A COVARIANCE REGRESSION MODEL

Peter D. Hoff and Xiaoyue Niu

University of Washington and Penn State University

Abstract: Classical regression analysis relates the expectation of a response variable to a linear combination of explanatory variables. In this article, we propose a covariance regression model that parameterizes the covariance matrix of a multivariate response vector as a parsimonious quadratic function of explanatory variables. The approach is analogous to the mean regression model, and is similar to a factor analysis model in which the factor loadings depend on the explanatory variables. Using a random-effects representation, parameter estimation for the model is straightforward using either an EM-algorithm or an MCMC approximation via Gibbs sampling. The proposed methodology provides a simple but flexible representation of heteroscedasticity across the levels of an explanatory variable, improves estimation of the mean function and gives better calibrated prediction regions when compared to a homoscedastic model.

Key words and phrases: Heteroscedasticity, positive definite cone, random effects.

1. Introduction

Estimation of a conditional mean function $\boldsymbol{\mu}_{\mathrm{x}} = \mathrm{E}\left[\boldsymbol{y}|\boldsymbol{x}\right]$ is a well studied data-analysis task for which there are a large number of statistical models and procedures. Less studied is the problem of estimating a covariance function $\boldsymbol{\Sigma}_{\mathrm{x}} = \mathrm{Var}\left[\boldsymbol{y}|\boldsymbol{x}\right]$ across a range of values for an explanatory $\boldsymbol{x}$-variable. In the univariate case, several procedures assume that the variance can be expressed as a function of the mean, i.e. $\sigma_{\mathrm{x}}^2 = g(\mu_{\mathrm{x}})$ for some known function g (see, for example, Carroll, Ruppert, and Holt (1982)). In many such cases the data can be represented by a generalized linear model with an appropriate variance function, or perhaps the data can be transformed to a scale for which the variance is constant as a function of the mean (Box and Cox (1964)). Other approaches separately parameterize the mean and variance, either giving a linear model for the standard deviation (Rutemiller and Bowers (1968)) or forcing the variance to be non-negative via a link function (Smyth (1989)). In situations where the explanatory variable x is continuous and the variance function is assumed to be smooth, Carroll (1982) and Müller and Stadtmüller (1987) propose and study kernel estimates of the variance function.

Models for multivariate heteroscedasticity have been developed in the context of multivariate time series, for which a variety of multivariate "autoregressive conditionally heteroscedastic" (ARCH) models have been studied (Engle and Kroner (1995); Fong, Li, and An (2006)). However, the applicability of such models is limited to situations where the heteroscedasticity is temporal in nature. A recent approach by Yin et al. (2010) uses a kernel estimator to allow $\boldsymbol{\Sigma}_\mathrm{x}$ to vary smoothly with $\boldsymbol{x}$. However, their focus is on a single continuous univariate explanatory variable, and it is not clear how to generalize such an approach to allow for discrete or categorical predictors. For many applications, it is desirable to construct a covariance function $\{\boldsymbol{\Sigma}_\mathrm{x} : \boldsymbol{x} \in \mathcal{X}\}$ for which the domain of the explanatory $\boldsymbol{x}$-variable is the same as in mean regression, that is, the explanatory vector can contain continuous, discrete and categorical variables. With this goal in mind, Chiu, Leonard, and Tsui (1996) suggested modeling the elements of the logarithm of the covariance matrix, $\boldsymbol{\Phi}_\mathrm{x} = \log \boldsymbol{\Sigma}_\mathrm{x}$, as linear functions of the explanatory variables, so that $\phi_{j,k,\mathrm{x}} = \boldsymbol{\beta}_{j,k}^T \boldsymbol{x}$ for unknown coefficients $\boldsymbol{\beta}_{j,k}$. This approach makes use of the fact that the only constraint on $\boldsymbol{\Phi}_\mathrm{x}$ is that it be symmetric. However, as the authors note, parameter interpretation for this model is difficult. For example, a submatrix of $\boldsymbol{\Sigma}_\mathrm{x}$ is not generally the matrix exponential of the same submatrix of $\boldsymbol{\Phi}_\mathrm{x}$, and so the elements of $\boldsymbol{\Phi}_\mathrm{x}$ do not directly relate to the corresponding covariances in $\boldsymbol{\Sigma}_\mathrm{x}$. Additionally, the number of parameters in this model can be quite large. For $\boldsymbol{y} \in \mathbb{R}^p$ and $\boldsymbol{x} \in \mathbb{R}^q$, the model involves a separate q-dimensional vector of coefficients for each of the $p(p + 1)/2$ unique elements of $\boldsymbol{\Phi}_\mathrm{x}$, thus requiring $q \times p(p + 1)/2$ parameters to be estimated.

