}_{X^*X^*}(\phi,\sigma^2)\boldsymbol{\epsilon}^* \\
&\quad - \boldsymbol{\Sigma}_{XX^*}\left(\phi^N,\sigma^{2,N}\right)\boldsymbol{\Sigma}^{-1}_{X^*X^*}\left(\phi^N,\sigma^{2,N}\right)\Big(\boldsymbol{\epsilon}^* + \alpha_2\left(\mathbf{S_2} - \mu_{S_2}\mathbf{1}\right)\Big) \\
&= \alpha_2\left(\mathbf{S_2} - \mu_{S_2}\mathbf{1}\right) - \boldsymbol{\Sigma}_{XX^*}\left(\phi^N,\sigma^{2,N}\right)\boldsymbol{\Sigma}^{-1}_{X^*X^*}\left(\phi^N,\sigma^{2,N}\right)\Big(\alpha_2\left(\mathbf{S_2} - \mu_{S_2}\mathbf{1}\right)\Big) \\
&\quad + \Big[\boldsymbol{\Sigma}_{XX^*}(\phi,\sigma^2)\boldsymbol{\Sigma}^{-1}_{X^*X^*}(\phi,\sigma^2) - \boldsymbol{\Sigma}_{XX^*}\left(\phi^N,\sigma^{2,N}\right)\boldsymbol{\Sigma}^{-1}_{X^*X^*}\left(\phi^N,\sigma^{2,N}\right)\Big]\boldsymbol{\epsilon}^*
\end{aligned}$$

Now using this expression, $\mathcal{U}_{c,\textsf{model mis}}$ is a function of two random vectors, $(\mathbf{S_2}, \mathbf{S_2}^*)$ and $\boldsymbol{\epsilon}^*$. First, $\boldsymbol{\mu}_{Uc} = \mathbf{0}$ because $E[(\mathbf{S_2} - \mu_{S_2}\mathbf{1})] = \mathbf{0}$ and $E[\boldsymbol{\epsilon}^*] = \mathbf{0}$. Then, the variance of $\mathcal{U}_{c,\textsf{model mis}}$ is

$$\boldsymbol{\Sigma}_{Uc} = \alpha_2^2\boldsymbol{\Sigma}_{S_2S_2} - \alpha_2\boldsymbol{\Sigma}_{S_2S_2^*}\mathbf{A}^\intercal - \alpha_2\mathbf{A}\boldsymbol{\Sigma}_{S_2^*S_2} + \mathbf{B}\boldsymbol{\Sigma}_{X^*X^*}\mathbf{B}^\intercal + \mathbf{A}\boldsymbol{\Sigma}_{S_2^*S_2^*}\mathbf{A}^\intercal$$

where

$$\mathbf{A} = \Big[\boldsymbol{\Sigma}_{XX^*}(\phi,\sigma^2)\boldsymbol{\Sigma}^{-1}_{X^*X^*}(\phi,\sigma^2) - \boldsymbol{\Sigma}_{XX^*}\left(\phi^N,\sigma^{2,N}\right)\boldsymbol{\Sigma}^{-1}_{X^*X^*}\left(\phi^N,\sigma^{2,N}\right)\Big]$$

and

$$\mathbf{B} = \mathbf{\Sigma}_{XX^*}\left(\phi^N, \sigma^{2,N}\right)\mathbf{\Sigma}_{X^*X^*}^{-1}\left(\phi^N, \sigma^{2,N}\right).$$

Now that we have $\mathcal{U}_{c,\text{model mis}} \sim MVN(\mathbf{0}, \mathbf{\Sigma}_{Uc})$, we can simulate remeasurement from this distribution for SIMEX.

2.4 SIMEX for correlated Berkson and Classical Errors

The SIMEX method was introduced by Cook and Stefanski (1994) and the method provides an approximate general correction for the effect of classical measurement errors on the estimation of a parameter of interest. SIMEX has two steps: a simulation step (SIM) where simulated error is added to the mismeasured exposures in increasing amounts, and an extrapolation step (EX) where a trend is fit to the distribution of parameter estimates over the increasing error levels and extrapolated back to the case of no error.

Briefly, suppose we observe an exposure with classical measurement error, $W_i = X_i + \mathcal{U}_{c,i}$ where
$\mathcal{U}_{c,i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_c^2)$. Then in the simulation step, for each λ we can generate psuedo-datasets $r = 1, \ldots, R$, such that

$$W_i^{(r)}(\lambda) = W_i + \sqrt{\lambda}\mathcal{U}_{c,i}^{(r)}, \text{ where } \mathcal{U}_{c,i}^{(r)} \overset{iid}{\sim} N(0, \sigma_u^2)$$

Thus,

$$E\left[W_i^{(r)}(\lambda)|X_i\right] = E[W_i|X_i] + \sqrt{\lambda}E[\mathcal{U}_{c,i}] = X_i$$

and

$$Var\left[W_i^{(r)}(\lambda)|X_i\right] = Var[W_i|X_i] + \lambda Var[\mathcal{U}_{c,i}] = (1+\lambda)\sigma_u^2$$

The key requirement of SIMEX is that the mean squared error $MSE[\mathbf{W}(\lambda)|\mathbf{X}] = E[\mathbf{W}(\lambda)|\mathbf{X}] \rightarrow$

0 as $\lambda \to -1$, which enables us to extrapolate to the case of no error. The conditions that $Var[W_i^{(r)}(\lambda)|X_i] \to 0$ and $E[W_i^{(r)}(\lambda)|X_i] \to X_i$ ensure the condition that the MSE converges to 0.

Once we have simulated the B remeasured pseudo-datasets for each λ, we estimate the parameter of interest for each dataset and average over the B simulations to get an estimate $\widehat{\theta}(\lambda)$. We then fit a trend to $\widehat{\theta}(\lambda)$ versus λ, typically using a linear or quadratic model. The predicted value of this trend at $\lambda = -1$ is our estimate of the parameter θ under no measurement error.

Mixtures of Berkson and classical errors in the independent identically distributed case are introduced briefly in Carroll et al. (2006), and have also been considered in several papers under a Bayesian framework.(Mallick et al., 2002; Li et al., 2007) In the uncorrelated case, both the Berkson errors and the Classical errors are assumed to be independent and identically distributed by the Normal distribution, with variances σ^2_{Ub} and σ^2_{Uc} respectively.

One could assume that the error variances of σ^2_b, σ^2_c are known, but typically in measurement error problems, external validation data with measurements of both the true exposure and the mismeasured exposure are used to estimate measurement error variances. One key issue which has been discussed in this setup is the nonidentifiability of the amount of uncertainty that is Berkson versus classical.(Mallick et al., 2002; Li et al., 2007) While external validation data would allow the estimation of the to total error variance, the relative proportions of Berkson and classical errors cannot be determined. In those studies, the authors perform sensitivity analyses regarding the percentage of variance assumed to be classical versus Berkson to deal with the identifiability issue. A SIMEX approach to the iid Normal verson of this shared uncertainty problem is briefly introduced in Carroll et al. (2006), but has not been well-studied. Schafer et al. (1999) considers SIMEX for this uncorrelated case, but mainly focus on separately considering the cases of classical errors and Berkson errors.

We now present an approach to extend the current SIMEX methodology to the case when the exposures are correlated over space, and the Berkson and classical errors are correlated over space as well.

2.4.1 SIMEX for mixtures of correlated classical and Berkson errors

Suppose we consider the latent variable $\mathcal{L}$ to be a suitable surrogate for X, since using $\mathcal{L}$ would provide an unbiased estimate of X by Berkson error properties. Then we can require that as $\lambda \to -1$, $E[W_i^{(r)}(\lambda)|\mathcal{L}_i] \to \mathcal{L}_i$ and $Var[W_i^{(r)}(\lambda)|\mathcal{L}_i] \to 0$. Under this scenario, we can simplify the re-measurement simulation as follows,

$$W_i^{(r)}(\lambda) = W_i + \sqrt{\lambda}\mathcal{U}_{c,i}^{(r)}$$

where $U_c \sim MVN\big(0, \Sigma_c\big)$. This allows the classical errors to be correlated or uncorrelated and assumes as usual that Σ_c is known. This approach is thus analogous to the classical-only error SIMEX approach.

2.5 Simulation Study

2.5.1 Bias Simulations for Scenario I and II

We conducted a simulation study to explore the degree of bias for different numbers of monitors. In scenario I, we assume a constant mean model, so all of the exposure variability comes from the spatially correlated residuals. First, we consider a smooth surface which would satisfy the smoothness conditions needed in our Taylor expansion. For our Matern covariance function, we chose a smoothness of $\kappa = 3$, and we chose covariance parameters $\phi = 0.2$ for the range and $\sigma^2 = 0.5$ for the variance. An example of a surface is shown in Figure 2.1. The results for this smooth exposure surface for Scenario I are given in Table 2.1. Even for small numbers of monitors, we see no bias.