Another clever reparameterization-based approach to covariance regression modeling was provided by Pourahmadi (1999), who suggested modeling the unconstrained elements of the Cholesky decomposition of $\boldsymbol{\Sigma}_\mathrm{x}^{-1}$ as linear functions of $\boldsymbol{x}$. The parameters in this model have a natural interpretation: The first $j - 1$ parameters in the jth row of the Cholesky decomposition relate to the conditional distribution of y_j given $y_1, \ldots, y_{j-1}$. This model is not invariant to reorderings of the elements of $\boldsymbol{y}$, and so is most appropriate when there is a natural order to the variables, such as with longitudinal data. Like the logarithmic covariance model of Chiu, Leonard, and Tsui (1996), the general form of the Cholesky factorization model requires $q \times p(p + 1)/2$ parameters to be estimated.

In this article we develop a simple parsimonious alternative to these reparameterization-based approaches. The covariance regression model we consider is based on an analogy with linear regression, and is given by $\boldsymbol{\Sigma}_\mathrm{x} = \boldsymbol{\Psi} + \boldsymbol{B}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{B}^T$, where $\boldsymbol{\Psi}$ is positive definite and $\boldsymbol{B}$ is a $p \times q$ real matrix. As a function of $\boldsymbol{x}$, $\boldsymbol{\Sigma}_\mathrm{x}$ is a curve within the cone of positive definite matrices. The $q \times p$ parameters of $\boldsymbol{B}$ have a direct interpretation in terms of how heteroscedasticity co-occurs among the p variables of $\boldsymbol{y}$. Additionally, the model has a random-effects representation,

allowing for straightforward maximum likelihood parameter estimation using the EM-algorithm, and Bayesian inference via Gibbs sampling. In the presence of heteroscedasticity, use of this covariance regression model can improve estimation of the mean function, characterize patterns of non-constant covariance, and provide prediction regions that are better calibrated than regions provided by homoscedastic models.

A geometric interpretation of the proposed model is developed in Section 2, along with a representation as a random-effects model. Section 3 discusses methods of parameter estimation and inference, including an EM-algorithm for obtaining maximum likelihood estimates (MLEs), an approximation to the covariance matrix of the MLEs, and a Gibbs sampler for Bayesian inference. A simulation study is presented in Section 4 that evaluates the estimation error of the regression coefficients in the presence of heteroscedasticity, the power of a likelihood ratio test of heteroscedasticity, as well as the coverage rates for approximate confidence intervals for model parameters. Section 5 considers an extension of the basic model to accommodate more complex patterns of heteroscedasticity, and Section 6 illustrates the model in an analysis of bivariate data on children's height and lung function. In this example it is shown that a covariance regression model provides better-calibrated prediction regions than a constant variance model. Section 7 provides a summary.

2. A Covariance Regression Model

Let $\boldsymbol{y} \in \mathbb{R}^p$ be a random multivariate response vector and $\boldsymbol{x} \in \mathbb{R}^q$ be a vector of explanatory variables. Our goal is to provide a parsimonious model and estimation method for $\mathrm{Cov}\,[\boldsymbol{y}|\boldsymbol{x}] = \boldsymbol{\Sigma}_{\mathrm{x}}$, the conditional covariance matrix of $\boldsymbol{y}$ given $\boldsymbol{x}$. We begin by analogy with linear regression. The simple linear regression model expresses the conditional mean $\boldsymbol{\mu}_{\mathrm{x}} = \mathrm{E}\,[\boldsymbol{y}|\boldsymbol{x}]$ as $\boldsymbol{b}+\boldsymbol{B}\boldsymbol{x}$, an affine function of $\boldsymbol{x}$. This model restricts the p-dimensional vector $\boldsymbol{\mu}_{\mathrm{x}}$ to a q-dimensional subspace of $\mathbb{R}^p$. The set of $p \times p$ covariance matrices is the cone of positive semidefinite matrices. This cone is convex and thus closed under addition. The simplest version of our proposed covariance regression model expresses $\boldsymbol{\Sigma}_{\mathrm{x}}$ as

$$\boldsymbol{\Sigma}_{\mathrm{x}} = \boldsymbol{\Psi} + \boldsymbol{B}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{B}^T, \tag{2.1}$$

where $\boldsymbol{\Psi}$ is a $p \times p$ positive-definite matrix and $\boldsymbol{B}$ is a $p \times q$ matrix. The resulting covariance function is positive definite for all $\boldsymbol{x}$, and expresses the covariance as equal to a "baseline" covariance matrix $\boldsymbol{\Psi}$ plus a rank-1, $p \times p$ positive definite matrix that depends on $\boldsymbol{x}$. The model given in (2.1) is in some sense a natural generalization of mean regression to a model for covariance matrices. A vector mean function lies in a vector (linear) space, and is expressed as a linear map from $\mathbb{R}^q$ to $\mathbb{R}^p$. The covariance matrix function lies in the cone of positive definite

matrices, where the natural group action is matrix multiplication on the left and right. The covariance regression model expresses the covariance function via such a map from the $q \times q$ cone to the $p \times p$ cone.

2.1. Model flexibility and geometry

Letting $\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_p\}$ be the rows of $\boldsymbol{B}$, the covariance regression model gives

$$\text{Var}\,[y_j|\boldsymbol{x}] = \psi_{j,j} + \boldsymbol{b}_j^T \boldsymbol{x}\boldsymbol{x}^T \boldsymbol{b}_j, \tag{2.2}$$

$$\text{Cov}\,[y_j, y_k|\boldsymbol{x}] = \psi_{j,k} + \boldsymbol{b}_j^T \boldsymbol{x}\boldsymbol{x}^T \boldsymbol{b}_k. \tag{2.3}$$

The parameterization of the variance suggests that the model requires the variance of each element of $\boldsymbol{y}$ to be increasing in the elements of $\boldsymbol{x}$, as the minimum variance is obtained when $\boldsymbol{x} = 0$. This constraint can be alleviated by including an intercept term so that the first element of $\boldsymbol{x}$ is 1. For example, in the case of a single scalar explanatory variable x, we abuse notation slightly and write $\boldsymbol{x} = (1, x)^T$, $\boldsymbol{b}_j = (b_{0,j}, b_{1,j})^T$, giving

$$\text{Var}\,[y_j|\boldsymbol{x}] = \psi_{j,j} + (b_{0,j} + b_{1,j}x)^2,$$

$$\text{Cov}\,[y_j, y_k|\boldsymbol{x}] = \psi_{j,k} + (b_{0,j} + b_{1,j}x)(b_{0,k} + b_{1,k}x).$$

For any given finite interval $(c, d) \subset \mathbb{R}$, there exist parameter values $(b_{0,j}, b_{1,j})$ so that the variance of y_j is either increasing or decreasing in x for $x \in (c, d)$.

We now consider the geometry of the covariance regression model. For each $\boldsymbol{x}$, the model expresses $\boldsymbol{\Sigma}_\text{x}$ as a point $\boldsymbol{\Psi}$ inside the positive-definite cone plus a rank-1 positive-semidefinite matrix $\boldsymbol{B}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{B}^T$. The latter matrix is a point on the boundary of the cone, so the range of $\boldsymbol{\Sigma}_\text{x}$ as a function of $\boldsymbol{x}$ can be seen as a submanifold of the boundary of the cone, but "pushed into" the cone by an amount $\boldsymbol{\Psi}$. Figure 1 represents this graphically for the simplest of cases in which $p = 2$ and there is just a single scalar explanatory variable x. In this case, each covariance matrix can be expressed as a three-dimensional vector $(\sigma_1^2, \sigma_2^2, \sigma_{1,2})$ such that

$$\sigma_1^2 \geq 0 \ , \ \sigma_2^2 \geq 0 \ , \ |\sigma_{1,2}| \leq \sigma_1\sigma_2.$$

The set of such points constitutes the positive semidefinite cone, whose boundary is shown by the outer surfaces in the two plots in Figure 1. The range of $\boldsymbol{B}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{B}^T$ over all $\boldsymbol{x}$ and matrices $\boldsymbol{B}$ includes the set of all rank-1 positive definite matrices, which is simply the boundary of the cone. Thus the possible range of $\boldsymbol{\Psi} + \boldsymbol{B}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{B}^T$ for a given $\boldsymbol{\Psi}$ is simply the boundary of the cone, translated by an amount $\boldsymbol{\Psi}$. Such a translated cone is shown from two perspectives in Figure 1. For a given $\boldsymbol{\Psi}$ and $\boldsymbol{B}$, the covariance regression model expresses $\boldsymbol{\Sigma}_\text{x}$ as a curve on this translated boundary. A few such curves for six different values of $\boldsymbol{B}$ are shown in black in Figure 1.
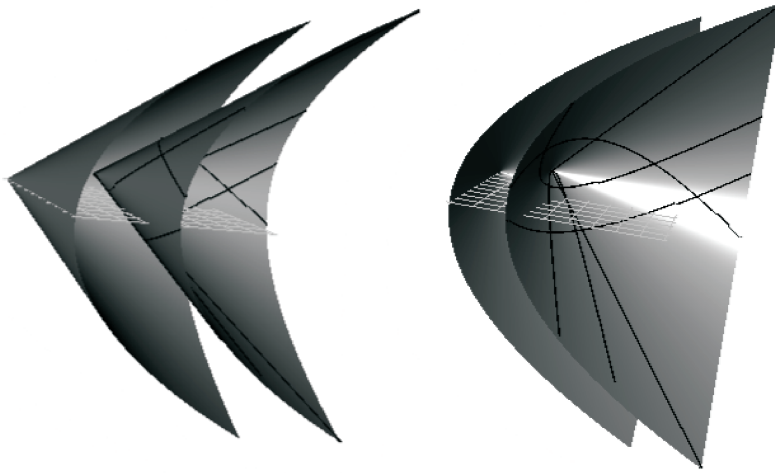
Figure 1. The positive-definite cone and a translation, from two perspectives. The outer surface is the boundary of the the positive definite cone, and the inner cone is the boundary plus a positive definite matrix $\boldsymbol{\Psi}$. Black curves on the inner cone represent covariance regression curves $\boldsymbol{\Psi} + \boldsymbol{Bxx}^T\boldsymbol{B}^T$ for different values of $\boldsymbol{B}$.