Next, we consider a rough surface which would not satisfy the smoothness conditions

Table 2.1: Simulation results for smooth and rough exposure surfaces for Scenarios I and II with different number of monitors m

Smooth Surface								Rough Surface							
S	m	Exposure	Bias	empir SE	model SE	MSE	Coverage	S	m	Exposure	Bias	empir SE	model SE	MSE	Coverage
I	20	True X	0.001	0.082	0.076	0.007	94.200	I	20	True X	-0.001	0.057	0.055	0.003	93.865
I	20	$g(\theta; X)$	-0.003	0.108	0.078	0.012	86.400	I	20	$g(\theta; X)$	-0.007	0.179	0.080	0.032	63.190
I	20	$g(\hat{\theta}; X)$	0.004	0.111	0.080	0.012	85.200	I	20	$g(\hat{\theta}; X)$	0.058	0.220	0.089	0.052	57.055
I	40	True X	-0.000	0.086	0.076	0.007	94.000	I	40	True X	-0.000	0.057	0.054	0.003	94.990
I	40	$g(\theta; X)$	0.001	0.088	0.076	0.008	93.000	I	40	$g(\theta; X)$	0.002	0.104	0.067	0.011	80.962
I	40	$g(\hat{\theta}; X)$	0.002	0.088	0.076	0.008	93.000	I	40	$g(\hat{\theta}; X)$	0.024	0.114	0.070	0.014	78.557
II	20	True X	0.001	0.044	0.043	0.002	94.990	II	20	True X	0.002	0.038	0.038	0.001	94.400
II	20	$g(\theta; X)$	0.002	0.055	0.044	0.003	89.379	II	20	$g(\theta; X)$	0.002	0.040	0.038	0.002	93.600
II	20	$g(\hat{\theta}; X)$	0.002	0.059	0.044	0.003	87.976	II	20	$g(\hat{\theta}; X)$	0.002	0.042	0.038	0.002	92.200
II	40	True X	0.002	0.043	0.043	0.002	94.400	II	40	True X	0.002	0.038	0.038	0.001	95.000
II	40	$g(\theta; X)$	0.003	0.044	0.043	0.002	93.600	II	40	$g(\theta; X)$	0.002	0.038	0.038	0.001	95.000
II	40	$g(\hat{\theta}; X)$	0.003	0.045	0.043	0.002	93.800	II	40	$g(\hat{\theta}; X)$	0.002	0.038	0.038	0.001	94.600

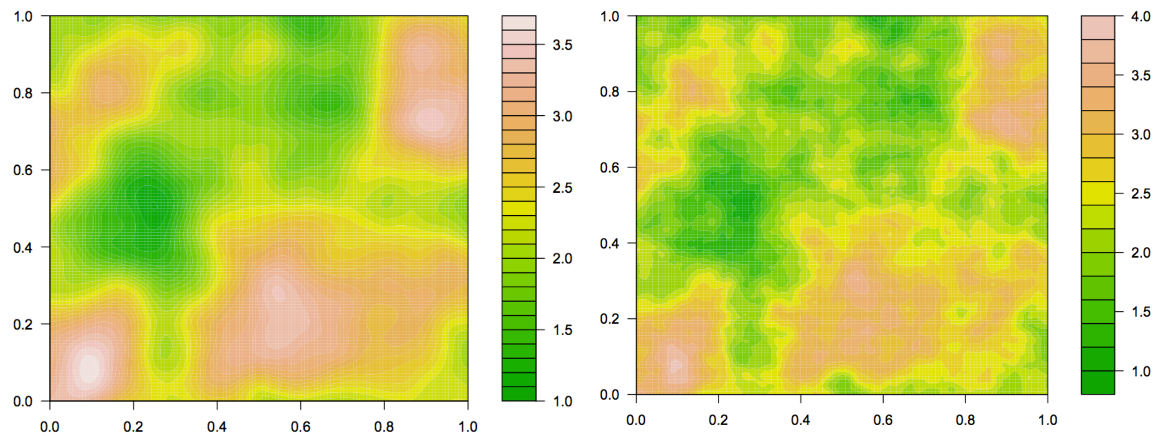

Figure 2.1: Example of smooth surface and rough exposure surface with Matern covariance. This realization is on a (0,1) grid with range $\phi = 0.2$ and variance $\sigma^2 = 0.5$. The smoother surface (left) has $\kappa = 3$ and the rougher surface (right) has $\kappa = 1$.

needed in our Taylor expansion. For our Matern covariance function, we chose a smoothness of $\kappa = 1$, and we chose covariance parameters $\phi = 0.2$ for the range and $\sigma^2 = 0.5$ for the variance. An example of a surface is shown in Figure 2.1. The results for this rough exposure surface for Scenario I are given in Table 2.2.

This is a particularly difficult case to estimate, because we use a small number of monitors and we have a rough surface. We see a small amount of bias, particularly in the case of only 20 or 30 monitors. Interestingly, the direction of this bias here is upward. This may be an interesting artifact of using a particularly sparse dataset and a rough surface which is difficult to estimate.

In Scenario II, we consider a scenario where the spatial covariates have rougher surfaces and the residual variation is smoother. For our Matern covariance function of the residuals, we chose a smoothness of $\kappa = 3$, and we chose covariance parameters $\phi = 0.2$ for the range and $\sigma^2 = 0.5$ for the variance. The results for this smooth exposure surface for Scenario I are given in Table 2.3.

Table 2.2: Simulation results for for Scenario III, misspecified model and correction by spatial SIMEX with different number of monitors m

Scenario	m	Exposure	Bias	SD	modelSE	MSE	Coverage
III	50	True	0.000	0.036	0.034	0.001	93.6
III	50	$g(\theta; X)$	0.000	0.036	0.034	0.001	93.6
III	50	$g(\hat{\theta}^N; X)$	-0.203	0.182	0.039	0.074	22.2
III	50	spatial SIMEX, linear	-0.072	0.211	0.241	0.050	90.6
III	50	spatial SIMEX, quad	0.026	0.254	0.285	0.065	91.9

2.5.2 Simulations for Bias in Scenario III and spatial SIMEX correction

We conducted a simulation study to evaluate performance of the SIMEX correction method for correlated Berkson and classical errors. We see that the bias was approximately corrected. In the supplementary material, we show in Table 2.1 that the naive model parameters in the exposure model converged approximately to the derivations in Section 3.

Examples of the SIMEX extrapolations for several simulations are shown in a Figure 2.1 in the Supplementary material.

Standard Error estimation when Σ_c is known

Using the true Σ_c classical variance from our derivations, we implemented our spatial SIMEX procedure with a bootstrap standard error estimate. To implement the bootstrap standard error, for each simulation i, start by implementing SIMEX and extrapolate to $\hat{\beta}_{\text{SIMEX},i}$. Then, for $k = 1, ..., 200$ bootstrap samples: (1) resample monitor data, (2) fit the initial exposure model to the monitoring data (3) then repeat the entire SIMEX procedure using these new predictions. The s.e. estimate for $\hat{\beta}_{\text{SIMEX},i}$, is then computed by the standard deviation of the 200 bootstrap estimates $\hat{\beta}^{(k)}_{\text{SIMEX},i}$. Table 2.2 shows the results of the implementation using this standard error estimation.

Table 2.3: Simulation results for for Scenario III, misspecified model and correction by spatial SIMEX with crude Σ_c estimation

Scenario	p	Exposure	Bias	SD	modelSE	MSE	Coverage
Misspec I	1.00	beta.x.truex	0.001	0.036	0.034	0.001	93.600
Misspec I	1.00	beta.x.g	0.001	0.036	0.034	0.001	93.600
Misspec I	1.00	beta.x.gNhat	-0.200	0.180	0.039	0.072	22.400
Misspec I	1.00	beta.x.SIMEX.linear	-0.068	0.213	0.228	0.050	91.260
Misspec I	1.00	beta.x.SIMEX.quad	0.067	0.322	0.336	0.108	87.195
Misspec I	0.90	beta.x.SIMEX.linear	-0.075	0.211	0.228	0.050	91.057
Misspec I	0.90	beta.x.SIMEX.quad	0.044	0.308	0.331	0.097	87.805
Misspec I	0.80	beta.x.SIMEX.linear	-0.076	0.208	0.227	0.049	91.572
Misspec I	0.80	beta.x.SIMEX.quad	0.028	0.294	0.324	0.087	89.066
Misspec I	0.50	beta.x.SIMEX.linear	-0.112	0.200	0.225	0.053	89.634
Misspec I	0.50	beta.x.SIMEX.quad	-0.058	0.247	0.304	0.064	91.260

Simulations using crude Σ_c estimation

We conducted a simulation study to evaluate performance of the SIMEX correction method when Σ_c is estimated by cross-validation on held-out monitors. For a set of held-out monitors, we used the remaining monitors to fit the same misspecified spatial model, and then predicted the measurement at the held-out monitors. We then fit a spatial model to the set of m held-out residuals and used that estimate as our $\hat{\Sigma}_{total}$. We then used this estimate for $\hat{\Sigma}_c$ in SIMEX, either directly or by multiplying $\hat{\Sigma}_c = p\hat{\Sigma}_{total}$ for different proportions of Classical to Berkson error. We implemented 1-fold, 2-fold, 5-fold, and leave-one-out CV. Increasing the fold of the CV seemed to make the spatial component of the residuals hard to estimate, hence we used 1-fold CV.

2.6 Data Example: Association between air pollution and low birthweight

We applied our spatial SIMEX method to a study of birthweight and particulate matter exposure during pregnancy in Massachusetts. The objective of the study was to estimate the association between birthweight and $PM_{2.5}$ exposure during the second and third trimesters. The study population included all singleton live births in Massachusetts from the Massachusetts Birth Registry during 2008 (January 1 to December 31), a total of 70,340 births. The residential address of each mother at time of birth was geocoded as described in Kloog et al. (2012).

$PM_{2.5}$ measurements during 2007 and 2008 were obtained from 40 monitoring sites in Massachusetts as part of the EPA (Environmental Protection Agency) and IMPROVE (Interagency Monitoring of Protected Visual Environments) monitoring networks.(Kloog et al., 2011)

Individual-level data on the mother and baby were obtained through the Massachusetts Birth Registry. Confounders included in the health model were maternal age, gestational age, number of cigarettes smoked during and before pregnancy, chronic conditions of mother or conditions of pregnancy (lung disease, hypertension, gestational diabetes and non-gestational diabetes), and socioeconomic measures (mother's race, mother's years of education, and the Kotelchuck index of adequacy of prenatal care utilization (APNCU). Area-level socioeconomic status was controlled by census-tract median household income using data from the United States Census Bureau of 2000 for each census tract in Massachusetts.