The parameters in the covariance regression model are generally identifiable given sufficient variability in the regressor $\boldsymbol{x}$, at least up to sign changes of $\boldsymbol{B}$. To see this, consider the simple case of a single scalar explanatory variable x. Abusing notation slightly, let $\boldsymbol{x} = (1, x)^T$ so that (2.1) becomes

$$\boldsymbol{\Sigma}_{\mathrm{x}}(\boldsymbol{\Psi}, \boldsymbol{B}) = \boldsymbol{\Psi} + \boldsymbol{b}_1\boldsymbol{b}_1^T + (\boldsymbol{b}_1\boldsymbol{b}_2^T + \boldsymbol{b}_2\boldsymbol{b}_1^T)x + \boldsymbol{b}_2\boldsymbol{b}_2^T x^2.$$

Now suppose that $(\tilde{\boldsymbol{\Psi}}, \tilde{\boldsymbol{B}})$ are such that $\boldsymbol{\Sigma}_{\mathrm{x}}(\boldsymbol{\Psi}, \boldsymbol{B}) = \boldsymbol{\Sigma}_{\mathrm{x}}(\tilde{\boldsymbol{\Psi}}, \tilde{\boldsymbol{B}})$ for all $x \in \mathbb{R}$. Setting $x = 0$ indicates that $\boldsymbol{\Psi} + \boldsymbol{b}_1\boldsymbol{b}_1^T = \tilde{\boldsymbol{\Psi}} + \tilde{\boldsymbol{b}}_1\tilde{\boldsymbol{b}}_1^T$. Considering $x = \pm 1$ implies that $\boldsymbol{b}_2\boldsymbol{b}_2^T = \tilde{\boldsymbol{b}}_2\tilde{\boldsymbol{b}}_2^T$ and thus that $\tilde{\boldsymbol{b}}_2 = \pm\boldsymbol{b}_2$. If $\boldsymbol{b}_2 \neq \boldsymbol{0}$, we have $\boldsymbol{b}_1\boldsymbol{b}_2^T + \boldsymbol{b}_2\boldsymbol{b}_1^T = \tilde{\boldsymbol{b}}_1\tilde{\boldsymbol{b}}_2^T + \tilde{\boldsymbol{b}}_2\tilde{\boldsymbol{b}}_1^T$, which implies that $\tilde{\boldsymbol{B}} = \pm\boldsymbol{B}$ and $\tilde{\boldsymbol{\Psi}} = \boldsymbol{\Psi}$. Thus these parameters are identifiable, at least given an adequate range of x-values.

2.2. Random-effects representation

The covariance regression model also has an interpretation as a type of random-effects model. Consider a model for observed data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ of the

form

$$\boldsymbol{y}_i = \boldsymbol{\mu}_{\mathrm{x}_i} + \gamma_i \times \boldsymbol{B}\boldsymbol{x}_i + \boldsymbol{\epsilon}_i, \tag{2.4}$$
$$\mathrm{E}\left[\boldsymbol{\epsilon}_i\right] = \boldsymbol{0} \ , \ \mathrm{Cov}\left[\boldsymbol{\epsilon}_i\right] \ = \ \boldsymbol{\Psi},$$
$$\mathrm{E}\left[\gamma_i\right] = 0 \ , \ \mathrm{Var}\left[\gamma_i\right] \ = \ 1 \ , \ \mathrm{E}\left[\gamma_i \times \boldsymbol{\epsilon}_i\right] \ = \ \boldsymbol{0}.$$

The resulting covariance matrix for $\boldsymbol{y}_i$ given $\boldsymbol{x}_i$ is then

$$\begin{aligned}
\mathrm{E}\left[(\boldsymbol{y}_i - \boldsymbol{\mu}_{\mathrm{x}_i})(\boldsymbol{y}_i - \boldsymbol{\mu}_{\mathrm{x}_i})^T\right] &= \mathrm{E}\left[\gamma_i^2 \boldsymbol{B}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{B}^T + \gamma_i(\boldsymbol{B}\boldsymbol{x}_i\boldsymbol{\epsilon}_i^T + \boldsymbol{\epsilon}_i\boldsymbol{x}_i^T\boldsymbol{B}^T) + \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T\right] \\
&= \boldsymbol{B}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{B}^T + \boldsymbol{\Psi} \\
&= \boldsymbol{\Sigma}_{\mathrm{x}_i}.
\end{aligned}$$

The model given in (2.4) can be thought of as a factor analysis model in which the latent factor for unit i is restricted to be a multiple of unit's explanatory vector $\boldsymbol{x}_i$. To see how this impacts the variance, let $\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_p\}$ be the rows of $\boldsymbol{B}$. Then (2.4) can be expressed as

$$\begin{pmatrix} y_{i,1} - \mu_{\mathrm{x}_i,1} \\ \vdots \\ y_{i,p} - \mu_{\mathrm{x}_i,p} \end{pmatrix} = \gamma_i \times \begin{pmatrix} \boldsymbol{b}_1^T\boldsymbol{x}_i \\ \vdots \\ \boldsymbol{b}_p^T\boldsymbol{x}_i \end{pmatrix} + \begin{pmatrix} \epsilon_{i,1} \\ \vdots \\ \epsilon_{i,p} \end{pmatrix}. \tag{2.5}$$

We can interpret γ_i as describing additional unit-level variability beyond that represented by $\boldsymbol{\epsilon}_i$. The vectors $\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_p\}$ describe how this additional variability is manifested across the p different response variables. Small values of $\boldsymbol{b}_j$ indicate little heteroscedasticity in y_j as a function of $\boldsymbol{x}$. Vectors $\boldsymbol{b}_j$ and $\boldsymbol{b}_k$ being either in the same or opposite direction indicates that y_j and y_k become more positively or more negatively correlated, respectively, as their variances increase.

With the random-effects representation, the covariance regression model can be seen as similar in spirit to the random-effects model for longitudinal data discussed in Scott and Handcock (2001). There, the covariance among a set of repeated measurements $\boldsymbol{y}_i$ from a single individual i were modeled as $\boldsymbol{y}_i = \boldsymbol{\mu}_i + \gamma_i\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$, where $\boldsymbol{X}_i$ is an observed design matrix for the repeated measurements, and γ_i is a mean-zero unit variance random effect. In their longitudinal data application, $\boldsymbol{X}_i$ was constructed from a set of basis functions evaluated at the observed time points, and $\boldsymbol{\beta}$ represented unknown weights. This model induces a covariance matrix of $\boldsymbol{X}_i\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{X}_i^T + \mathrm{Cov}[\boldsymbol{\epsilon}_i]$ among the observations common to an individual. For the problem we are considering, where the explanatory variables are shared among all p observations of a given unit (i.e. the rows of $\boldsymbol{X}_i$ are identical and equal to $\boldsymbol{x}_i$), the covariance matrix induced by Scott and Handcock's model reduces to $(\boldsymbol{x}_i^T\boldsymbol{\beta})^2\boldsymbol{1}\boldsymbol{1}^T + \mathrm{Cov}[\boldsymbol{\epsilon}_i]$, which is much more restrictive than (2.4).

The family of linear regression models is closed under linear transformations of the outcome and explanatory variables, and the same holds for the covariance regression model, as can be seen as follows: Suppose $\mathrm{E}\,[\boldsymbol{y}|\boldsymbol{x}] = \boldsymbol{A}\boldsymbol{x}$ and $\mathrm{Cov}\,[\boldsymbol{y}|\boldsymbol{x}] = \boldsymbol{B}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{B}^T + \boldsymbol{C}\boldsymbol{C}^T$, where $\boldsymbol{\Psi} = \boldsymbol{C}\boldsymbol{C}^T$ is positive definite. Via the random-effects representation, we can write $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \gamma \times \boldsymbol{B}\boldsymbol{x} + \boldsymbol{C}\boldsymbol{\epsilon}$. Letting $\tilde{\boldsymbol{y}} = \boldsymbol{D}(\boldsymbol{y} - \boldsymbol{e})$ and $\tilde{\boldsymbol{x}} = \boldsymbol{F}(\boldsymbol{x} - \boldsymbol{g})$ for invertible $\boldsymbol{D}$ and $\boldsymbol{F}$, we have