A universal kriging model was assumed, with a Matern covariance structure for the residuals. The mean function for the kriging model included a linear trend for land use covariates: distance to level A1 road (primary highway with limited access), distance to known particulate matter emission source, and average traffic density. Details of these land use covariates have been described previously.(Kloog et al., 2011) These kriging models were fit to the monthly average $PM_{2.5}$ concentrations at the monitoring sites during 2007 and 2008 and kriging predictions were estimated at the address of each birth. Exposure during the third trimester of pregnancy was estimated by averaging the monthly $PM_{2.5}$ concentrations over the 90 days prior to the delivery date. Exposure during the second trimester of pregnancy was estimated by averaging the monthly $PM_{2.5}$ concentrations over the days 91 to 180 prior to the delivery date.

We fit separate linear health effect models using predicted exposure to $PM_{2.5}$ for the second trimester and third trimester adjusting for confounders. Without correcting for measurement error, we found negative associations between birthweight and second trimester exposure to $PM_{2.5}$ and third trimester exposure to $PM_{2.5}$. The estimated association with birthweight, without correcting for measurement error, was -5.04 grams per 1 ug/m3 for predicted second trimester $PM_{2.5}$ exposure, 95% confidence interval $(-8.02, -2.05)$. The estimated association with birthweight was -3.49 grams for a 1 ug/m3 increase in average $PM_{2.5}$ exposure during the third trimester, $95\% CI : (-6.08, -0.89)$.

We applied our proposed spatial SIMEX correction method to this data, using 50 SIMEX remeasurement steps and 50 bootstrap resampling steps (within which 50 SIMEX remeasurement steps were needed). We used a quadratic extrapolation function and assumed that 80% of the correlated error was classical.

The estimated association between $PM_{2.5}$ and birthweight when corrected by spatial SIMEX was -8.06 grams per 1 ug/m3 for predicted second trimester $PM_{2.5}$ exposure with 95% confidence interval $(-8.66, -7.45)$. The estimated association between $PM_{2.5}$ and birthweight when corrected by spatial SIMEX was -5.07 grams for a 1 ug/m3 increase in average $PM_{2.5}$ exposure during the third trimester, $95\% CI : (-6.54, -3.61)$. Figure 2.3 in the Supplement shows both the quadratic and linear extrapolations for the second trimester birthweight exposure.

2.7 Discussion and Conclusions

In this paper, we have conducted a bias analysis of several key scenarios in exposure modeling of air pollution. We have shown that when the exposure model is misspecified by omitting an important covariate, that can induce notable downward bias. We have proposed a new spatial SIMEX approach to adjust for bias in the presence of model misspecification. We demonstrated that this bias due to exposure model misspecification can be approximately corrected by this spatial SIMEX procedure. We have also shown that in the case of a correctly specified exposure model, the degree of bias is typically negligible. Hence, this work has demonstrated that with respect to bias, model misspecification is a much bigger problem than parameter estimation.

Previous research in this area has suggested that using the plug-in estimator typically induces little bias, and authors have advocated for using the plug-in estimator to estimate the effect size and then adjusting the standard errors to account for the additional variability in using the exposure predictions.(Szpiro et al., 2011; Madsen et al., 2008; Lopiano et al., 2011; Gryparis et al., 2009) However, those papers investigate bias in simulation studies only by fitting the correct exposure model used to generate the data. Our findings in Section 3 for the bias of exposure model scenarios I and II are consistent with these previous studies, as we also found in simulations that the degree of bias is small when the correct exposure model is specified. In practice, however, the underlying exposure model that generates air pollution levels in any given region is not exactly known. In addition, current approaches for correcting the standard errors of estimates also rely on the assumption that the exposure model is correctly specified.(Szpiro et al., 2011; Madsen et al., 2008) We have approached this problem using both analytical methods and simulation studies, and we have presented a much more thorough bias analysis than what has been considered in previous work. In particular, we have extended to the case of model misspecification, which is important since exposure models are not believed to be perfect.

This work also points to a few practical considerations which are important in order to help with the implementation of this spatial SIMEX method in a realistic setting. As in other SIMEX procedures, the classical error variance is needed to generate the simulated re-measurements for the bias correction, and that variance is assumed to be known. In Section 3, we derived the particular form of the classical error variance for scenario III, but in practice the exact classical error variance would not be known and finding a way to estimate that variance may be difficult. The other key practical consideration in the estimation of the classical error variance is the nonidentifiability issue created by the mixture of Berkson and classical errors. Typically in measurement error problems, external validation data with measurements of both the true exposure and the mismeasured exposure are used to estimate the measurement error variance. However, previous studies looking at mixtures of independent and identically distributed Berkson and classical errors have noted that the amount of uncertainty that is Berkson versus classical is not identifiable.(Mallick et al., 2002; Li et al., 2007) While external validation data would allow the estimation of the total error variance, the relative proportions of Berkson and classical errors cannot be determined. In Li et al. (2007), the authors perform sensitivity analyses by considering a range of values for the percentage of variance assumed to be classical versus Berkson to deal with the identifiability issue.

Our work in this paper points to a number of areas of future research which would be of interest to advance the understanding of the impacts of model uncertainty in air pollution exposure estimation and enable straightforward implementation of measurement error correction methods in epidemiological studies of air pollution and health. First, our work suggests a need to further examine issues of model misspecification in land use regression and kriging models for air pollution exposure. We have considered one scenario of how model misspecification could arise, and we saw that notable bias was induced in that case. Many other scenarios of model misspecification may also arise, and thus it is important to consider the sources of model uncertainty and study how those scenarios may induce bias. Second, it would be helpful to study the practical issues of the implementation, including methods for estimating the classical error variance given the issue of identifiability, as mentioned above, as well as the robustness of spatial SIMEX to incorrect estimation of the classical error variance.

This work examines aspects of exposure modeling of air pollution for health effect studies and provides some insight into what are the key sources of model uncertainty which may induce bias in the estimation of health effects. Understanding the impacts of model uncertainty when constructing land use regression and kriging models is of fundamental importance to studies of air pollution and health. In particular, these land use regression and kriging models are becoming widely used in the area of air pollution exposure modeling, and the results of these epidemiology studies are evaluated by regulatory agencies and ultimately used to inform policy decisions. If air pollution epidemiology studies employ procedures which bias the effect estimates in a particular direction, then it is important to correct these biases to inform correct scientific conclusions about the effects of air pollution.

2.8 References

M. Brauer, G. Hoek, P. van Vliet, K. Meliefste, P. Fischer, U. Gehring, J. Heinrich, J. Cyrys, T. Bellander, M. Lewne, and B. Brunekreef. Estimating long-term average particulate air pollution concentrations: application of traffic indicators and geographic information systems. *Epidemiology*, 14(2):228239, 2003.

R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman & Hall, New York, New York, 2nd edition, 2006.

J. Clougherty, R. Wright, L. Baxter, and J. Levy. Land use regression modeling of intra-urban residential variability in multiple traffic-related air pollutants. *Environ Health*, 7 (17), 2008.

J. R. Cook and L. A. Stefanski. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89:13141328, 1994.

A. Gryparis, C. J. Paciorek, A. Zeka, J. Schwartz, and B. A. Coull. Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics*, 10:258–274, 2009.

I. Kloog, P. Koutrakis, B. Coull, H. Lee, and J. Schwartz. Temporally and spatially resolved pm2. 5 exposures for epidemiological studies using satellite aerosol optical depth measurements. *Atmospheric Environment*, 45:6267–6275, 2011.

I. Kloog, S. J. Melly, W. L. Ridgway, B. A. Coull, and J. Schwartz. Using new satellite based exposure methods to study the association between pregnancy pm2.5 exposure, premature birth and birth weight in massachusetts. *Environmental Health*, 11:40, 2012.

Y. Li, A. Guolo, O. Hoffman, and R. Carroll. Shared uncertainty in measurement error problems, with application to nevada test site fallout data. *Biometrics*, 63:1226–1236, 2007.

K. Lopiano, L. Young, and C. Gotway. A comparison of errors in variables methods for use in regression models with spatially misaligned data. *Statistical Methods in Medical Research*, 20:29–47, 2011.

L. Madsen, D. Ruppert, and N. Altman. Regression with spatially misaligned data. *Environmetrics*, 19:453–467, 2008.

B. Mallick, O. Hoffman, and R. Carroll. Semiparametric regression modeling with mixtures of berkson and classical error, with application to fallout from the nevada test site. *Biometrics*, 58:13–20, 2002.

J. Moller, editor. *Spatial Statistics and Computational Methods*, volume 173 of *Lecture Notes in Statistics*. Springer-Verlag, New York, New York, 2003.

D. Schafer, L. Stefanski, and R. J. Carroll. Consideration of measurement errors in the international radiation study of cervical cancer, in uncertainties in radiation dosimetry and their impact on dose-response analyses. *National Institutes of Health Publication*, 1999.

A. Szpiro, L. Sheppard, and T. Lumley. Efficient measurement error correction with spatially misaligned data. *Biostatistics*, 12:610–623, 2011.

N. Wang, X. Lin, R. Gutierrez, and R. Carroll. Bias analysis and simex approach in generalized linear mixed measurement error models. *Journal of the American Statistical Association*, 93:249–261, 1998.

2.A Appendix. Details of naive model parameter derivations

2.A.1 Derivation of naive model mean parameters in scenario III

Here are the details to solve equation (5) for the naive model mean parameters $\boldsymbol{\theta}_{1,N} = (\alpha_{0,N}, \alpha_{1,N})$. We first rewrite equation (5) as

$$
\begin{aligned}
\mathbb{E}\big[\mathbf{S}_N^{\mathsf{T}}\mathbf{V}_N^{-1}(\mathbf{X}^* - \mathbf{S}_N\boldsymbol{\theta}_{1,N})\big] &= 0 \\
\Leftrightarrow\ \mathbb{E}\big[\mathbf{S}_N^{\mathsf{T}}\mathbf{V}_N^{-1}\mathbf{X}^*\big] &= \mathbb{E}\big[\mathbf{S}_N^{\mathsf{T}}\mathbf{V}_N^{-1}\mathbf{S}_N\boldsymbol{\theta}_{1,N}\big] \\
\Leftrightarrow\ \mathbb{E}\big[\mathbb{E}\{\mathbf{S}_N^{\mathsf{T}}\mathbf{V}_N^{-1}\mathbf{X}^*|\mathbf{S}\}\big] &= \mathbb{E}\big[\mathbf{S}_N^{\mathsf{T}}\mathbf{V}_N^{-1}\mathbf{S}_N\boldsymbol{\theta}_{1,N}\big] \\
\Leftrightarrow\ \mathbb{E}\big[\mathbf{S}_N^{\mathsf{T}}\mathbf{V}_N^{-1}\mathbf{S}\big]\boldsymbol{\theta}_1 &= \mathbb{E}\big[\mathbf{S}_N^{\mathsf{T}}\mathbf{V}_N^{-1}\mathbf{S}_N\big]\boldsymbol{\theta}_{1,N}
\end{aligned}
$$

Written in matrix notation, the last line is

$$
\mathbb{E}\begin{pmatrix} \mathbf{1}^{\mathsf{T}}\{\mathbf{V}_N^{-1}\}\mathbf{1} & \mathbf{1}^{\mathsf{T}}\{\mathbf{V}_N^{-1}\}\mathbf{S}_1 & \mathbf{1}^{\mathsf{T}}\{\mathbf{V}_N^{-1}\}\mathbf{S}_2 \\ \mathbf{S}_1^{\mathsf{T}}\{\mathbf{V}_N^{-1}\}\mathbf{1} & \mathbf{S}_1^{\mathsf{T}}\{\mathbf{V}_N^{-1}\}\mathbf{S}_1 & \mathbf{S}_1^{\mathsf{T}}\{\mathbf{V}_N^{-1}\}\mathbf{S}_2 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix}
$$

$$
= \mathbb{E}\begin{pmatrix} \mathbf{1}^{\mathsf{T}}\{\mathbf{V}_N^{-1}\}\mathbf{1} & \mathbf{1}^{\mathsf{T}}\{\mathbf{V}_N^{-1}\}\mathbf{S}_1 \\ \mathbf{S}_1^{\mathsf{T}}\{\mathbf{V}_N^{-1}\}\mathbf{1} & \mathbf{S}_1^{\mathsf{T}}\{\mathbf{V}_N^{-1}\}\mathbf{S}_1 \end{pmatrix} \begin{pmatrix} \alpha_{0,N} \\ \alpha_{1,N} \end{pmatrix}
$$

Now this is a set of two equations and two unknowns,

$$
\begin{aligned}
a\alpha_0 + b\alpha_1 + c\alpha_2 &= a\alpha_{0,N} + b\alpha_{1,N} \\
b\alpha_0 + d\alpha_1 + e\alpha_2 &= b\alpha_{0,N} + d\alpha_{1,N}
\end{aligned}
$$

where a, b, c, d, e are defined to be the following expectations

$$
\begin{aligned}
a &= \mathbb{E}[\mathbf{1}^\mathsf{T}\{\mathbf{V}_N^{-1}\}\mathbf{1}] = \mathrm{tr}\{\mathbf{V}_N^{-1}\} \\
b &= \mathbb{E}[\mathbf{1}^\mathsf{T}\{\mathbf{V}_N^{-1}\}\mathbf{S}_1] = \mu_{S_1}\mathrm{tr}\{\mathbf{V}_N^{-1}\} \\
c &= \mathbb{E}[\mathbf{1}^\mathsf{T}\{\mathbf{V}_N^{-1}\}\mathbf{S}_2] = \mu_{S_2}\mathrm{tr}\{\mathbf{V}_N^{-1}\} \\
d &= \mathbb{E}[\mathbf{S}_1^\mathsf{T}\{\mathbf{V}_N^{-1}\}\mathbf{S}_1] = \mathrm{tr}\{\mathbf{V}_N^{-1}\Sigma_{S_1}\} + \mu_{S_1}^2\mathrm{tr}\{\mathbf{V}_N^{-1}\} \\
e &= \mathbb{E}\big[\mathbf{S}_1^\mathsf{T}\{\mathbf{V}_N^{-1}\}\mathbf{S}_2] = \mathbb{E}\big[\mathbb{E}\big[\mathbf{S}_1^\mathsf{T}\{\mathbf{V}_N^{-1}\}\mathbf{S}_2|\mathbf{S}_2\big]\big] = \mu_{S_1}\mathbf{1}^\mathsf{T}\mathbb{E}\big[\{\mathbf{V}_N^{-1}\}\mathbf{S}_2\big] = \mu_{S_1}\mu_{S_2}\mathrm{tr}\{\mathbf{V}_N^{-1}\}
\end{aligned}
$$

(Note that for term e, we use the property that $\mathbf{S}_1$ and $\mathbf{S}_2$ are independent. Thus, if $\mathbf{S}_2$ were actually a confounder, this step would be different because we would need the conditional expecation $\mathbb{E}[\mathbf{S}_1|\mathbf{S}_2]$.)

Rearranging terms to solve for $\alpha_{0,N}$ and $\alpha_{1,N}$,

$$
\begin{aligned}
\alpha_{0,N} &= \alpha_0 + \frac{cd - be}{ad - b^2}\cdot\alpha_2 \\
\alpha_{1,N} &= \alpha_1 + \frac{bc - ae}{b^2 - ad}\cdot\alpha_2
\end{aligned}
$$

These terms reduce

$$
\begin{aligned}
\frac{cd - be}{ad - b^2} &= \frac{\mu_{S_2}\mathrm{tr}\{\mathbf{V}_N^{-1}\}\big[\mathrm{tr}\{\mathbf{V}_N^{-1}\Sigma_{S_1}\} + \mu_{S_1}^2\mathrm{tr}\{\mathbf{V}_N^{-1}\}\big] - \mu_{S_1}\mathrm{tr}\{\mathbf{V}_N^{-1}\}\mu_{S_1}\mu_{S_2}\mathrm{tr}\{\mathbf{V}_N^{-1}\}}{\mathrm{tr}\{\mathbf{V}_N^{-1}\}\big[\mathrm{tr}\{\mathbf{V}_N^{-1}\Sigma_{S_1}\} + \mu_{S_1}^2\mathrm{tr}\{\mathbf{V}_N^{-1}\}\big] - \mu_{S_1}^2\mathrm{tr}\{\mathbf{V}_N^{-1}\}^2} = \mu_{S_2} \\
\frac{bc - ae}{b^2 - ad} &= \frac{\mu_{S_1}\mathrm{tr}\{\mathbf{V}_N^{-1}\}\mu_{S_2}\mathrm{tr}\{\mathbf{V}_N^{-1}\} - \mathrm{tr}\{\mathbf{V}_N^{-1}\}\mu_{S_1}\mu_{S_2}\mathrm{tr}\{\mathbf{V}_N^{-1}\}}{\mu_{S_1}^2\mathrm{tr}\{\mathbf{V}_N^{-1}\}^2 - \mathrm{tr}\{\mathbf{V}_N^{-1}\}\big[\mathrm{tr}\{\mathbf{V}_N^{-1}\Sigma_{S_1}\} + \mu_{S_1}^2\mathrm{tr}\{\mathbf{V}_N^{-1}\}\big]} = 0
\end{aligned}
$$

Hence, the naive model mean parameters are

$$\alpha_{0,N} = \alpha_0 + \mu_{S_2}\alpha_2$$
$$\alpha_{1,N} = \alpha_1$$

2.A.2 Derivation of naive variance model parameters in scenario III

We want to derive the naive model variance parameters $\boldsymbol{\theta}_{2,N} = (\phi_N, \sigma_N^2)$. We solve eqn (6) for the variance model parameters ϕ_N, σ_N^2. Here, we cannot derive closed-form solutions, but we can derive equations which can be solved numerically.

Using the identity for the expectation of a quadratic form for the random vector $[\mathbf{X}^* - \mathbf{S}\boldsymbol{\theta}_N]$, We can rewrite the first term of eqn (2) as

$$\begin{aligned}
&\mathbb{E}\Big\{\big[\mathbf{X}^* - \mathbf{S}\boldsymbol{\theta}_N\big]^\mathsf{T}\mathbf{V}_N^{-1}\frac{\partial \mathbf{V}_N}{\partial \theta_{2,l,N}}\mathbf{V}_N^{-1}\big[\mathbf{X}^* - \mathbf{S}\boldsymbol{\theta}_N\big]\Big\} \\
&= \mathbb{E}\big[\mathbf{X}^* - \mathbf{S}\boldsymbol{\theta}_N\big]^\mathsf{T}\mathbf{V}_N^{-1}\frac{\partial \mathbf{V}_N}{\partial \theta_{2,l,N}}\mathbf{V}_N^{-1}\mathbb{E}\big[\mathbf{X}^* - \mathbf{S}\boldsymbol{\theta}_N\big] + \mathsf{trace}\Big\{\mathbf{V}_N^{-1}\frac{\partial \mathbf{V}_N}{\partial \theta_{2,l,N}}\mathbf{V}_N^{-1}\mathrm{Var}\big[\mathbf{X}^* - \mathbf{S}\boldsymbol{\theta}_N\big]\Big\}
\end{aligned}$$

Computing the expectation and covariance of the random vector $[\mathbf{X}^* - \mathbf{S}\boldsymbol{\theta}_N]$, we have

$$\begin{aligned}
\mathbb{E}\big[\mathbf{X}^* - \mathbf{S}\boldsymbol{\theta}_N\big] &= \mathbb{E}\big[\mathbb{E}\big[\mathbf{X}^* - \mathbf{S}\boldsymbol{\theta}_N\big|\mathbf{S}\big]\big] \\
&= \mathbf{0}
\end{aligned}$$

and

$$\mathrm{Var}\big[\mathbf{X}^* - \mathbf{S}\boldsymbol{\theta}_N\big] = \boldsymbol{\Sigma}_{X^*X^*}(\boldsymbol{\theta}_2) + \theta_1^2\mu_{S_1}^2\mathbf{1}\mathbf{1}^\mathsf{T}$$

Then we can plug in those expressions to eqn (6) and compute the iterated expectation to get a set of two equations and two unknowns which can be solved numerically.