$$\boldsymbol{y} = \boldsymbol{D}^{-1}\tilde{\boldsymbol{y}} + \boldsymbol{e} = \boldsymbol{A}(\boldsymbol{F}^{-1}\tilde{\boldsymbol{x}} + \boldsymbol{g}) + \gamma \times \boldsymbol{B}(\boldsymbol{F}^{-1}\tilde{\boldsymbol{x}} + \boldsymbol{g}) + \boldsymbol{C}\boldsymbol{\epsilon} \;,\; \text{giving}$$
$$\tilde{\boldsymbol{y}} = [\boldsymbol{D}\boldsymbol{A}\boldsymbol{F}^{-1}]\tilde{\boldsymbol{x}} + \gamma \times [\boldsymbol{D}\boldsymbol{B}\boldsymbol{F}^{-1}]\tilde{\boldsymbol{x}} + [\boldsymbol{D}\boldsymbol{C}]\boldsymbol{\epsilon}$$
$$= \tilde{\boldsymbol{A}}\tilde{\boldsymbol{x}} + \gamma \times \tilde{\boldsymbol{B}}\tilde{\boldsymbol{x}} + \tilde{\boldsymbol{C}}\boldsymbol{\epsilon},$$

which is a member of the class of covariance regression models.

3. Parameter Estimation and Inference

In this section we consider parameter estimation based on data $\boldsymbol{Y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$ observed under conditions $\boldsymbol{X} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_n^T)^T$. We assume normal models for all error terms:

$$\gamma_1, \ldots, \gamma_n \overset{\text{i.i.d.}}{\sim} \mathrm{normal}(0, 1), \tag{3.1}$$
$$\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n \overset{\text{i.i.d.}}{\sim} \mathrm{multivariate\ normal}(\boldsymbol{0}, \boldsymbol{\Psi}),$$
$$\boldsymbol{y}_i = \boldsymbol{\mu}_{\mathrm{x}_i} + \gamma_i \times \boldsymbol{B}\boldsymbol{x}_i + \boldsymbol{\epsilon}_i.$$

Let $\boldsymbol{E} = (\boldsymbol{e}_1^T, \ldots, \boldsymbol{e}_n^T)^T$ be the matrix of residuals for a given mean function $\{\boldsymbol{\mu}_{\mathrm{x}}, \boldsymbol{x} \in \mathcal{X}\}$. The log-likelihood of the covariance parameters $(\boldsymbol{B}, \boldsymbol{\Psi})$ based on $\boldsymbol{E}$ and $\boldsymbol{X}$ is

$$l(\boldsymbol{\Psi}, \boldsymbol{B} : \boldsymbol{E}, \boldsymbol{X}) = c - \frac{1}{2}\sum_i \log|\boldsymbol{\Psi} + \boldsymbol{B}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{B}| - \frac{1}{2}\sum_i \mathrm{tr}[(\boldsymbol{\Psi} + \boldsymbol{B}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{B}^T)^{-1}\boldsymbol{e}_i\boldsymbol{e}_i^T].$$
$$\tag{3.2}$$

After some algebra, it can be shown that the maximum likelihood estimates of $\boldsymbol{\Psi}$ and $\boldsymbol{B}$ satisfy the following equations:

$$\sum_i \hat{\boldsymbol{\Sigma}}_{\mathrm{x}_i}^{-1} = \sum_i \hat{\boldsymbol{\Sigma}}_{\mathrm{x}_i}^{-1}\boldsymbol{e}_i\boldsymbol{e}_i^T\hat{\boldsymbol{\Sigma}}_{\mathrm{x}_i}^{-1},$$
$$\sum_i \hat{\boldsymbol{\Sigma}}_{\mathrm{x}_i}^{-1}\hat{\boldsymbol{B}}\boldsymbol{x}_i\boldsymbol{x}_i^T = \sum_i \hat{\boldsymbol{\Sigma}}_{\mathrm{x}_i}^{-1}\boldsymbol{e}_i\boldsymbol{e}_i^T\hat{\boldsymbol{\Sigma}}_{\mathrm{x}_i}^{-1}\hat{\boldsymbol{B}}\boldsymbol{x}_i\boldsymbol{x}_i^T,$$

where $\hat{\boldsymbol{\Sigma}}_{\mathrm{x}} = \hat{\boldsymbol{\Psi}} + \hat{\boldsymbol{B}}\boldsymbol{x}\boldsymbol{x}^T\hat{\boldsymbol{B}}^T$. While not providing closed-form expressions for $\hat{\boldsymbol{\Psi}}$ and $\hat{\boldsymbol{B}}$, these equations indicate that the MLEs give a covariance function $\hat{\boldsymbol{\Sigma}}_{\mathrm{x}_i}^{-1}$ that, loosely speaking, acts "on average" as a pseudo-inverse for $\boldsymbol{e}_i\boldsymbol{e}_i^T$.