Chapter 3

Consequences of kriging and land use regression for PM2.5 predictions in health effects analyses: Insights into spatial variability using high-resolution satellite data.

Stacey E. Alexeeff[1], Joel Schwartz[2], Itai Kloog[2], Alexandra Chudnovsky[2], and Brent A. Coull[1]

[1] Department of Biostatistics, Harvard School of Public Health

[2] Department of Environmental Health, Harvard School of Public Health

3.1 Introduction

There is strong epidemiological evidence that both short-term and long-term exposures to air pollution are related to cardiovascular morbidity and mortality.(Brook et al., 2010) In particular, much of the air pollution research shows that exposure to ambient particulate matter (PM) with aerodynamic diameter $\leq 2.5\mu/m^3$ ($PM_{2.5}$) is associated with many adverse cardiovascular outcomes. Within a given city region, ambient levels of $PM_{2.5}$ often vary, and traffic sources may contribute to this variation.(Brauer et al., 2003; Clougherty et al., 2008) However, levels of $PM_{2.5}$ are typically measured only at a small number of stationary monitoring sites which may only capture a small part of the overall regional heterogeneity.

Because it is not usually feasible to obtain exposure recordings at each study subject's residence over an entire study period, researchers often use existing pollution monitoring networks to gather data on the variability in $PM_{2.5}$ levels over space, and then build prediction models based on these data that can be used to estimate location-specific exposures throughout the study region. Thus, spatial modeling of air-pollution levels is becoming widespread in air pollution epidemiology research. Ordinary kriging (with a constant mean) and universal kriging (with a mean function that depends on spatial covariates) have been used to predict PM2.5 exposures and study relationships with health, such as the assessment of the short-term relationship between PM2.5 and cardiac responses(Liao et al., 2006) and chronic health effects between PM2.5 and cancer mortality.(Jerrett et al., 2005)

The use of predicted exposures in a health effects analysis is an example of measurement error because the predicted exposures represent an imperfect surrogate of the true exposure. The most common use of predicted exposures in a health effects analysis is the direct plug-in of the individual-specific exposure estimates. This approach treats the exposures as known, without acknowledgement of the uncertainty in the prediction process. In general, ignoring this exposure measurement error can lead to biased health effect estimates and overstated confidence in the resulting risk assessments.(Carroll et al., 2006; Gryparis et al., 2009)

The degree to which this exposure measurement error may be affecting health effect analyses is largely unknown because of inherent lack of validation data to study such an issue. In particular, a complete spatial picture has not been available. Cross-validation methods can be used to assess model prediction performance only at the small number of locations where monitoring data is available. These assessments of prediction error in the exposure modeling stage do not necessarily translate into knowledge of the impact of measurement error on the health effect estimates.

In the statistical literature, simulation studies have been primarily used to assess the degree of bias and variance corrections needed, and to evaluate the performance of statistical measurement error correction strategies. A number of papers have found in simulation studies that direct use of the predicted exposures often induces little to no bias.(Szpiro et al., 2011b; Madsen et al., 2008; Lopiano et al., 2011; Gryparis et al., 2009) However, those simulation studies only use known exposure surfaces and fit the correct exposure model used to generate the data. In real data scenarios, the actual performance of the naive plug-in estimator, the degree to which bias and variance adjustments need to be made, and the performance of the current adjustment methods is largely unknown. It is important to identify situations where the plug-in estimator may be unbiased versus situations where both a bias and variance adjustment may be necessary.

A gold standard for the fine-scale spatial distribution of air pollution throughout an entire region has never previously been available. Satellite data on aerosol optical depth (AOD) is now readily available and can be calibrated to reflect PM2.5 concentrations.(Kloog et al., 2011, 2012b) We propose that calibrated high-resolution satellite data could be viewed as a "silver standard" of comparison to evaluate the use of spatial air pollution predictions.

In this study, we investigate the practical implications for what happens when modeling real air pollution surfaces. We study the effects of measurement error on health effect estimates via a simulation study based on high resolution satellite data, where the true exposure surface is represented by calibrated satellite AOD data.

3.2 Materials and Methods

3.2.1 Satellite AOD Data.

Daily spectral AOD data was obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS) on the Terra and Aqua satellites for the year 2003. Further details about MODIS satellite aerosol data retrieval and validation have been published previously.(Remer et al., 2005; Levy et al., 2007) The daily data are freely available online through the NASA Web site.(NASA, 2012)

For this study, MODIS level 2 files from the Terra and Aqua satellites were obtained at the spatial resolution of 1 km $\times$ 1 km at nadir. Daily values of AOD were assigned to the grid cell where the AOD retrieval centroid was located. One feature of the AOD data is that some of the grid-specific AOD values are missing on some days due to cloud cover or snow cover.(Kloog et al., 2011) Thus, the spatial coverage of the AOD data varies considerably by day.

3.2.2 Air Pollution Monitors.

Data for daily $PM_{2.5}$ mass concentrations across New England (see Figure 3.1) for the year 2003 were obtained from the U.S. Environmental Protection Agency (EPA) Air Quality System (AQS) database as well as the IMPROVE (Interagency Monitoring of Protected Visual Environments) network. IMPROVE monitor sites are located in national parks and wilderness areas while EPA monitoring sites are located across New England including urban areas such as downtown Boston. There were 72 monitors with unique locations operating in New England during the study period.

3.2.3 Spatial and Temporal Covariates.

Spatial covariates included major roads, point emissions and area emissions. Data on the density of major roads was based on A1 roads (hard surface highways including Interstate and U.S. numbered highways, primary State routes, and all controlled access highways) data obtained through the US census 2000 topologically integrated geographic encoding and referencing system. Because the distributions of major roads were highly right-skewed, they were log transformed.

Temporal covariates included wind speed, humidity, visibility, height of the planetary boundary layer. All meteorological variables (temperature, wind speed, humidity, visibility) were obtained through the national climatic data center (NCDC). Further details on spatial and temporal covariates are given in Kloog et al. (2011) and Kloog et al. (2012b).

3.2.4 Calibration of AOD.

A description of the method used to calibrate the AOD values to represent $PM_{2.5}$ concentrations is given in Kloog et al. (2011) and Kloog et al. (2012b). Briefly, the relationship between $PM_{2.5}$ and AOD at the monitoring sites was modeled using a mixed-effects regression model where $PM_{2.5}$ was the dependent variable and AOD was a predictor. The model included spatial covariates for major roads, point emissions and area emissions, and temporal covariates for wind speed, visibility, height of the planetary boundary layer, with interactions between AOD and random intercepts for each day.

Kloog et al. (2011) also includes a third stage of modeling which imputes $PM_{2.5}$ at the missing AOD locations. In this study, we restricted to only days with ample AOD present to leverage the observed spatial variability in the data and minimize the use of known land-use regression models.

3.2.5 Simulation setup.

A simulation study was conducted to assess the performance of kriging and land use regression methods on a realistic representation of an air pollution surface. Separate simulation studies were conducted to consider studies of chronic health effects due to long-term air pollution exposures and acute health effects due to short-term air pollution exposures.

We considered two types of health effect models: a binary health outcome and a continuous health outcome. A linear regression health model was assumed for the continuous health outcome, where the outcome depends linearly on the exposure. For the binary health outcome, a logistic regression health model was assumed, where the outcome depends linearly on the exposure through a logit link function. No other confounding variables were included in the health model. We explored exposure models with four different covariance models (Matern function with different levels of smoothness, indexed by κ). We also contrasted two settings for the number of monitors where $m = 100$ is the realistic setting (although still higher than the actual number of monitors in this region during the study period), and $m = 500$ to represent a unrealistic "best case scenario" with much more spatial coverage to help highlight the problems due to sample size vs the problems due to model misspecification.

We restricted our simulation studies to the 32 days such that at least 50,000 grid-cells of AOD data were available from the satellite.

Chronic Effects Simulation

To emulate the setting of a health study of the chronic effects of particulate matter, we generated a chronic exposure surface by averaging the calibrated $PM_{2.5}$ data at each grid-cell over the 32 days of exposure. In this scenario, all subjects' exposures were sampled from this one common exposure surface. Thus, the spatial variability of the surface provided the only variability in the exposures of different subjects.

For each simulation, we generated 500 subjects' exposure and outcome measurements.

To assign the exposure, we first generated each subjects' residential location by population density. Population density sampling was approximated by using the geocoded locations of births during 2003 from a previous study.(Kloog et al., 2012a) We then assigned the corresponding calibrated $PM_{2.5}$ value at the subjects' residential location as the exposure. The health outcome was generated to depend on the assigned exposure using the chosen health model type with no confounders. The monitor locations were chosen by a random uniform distribution across the exposure surface, and the corresponding calibrated $PM_{2.5}$ value at the monitor location was used as the observed exposure. Using the measured exposure at the monitor locations, the kriging or land use model was fit to the data and chronic exposure predictions were generated at the residential locations of the subjects. The predicted exposures were then fit to the health outcomes to estimate the association.

Acute Effects Simulation

We designed our acute effects simulation to mimic the setting of a health study of the short-term effects of particulate matter. Using the 32 days of calibrated $PM_{2.5}$, we considered the exposure period of interest to be one day of $PM_{2.5}$ exposure. For each simulation, we generated 1,000 subjects' residential location by randomly sampling the day of the exposure and then sampling the health locations by population density, as in the chronic simulation. Once the date and grid-cell were randomly chosen, we assigned the corresponding calibrated $PM_{2.5}$ exposure at the grid-cell. The health outcomes were generated to depend on the assigned exposure using the chosen health model type with no confounders. We simulated 1000 subjects per simulation so that there were approximately 30 subjects sampled from each of the 32 days. The monitor locations were chosen by a random uniform distribution across the exposure surface, and the corresponding daily calibrated $PM_{2.5}$ value at the monitor location was used as the observed exposure for each day. Using the measured exposure at the monitor locations, the kriging or land use model was fit to the data by day and exposure predictions were generated for each day at the residential locations of the subjects. The predicted exposures were then fit to the health outcomes to estimate the association.

3.3 Results

The average daily $PM_{2.5}$ levels from the calibrated AOD data ranged from $1.98\mu g/m^3$ to $16.82\mu g/m^3$, with a mean of $7.47\mu g/m^3$. The $PM_{2.5}$ levels on all days at all locations ranged from $0.002\mu g/m^3$ to $20.0\mu g/m^3$. Between-day variability accounted for 92% of the total variation in $PM_{2.5}$ while the within-day variability accounted for 8% of the total variation in $PM_{2.5}$ levels. A table summarizing the daily mean, SD, and number of grid-cells for the $PM_{2.5}$ concentrations for each of the 32 days used in the study is given in the Supplementary Material section. Figure 3.1 shows the $PM_{2.5}$ levels for one date, Sept 10, 2003, and the $PM_{2.5}$ levels for the chronic average surface.

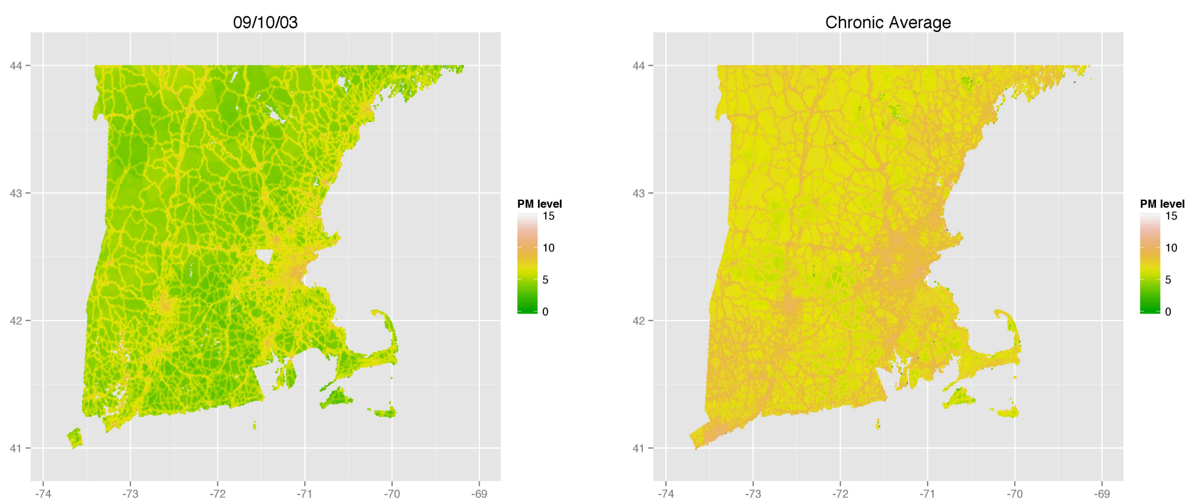

Figure 3.1: $PM_{2.5}$ concentrations with satellite grid-cells at 1km x 1km resolution for (a) one day September 10, 2003, (b) chronic average surface.

The results for the simulations of chronic effects of air pollution are shown in Tables 3.1 and 3.2, where Table 3.1 shows the results for a a linear model relating chronic air pollution exposure to a continuous health outcome, and Table 3.2 shows the results for a logistic model relating chronic air pollution exposure to a binary health outcome.

Table 3.1: Linear regression health effects with chronic exposure to air pollution, fit using (i) the true exposure, (ii) the predicted exposures using kriging model with constant mean, and (iii) a kriging model with land use terms.

Scenario	κ	m	mean β	empir SE	model SE	MSE	Coverage
Chronic, True X			1.001	0.029	0.030	0.001	95.1
Chronic,	0.5	100	1.602	0.871	0.181	1.121	31.0
Kriging only	1.0	100	1.541	0.846	0.163	1.008	31.8
	2.0	100	1.530	0.763	0.555	0.863	32.5
	3.0	100	1.533	0.775	0.529	0.885	32.7
	0.5	500	1.240	0.202	0.084	0.098	35.0
	1.0	500	1.232	0.204	0.086	0.096	37.9
	2.0	500	1.221	0.208	0.088	0.092	40.9
	3.0	500	1.213	0.208	0.088	0.089	42.3
Chronic,	0.5	100	1.051	0.141	0.047	0.023	47.4
land use	1.0	100	1.045	0.143	0.047	0.022	47.6
	2.0	100	1.042	0.144	0.047	0.022	48.0
	3.0	100	1.042	0.144	0.048	0.022	47.4
	0.5	500	1.014	0.077	0.038	0.006	67.9
	1.0	500	1.013	0.078	0.038	0.006	68.0
	2.0	500	1.013	0.079	0.038	0.006	68.0
	3.0	500	1.012	0.079	0.038	0.006	68.3

Compared to using the true chronic exposure, the kriging only models have notable upward bias and highly inflated empirical standard errors. We see this pattern in both the linear and logistic health effects models. The magnitude of the bias does not seem to vary by κ, the degree of smoothness of the assumed surface, yet the bias does diminish some when the number of monitors, m, increases. There is substantial under-coverage by the naive confidence intervals in the linear model, due to both the bias and the discrepancy between the naive model-based standard error and the empirical standard error. The naive standard error in the logistic model better reflects the empirical standard error, and despite the problems with bias in the logistic model, there is only slight under-coverage by the 95% confidence intervals when $m = 100$.

In both the linear and logistic health effect models, there was notable improvement in terms of both bias and standard errors in the universal kriging model which included two

Table 3.2: Logistic regression health effects with chronic exposure to air pollution, fit using (i) the true exposure, (ii) the predicted exposures using kriging model with constant mean, and (iii) a kriging model with land use terms.

Scenario	κ	m	OR	empir SE (β)	model SE (β)	MSE (β)	Coverage
Chronic, true X			2.028	0.167	0.165	0.028	95.2
Chronic,	0.5	100	2.573	0.770	0.514	0.656	89.3
Kriging only	1.0	100	2.526	0.955	0.469	0.966	89.8
	2.0	100	2.503	0.696	1.380	0.534	90.6
	3.0	100	2.492	0.671	1.415	0.498	90.0
	0.5	500	2.170	0.293	0.284	0.092	94.8
	1.0	500	2.148	0.296	0.287	0.092	95.2
	2.0	500	2.119	0.298	0.289	0.092	94.9
	3.0	500	2.104	0.299	0.289	0.092	95.2
Chronic,	0.5	100	2.111	0.270	0.228	0.076	91.0
land use	1.0	100	2.093	0.267	0.226	0.073	90.9
	2.0	100	2.088	0.263	0.226	0.071	91.4
	3.0	100	2.086	0.266	0.226	0.072	91.2
	0.5	500	2.076	0.215	0.196	0.047	93.8
	1.0	500	2.073	0.216	0.196	0.048	93.8
	2.0	500	2.072	0.216	0.196	0.048	93.7
	3.0	500	2.071	0.216	0.196	0.048	93.7

land use terms. These models exhibited only slight upward bias in the health effect estimates, although there was still significant under-coverage in the linear health effect model.

The results for the simulations of acute effects of air pollution are shown in Tables 3.3 and 3.4, where Table 3.3 shows the results for a linear model relating acute air pollution exposure to a continuous health outcome, and Table 3.4 shows the results for a logistic model relating acute air pollution exposure to a binary health outcome.

In the linear health effects setting, the kriging only models show no bias and only slightly inflated empirical standard errors compared to using the true acute exposure. Similarly, in the logistic health effects setting, the kriging only models show slight downward bias and no inflation of the empirical standard errors compared to using the true acute exposure.

In contrast, there was considerable downward bias and inflation of empirical standard errors in both the linear and logistic health effect setting for the spatial-temporal model with

Table 3.3: Linear regression health effects with acute exposure to air pollution, fit using (i) the true exposure, (ii) the predicted exposures using kriging model with constant mean, and (iii) a kriging model with land use terms.

Scenario	κ	m	mean β	empir SE	model SE	MSE	Coverage
Acute, True X			1.000	0.006	0.006	0.000	95.2
Acute, kriging only	0.5	100	1.021	0.016	0.013	0.001	61.2
	1.0	100	1.020	0.017	0.013	0.001	61.6
	2.0	100	1.020	0.017	0.013	0.001	62.0
	3.0	100	1.020	0.017	0.013	0.001	63.6
	0.5	500	1.024	0.012	0.011	0.001	41.8
	1.0	500	1.024	0.013	0.011	0.001	43.2
	2.0	500	1.024	0.013	0.011	0.001	44.3
	3.0	500	1.023	0.012	0.011	0.001	46.4
Acute, land use	0.5	100	0.451	0.068	0.023	0.307	0.0
	1.0	100	0.411	0.074	0.023	0.353	0.0
	2.0	100	0.365	0.078	0.023	0.410	0.0
	3.0	100	0.343	0.080	0.023	0.438	0.0
	0.5	500	0.641	0.036	0.021	0.130	0.0
	1.0	500	0.629	0.035	0.021	0.139	0.0
	2.0	500	0.607	0.035	0.022	0.155	0.0
	3.0	500	0.592	0.039	0.022	0.168	0.0

universal kriging which included both temporal and spatial land use terms. In addition, there was no coverage of the true effect by the naive confidence intervals in either the linear or the logistic setting, due to both the bias and the discrepancy between the naive model-based standard error and the empirical standard error.

3.4 Discussion

In this study, we found that there may be substantial bias of health effect estimates in models using exposures predicted by kriging or land use regression. We found that the direction of bias may be either toward or away from the null, and the degree of bias varies by the type of study, with some exposure predictions working well in certain situations. We also found substantial under-coverage where the true effect was often not included in the naive 95%

Table 3.4: Logistic regression health effects with acute exposure to air pollution, fit using (i) the true exposure, (ii) the predicted exposures using kriging model with constant mean, and (iii) a kriging model with land use terms.