While direct maximization of (3.2) is challenging, the random-effects representation of the model allows for parameter estimation via simple iterative

methods. In particular, maximum likelihood estimation via the EM algorithm is straightforward, as is Bayesian estimation using a Gibbs sampler to approximate the posterior distribution $p(\boldsymbol{\Psi}, \boldsymbol{B} | \boldsymbol{Y}, \boldsymbol{X})$. Both of these methods rely on the conditional distribution of $\{\gamma_1, \ldots, \gamma_n\}$ given $\{\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\Psi}, \boldsymbol{B}\}$. Straightforward calculations give

$$\{\gamma_i | \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\Psi}, \boldsymbol{B}\} \sim \text{normal}(m_i, v_i) \text{ , where}$$
$$v_i = (1 + \boldsymbol{x}_i^T \boldsymbol{B}^T \boldsymbol{\Psi}^{-1} \boldsymbol{B} \boldsymbol{x}_i)^{-1},$$
$$m_i = v_i (\boldsymbol{y}_i - \boldsymbol{\mu}_{\text{x}_i})^T \boldsymbol{\Psi}^{-1} \boldsymbol{B} \boldsymbol{x}_i.$$

A wide variety of modeling options exist for the mean function $\{\boldsymbol{\mu}_{\text{x}} : \boldsymbol{x} \in \mathcal{X}\}$. For ease of presentation, in the rest of this section we assume that the mean function is linear, i.e. $\boldsymbol{\mu}_{\text{x}} = \boldsymbol{A}\boldsymbol{x}$, using the same regressors as the covariance function. This assumption is not necessary, and in Section 6 an analysis is performed where the regressors for the mean and variance functions are distinct.

3.1. Estimation with the EM-algorithm

The EM-algorithm proceeds by iteratively maximizing the expected value of the complete data log-likelihood, $l(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\Psi}) = \log p(\boldsymbol{Y} | \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\Psi}, \boldsymbol{X}, \boldsymbol{\gamma})$, which is simply obtained from the multivariate normal density

$$-2l(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\Psi})$$
$$= np \log(2\pi) + n \log |\boldsymbol{\Psi}| + \sum_{i=1}^{n} (\boldsymbol{y}_i - [\boldsymbol{A} + \gamma_i \boldsymbol{B}]\boldsymbol{x}_i)^T \boldsymbol{\Psi}^{-1} (\boldsymbol{y}_i - [\boldsymbol{A} + \gamma_i \boldsymbol{B}]\boldsymbol{x}_i). \quad (3.3)$$

Given current estimates $(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{\Psi}})$ of $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\Psi})$, one step of the EM algorithm proceeds as follows: First, $m_i = \text{E}[\gamma_i | \hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{\Psi}}, \boldsymbol{y}_i]$ and $v_i = \text{Var}[\gamma_i | \hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{\Psi}}, \boldsymbol{y}_i]$ are computed and plugged into the likelihood (3.3), giving

$$-2\text{E}[l(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\Psi}) | \hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{\Psi}}]$$
$$= np \log(2\pi) + n \log |\boldsymbol{\Psi}| + \sum_{i=1}^{n} \text{E}[(\hat{\boldsymbol{e}}_i - \gamma_i \boldsymbol{B} \boldsymbol{x}_i)^T \boldsymbol{A}^{-1} (\hat{\boldsymbol{e}}_i - \gamma_i \boldsymbol{B} \boldsymbol{x}_i) | \hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{\Psi}}]$$

where $\hat{\boldsymbol{e}}_i = \boldsymbol{y}_i - \hat{\boldsymbol{A}}\boldsymbol{x}_i$ and

$$\text{E}[(\hat{\boldsymbol{e}}_i - \gamma_i \boldsymbol{B} \boldsymbol{x}_i)^T \boldsymbol{\Psi}^{-1} (\hat{\boldsymbol{e}}_i - \gamma_i \boldsymbol{B} \boldsymbol{x}_i) | \hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{\Psi}}]$$
$$= (\hat{\boldsymbol{e}}_i - m_i \boldsymbol{B} \boldsymbol{x}_i)^T \boldsymbol{\Psi}^{-1} (\hat{\boldsymbol{e}}_i - m_i \boldsymbol{B} \boldsymbol{x}_i) + v_i \boldsymbol{x}_i^T \boldsymbol{B}^T \boldsymbol{\Psi}^{-1} \boldsymbol{B} \boldsymbol{x}_i$$
$$= (\hat{\boldsymbol{e}}_i - m_i \boldsymbol{B} \boldsymbol{x}_i)^T \boldsymbol{\Psi}^{-1} (\hat{\boldsymbol{e}}_i - m_i \boldsymbol{B} \boldsymbol{x}_i) + s_i \boldsymbol{x}_i^T \boldsymbol{B}^T \boldsymbol{\Psi}^{-1} \boldsymbol{B} \boldsymbol{x}_i s_i,$$

with $s_i = v_i^{1/2}$. To maximize the expected log-likelihood, first construct the $2n \times 2q$ matrix $\tilde{\boldsymbol{X}}$ whose ith row is $(\boldsymbol{x}_i^T, m_i \boldsymbol{x}_i^T)$ and whose $(n + i)$th row is