Scenario	κ	m	OR	empir SE (β)	model SE (β)	MSE (β)	Coverage
Acute, True X			2.000	0.050	0.050	0.002	95.2
Acute, kriging	0.5	100	1.884	0.045	0.045	0.006	70.6
	1.0	100	1.885	0.047	0.046	0.006	69.2
	2.0	100	1.888	0.048	0.046	0.006	68.8
	3.0	100	1.889	0.048	0.046	0.006	70.4
	0.5	500	1.926	0.047	0.047	0.004	83.8
	1.0	500	1.923	0.047	0.046	0.004	82.6
	2.0	500	1.924	0.045	0.046	0.004	84.4
	3.0	500	1.920	0.046	0.046	0.004	82.3
Acute, land use	0.5	100	1.261	0.041	0.020	0.214	0.0
	1.0	100	1.234	0.042	0.019	0.235	0.0
	2.0	100	1.204	0.042	0.018	0.260	0.0
	3.0	100	1.190	0.042	0.017	0.271	0.0
	0.5	500	1.435	0.039	0.027	0.112	0.0
	1.0	500	1.414	0.037	0.026	0.122	0.0
	2.0	500	1.414	0.037	0.025	0.122	0.0
	3.0	500	1.376	0.038	0.025	0.141	0.0

confidence interval. We gained these insights into the spatial variability of $PM_{2.5}$ predictions by using high-resolution satellite data on aerosol optical depth, which were calibrated to reflect $PM_{2.5}$ concentrations.

In the chronic simulations where spatial variability provided the only source of variation in the exposures, kriging alone was insufficient to model and predict exposures. However, when only two land-use regression terms were added, the degree of bias was substantially decreased. This demonstrates that not all types of model misspecification lead to substantial bias in health effect estimates; we know that the chronic land use model with two land use terms and spatial covariance is not the true model, yet it performs relatively well. On the other hand, we still observe $4-5\%$ upward bias in when using these chronic land use exposure models in the realistic setting of $m = 100$ monitors.

The most surprising result is that in the acute setting, the daily kriging model worked well yet the model incorporating spatial and temporal covariates performed very poorly. The success of the kriging models in this setting is most likely related to the aforementioned fact that in the acute model, 92% of the variability of the true exposure is temporal, while only 8% of the variability is spatial. A recent brief report suggests that predicted exposures with higher R^2 in the exposure model may not always improve the quality of health effect estimates.(Szpiro et al., 2011a)

Interestingly, the overall performance of simulations did not vary by the covariance model chosen, as evidenced by similar results in each setting across varying κ. Note that $\kappa = 0.5$ corresponds to the exponential covariance structure. Hence, the choice of spatial covariance model may not play a strong role in the effectiveness of using exposure predictions in health effect analyses.

Other statistical studies assessing performance of kriging and land use regression models have not considered real surfaces which may not behave like smooth surfaces. In the current literature, authors have found in simulation studies that direct use of exposure predictions in health effects models often induces little to no bias.(Szpiro et al., 2011b; Madsen et al., 2008; Lopiano et al., 2011; Gryparis et al., 2009) However, those papers investigate bias in simulation studies only by fitting the correct exposure model used to generate the data. Madsen et al. (2008) and Szpiro et al. (2011b) assume smooth exposure surfaces that can be fit well using kriging methods, finding no need for bias correction, thus the papers focus on ways to estimate standard errors instead.

The issue of model misspecification in spatial exposure models has not been a focus of previous statistical research in the area of measurement error in air pollution epidemiology. The results of this simulation study suggest a direction of future statistical research needed to understand the implications of mis-specifying exposure models, to provide appropriate diagnostic procedures, and to implement effective measurement error correction strategies.

Limitations

Any simulation study of this nature will need to make some decisions about how to setup the simulations. Thus, there will always be some limitation of not considering every possible scenario one might think of. However, we have attempted to provide a range of simulations with varying degrees of temporal and spatial variability.

There are many other potential sources of measurement error in air pollution epidemiology studies not considered here. Zeger et al. (2000) provides a framework for considering a number of sources of exposure measurement error in air pollution research. We also assumed no confounding to fully distinguish the problems stemming from the measurement error inherent in exposure modeling from the issue of confounding. The combination of misspecified exposure models and incomplete control for confounding variables may introduce different problems and is not yet known.

The days in which AOD measurements are more complete represent days which are more sunny and the ground has less snow cover. Hence these days are not a representative sample of all days throughout the year. Thus, other days which are more cloudy may have a different spatial distribution due to differences in the height of the planetary boundary layer and other factors.

This study does not suggest that satellite calibrated AOD measurements are a perfect measure of true air pollution exposure. Rather, this study uses the daily satellite calibrated AOD measurements as representations of spatial variability in daily exposure surfaces of $PM_{2.5}$ to better understand how kriging may perform in a real life setting. Without the availability of considerably more spatial coverage of air pollution monitoring data, the question of how well satellite calibrated AOD measurement reflect true spatial variation in $PM_{2.5}$ exposures may be difficult to evaluate.

There remains no gold standard for the entire fine-scale spatial distribution of particulate matter throughout a region. While this study can lend insight into potential performance of kriging, land use regression, and spatio-temporal modeling by using a more realistic representation of a regional $PM_{2.5}$ surface, it is not generalizable to all possible true air pollution surfaces. Rather, these simulations serve as examples of potential impacts kriging and land use regression may perform better or worse.

Overall, this simulation study uses high-resolution satellite data to provide several settings with realistic exposure surfaces, and suggests that (i) kriging and land use regression models sometimes work well in health effect models but sometimes introduce substantial biases, (ii) the success in using modeled exposures varies by the spatial and temporal properties of the underlying data, and (iii) future statistical research is needed to better understand and deal with these issues.

3.5 References

M. Brauer, G. Hoek, P. van Vliet, K. Meliefste, P. Fischer, U. Gehring, J. Heinrich, J. Cyrys, T. Bellander, M. Lewne, and B. Brunekreef. Estimating long-term average particulate air pollution concentrations: application of traffic indicators and geographic information systems. *Epidemiology*, 14(2):228–239, 2003.

R. D. Brook, S. Rajagopalan, C. A. P. III, J. R. Brook, A. Bhatnagar, A. V. Diez-Roux, F. Holguin, Y. Hong, R. V. Luepker, M. A. Mittleman, A. Peters, D. Siscovick, S. C. S. Jr, L. Whitsel, and J. D. Kaufman. A.h.a. scientific statement. particulate matter air pollution and cardiovascular disease. *Circulation*, 121:2331–2378, 2010.

R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman & Hall, New York, New York, 2nd edition, 2006.

J. Clougherty, R. Wright, L. Baxter, and J. Levy. Land use regression modeling of intra-urban residential variability in multiple traffic-related air pollutants. *Environ Health*, 7: 17, 2008.

A. Gryparis, C. J. Paciorek, A. Zeka, J. Schwartz, and B. A. Coull. Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics*, 10:258–274, 2009.

M. Jerrett, R. T. Burnett, R. Ma, C. A. P. III, D. Krewski, K. B. Newbold, . G. Thurston, Y. Shi, N. Finkelstein, . E. E. Calle, , and M. J. Thun. Spatial analysis of air pollution and mortality in los angeles. *Epidemiology*, 16:727–736, 2005.

I. Kloog, P. Koutrakis, B. Coull, H. Lee, and J. Schwartz. Temporally and spatially resolved pm2. 5 exposures for epidemiological studies using satellite aerosol optical depth measurements. *Atmospheric Environment*, 45:6267–6275, 2011.

I. Kloog, S. J. Melly, W. L. Ridgway, B. A. Coull, and J. Schwartz. Using new satellite based exposure methods to study the association between pregnancy pm2.5 exposure, premature birth and birth weight in massachusetts. *Environmental Health*, 11:40, 2012a.

I. Kloog, F. Nordio, B. Coull, and J. Schwartz. Incorporating local land use regression and satellite aerosol optical depth in a hybrid model of spatiotemporal pm2.5 exposures in the mid-atlantic states. *Environmental Science and Technology*, 46:11913–11921, 2012b.

R. Levy, L. Remer, S. Mattoo, E. Vermote, and Y. Kaufman. Second-generation operational algorithm: Retrieval of aerosol properties over land from inversion of moderate resolution imaging spectroradiometer spectral reflectance. *Journal of Geophysical Research: Atmospheres*, 112:D13211, 2007.

D. Liao, D. J. Peuquet, Y. Duan, E. A. Whitsel, J. Dou, R. L. Smith, H.-M. Lin, J.-C. Chen, and G. Heiss. Gis approaches for the estimation of residential-level ambient pm concentrations. *Environmental Health Perspectives*, 114:1374–1380, 2006.

K. Lopiano, L. Young, and C. Gotway. A comparison of errors in variables methods for use in regression models with spatially misaligned data. *Statistical Methods in Medical Research*, 20:29–47, 2011.

L. Madsen, D. Ruppert, and N. Altman. Regression with spatially misaligned data. *Environmetrics*, 19:453–467, 2008.

NASA. The moderate resolution imaging spectroradiometer website. 2012.

L. A. Remer, Y. J. Kaufman, D. Tanr, S. Mattoo, D. A. Chu, J. V. Martins, R.-R. Li, C. Ichoku, R. C. Levy, R. G. Kleidman, T. F. Eck, E. Vermote, and B. N. Holben. The modis aerosol algorithm, products, and validation. *Journal of the Atmospheric Sciences*, 62:947–973, 2005.

A. Szpiro, C. J. Paciorek, and L. Sheppard. Does more accurate exposure prediction necessarily improve health effect estimates? *Biostatistics*, 22:680–685, 2011a.

A. Szpiro, L. Sheppard, and T. Lumley. Efficient measurement error correction with spatially misaligned data. *Biostatistics*, 12:610–623, 2011b.

S. L. Zeger, D. Thomas, F. Dominici, J. M. Samet, J. Schwartz, D. Dockery, and A. Cohen. Exposure measurement error in timeseries studies of air pollution: Concepts and consequences. *Environ Health Perspect*, 108:419–426, 2000.

Supplementary Figures and Tables.

Table 4.5: Exposure parameters in naive misspecified model.

Parameter	True value	True model		Naive model,	Naive model	
		Mean	empirical SE	theoretical	Mean	empirical SE
α_0	2.0	1.992	0.390	1.5	1.460	0.556
α_1	1.0	0.999	0.016	1.0	1.002	0.181
α_2	1.0	1.000	0.006	0.0	0.0	0.0
ϕ	0.2	0.194	0.026	0.14	0.100	0.058
σ_2	0.5	0.457	0.270	0.56	0.602	0.531
κ	3.0	3.0	0.0	3.0	3.0	0.0

Table 4.6: Mean and Standard Deviation of $PM_{2.5}$ concentration and number of 1km satellite gridpoints for each day in simulations.

Date	Mean	SD	n	Date	Mean	SD	n
01/01/03	10.38	1.53	57,414	09/10/03	5.08	1.07	66,861
04/17/03	5.45	0.85	54,198	09/11/03	5.48	1.25	52,165
04/25/03	5.1	0.85	61,725	09/17/03	6.38	1.22	57,939
04/28/03	10.04	1.19	63,447	10/03/03	2.6	0.79	63,958
05/03/03	3.89	0.97	63,476	10/07/03	8.74	0.99	62,323
05/17/03	5.73	1.13	66,254	10/08/03	16.82	1.29	51,517
05/18/03	6.57	1.25	58,435	10/11/03	15.58	1.72	52,055
05/19/03	8.18	1.22	68,243	10/13/03	4.78	1.06	66,377
06/17/03	3.51	1.26	60,201	10/30/03	6.55	0.94	50,036
06/23/03	8.25	1.54	51,392	11/08/03	3.74	0.75	61,573
06/24/03	16.34	2.08	59,135	11/09/03	5.99	0.87	66,142
08/20/03	14.07	1.21	54,960	11/10/03	10.88	1.22	60,781
08/23/03	4.45	0.97	66,140	11/14/03	2.36	1	50,385
08/24/03	1.98	0.86	64,550	11/15/03	3.64	0.85	62,221
09/07/03	5.66	1.12	61,196	11/24/03	10.67	1.71	56,299
09/08/03	4.61	1.43	50,006	11/27/03	15.73	1.59	60,641

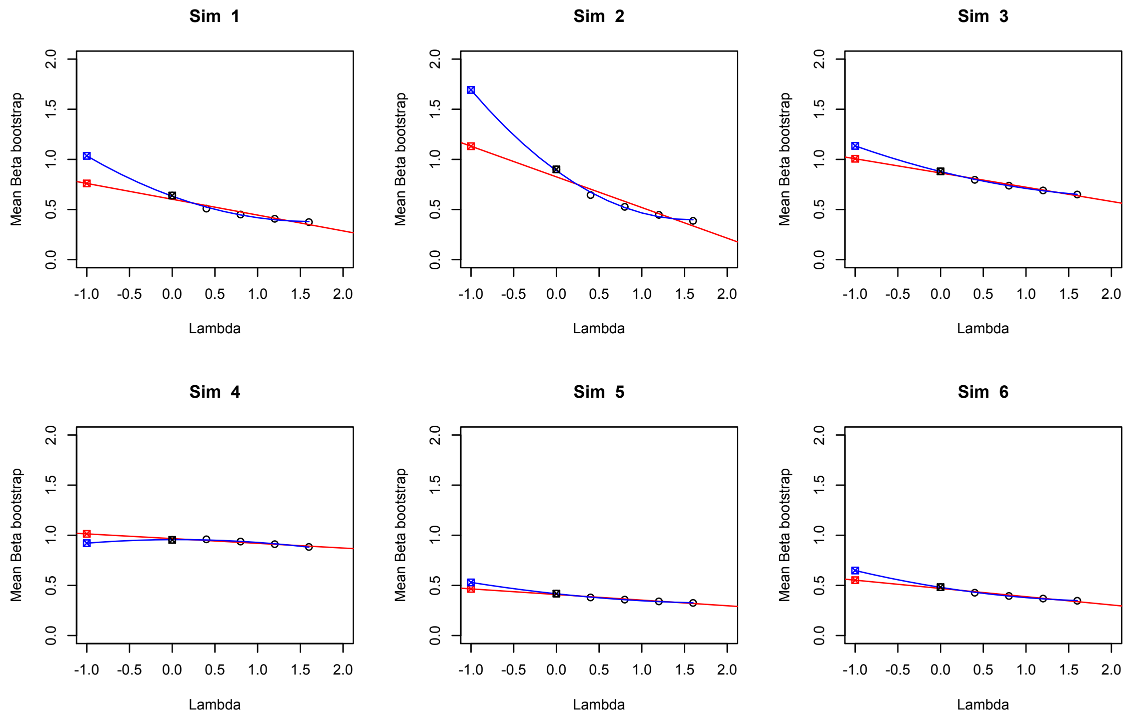

Figure 4.2: Examples of spatial SIMEX Extrapolations for six simulations.

Table 4.7: Results for sensitivity analysis in linear regression health effects with acute exposure to air pollution, fit using a kriging model with land use terms.

Scenario	k	m	Beta	empirical SE	model SE	MSE	Coverage
True X			1.000	0.005	0.006	0.000	95.6
Acute, land use	0.5	100	1.081	0.034	0.010	0.008	0.8
Acute, land use	1.0	100	1.110	0.047	0.011	0.014	0.6
Acute, land use	2.0	100	1.134	0.056	0.011	0.021	0.4
Acute, land use	3.0	100	1.145	0.060	0.012	0.025	0.4

Table 4.8: Results for sensitivity analysis in logistic regression health effects with acute exposure to air pollution, fit using a kriging model with land use terms.

Scenario	k	m	OR	empirical SE	model SE	MSE	Coverage
True X			2.000	0.050	0.050	0.002	95.2
Acute, land use	0.5	100	2.078	0.057	0.053	0.005	90.2
Acute, land use	1.0	100	2.124	0.065	0.054	0.008	82.0
Acute, land use	2.0	100	2.165	0.073	0.056	0.012	72.4
Acute, land use	3.0	100	2.184	0.077	0.057	0.014	68.4

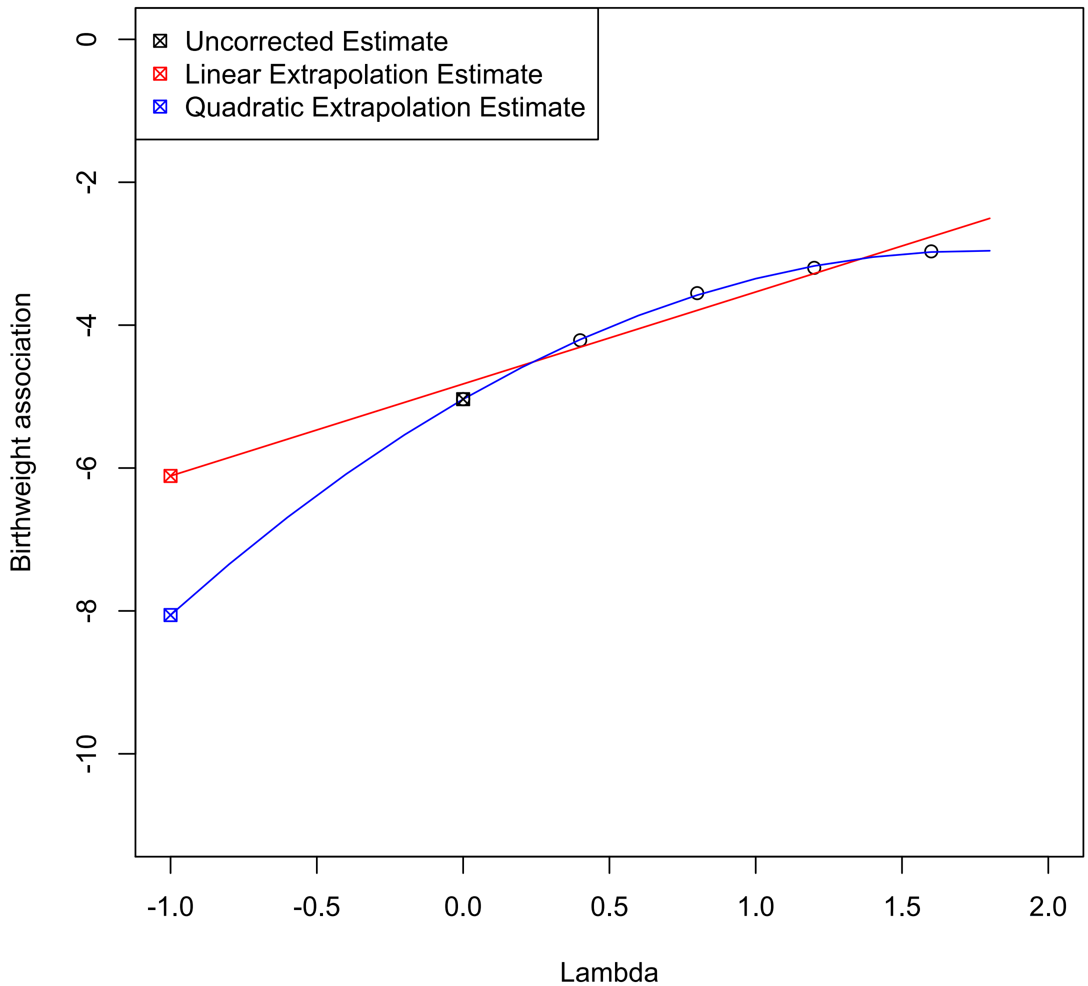

Figure 4.3: Spatial SIMEX Extrapolation for birthweight data.