$(\mathbf{0}_q^T, s_i \mathbf{x}_i^T)$, and let $\tilde{\mathbf{Y}}$ be the $2n \times p$ matrix given by $(\mathbf{Y}^T, \mathbf{0}_{n \times p}^T)^T$. The expected value of the complete data log-likelihood can then be written as

$$-2\mathrm{E}\left[l(\mathbf{A}, \mathbf{B}, \mathbf{\Psi})|(\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{\Psi}})\right] - np\log(2\pi) = n\log|\mathbf{\Psi}| + \mathrm{tr}([\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\mathbf{C}^T][\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\mathbf{C}^T]^T\mathbf{\Psi}^{-1})$$

with $\mathbf{C} = (\mathbf{A}, \mathbf{B})$. The next step of the EM algorithm obtains the new values $(\check{\mathbf{A}}, \check{\mathbf{B}}, \check{\mathbf{\Psi}})$ as the maximizers of this expected log-likelihood. Since the expected log-likelihood has the same form as the log-likelihood for normal multivariate regression, $(\check{\mathbf{A}}, \check{\mathbf{B}}, \check{\mathbf{\Psi}})$ are given by

$$(\check{\mathbf{A}}, \check{\mathbf{B}}) = \check{\mathbf{C}} = \tilde{\mathbf{Y}}^T \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1},$$

$$\check{\mathbf{\Psi}} = \frac{1}{n}(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\check{\mathbf{C}}^T)^T(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\check{\mathbf{C}}^T).$$

The procedure is then repeated until a desired convergence criterion has been met.

3.2. Confidence intervals via expected information

Approximate confidence intervals for model parameters can be provided by Wald intervals, i.e. the MLEs plus or minus a multiple of the standard errors. Standard errors can be obtained from the inverse of the expected information matrix evaluated at the MLEs. The log-likelihood given an observation $\mathbf{y}$ is $l(\mathbf{B}, \mathbf{\Psi} : \mathbf{y}) = \log p(\mathbf{y}|\mathbf{\Sigma}) = -(p\log 2\pi + \log|\mathbf{\Sigma}| + \mathbf{e}^T\mathbf{\Sigma}^{-1}\mathbf{e})/2$, where $\mathbf{e} = \mathbf{y} - \mathbf{A}\mathbf{x}$ and $\mathbf{\Sigma} = \mathbf{\Psi} + \mathbf{B}\mathbf{x}\mathbf{x}^T\mathbf{B}^T$. Likelihood derivatives with respect to $\mathbf{A}$ and $\mathbf{B}$ are

$$\dot{l}_{\mathrm{A}} = \frac{\partial l(\mathbf{A}, \mathbf{B}, \mathbf{\Psi} : \mathbf{y})}{\partial \mathbf{A}} = \mathbf{\Sigma}^{-1}\mathbf{e}\mathbf{x}^T,$$

$$\dot{l}_{\mathrm{B}} = \frac{\partial l(\mathbf{A}, \mathbf{B}, \mathbf{\Psi} : \mathbf{y})}{\partial \mathbf{B}} = -\frac{(\partial \log|\mathbf{\Sigma}|/\partial \mathbf{B} + \partial \mathbf{e}^T\mathbf{\Sigma}^{-1}\mathbf{e}/\partial \mathbf{B})}{2}$$

$$= -\mathbf{\Sigma}^{-1}\mathbf{B}\mathbf{x}\mathbf{x}^T + \mathbf{\Sigma}^{-1}\mathbf{e}\mathbf{e}^T\mathbf{\Sigma}^{-1}\mathbf{B}\mathbf{x}\mathbf{x}^T$$

$$= \mathbf{H}_{\mathrm{z}}\mathbf{B}\mathbf{x}\mathbf{x}^T,$$

where $\mathbf{H}_{\mathrm{z}} = \mathbf{\Sigma}^{-1/2}(\mathbf{z}\mathbf{z}^T - \mathbf{I})\mathbf{\Sigma}^{-1/2}$ and $\mathbf{z} = \mathbf{\Sigma}^{-1/2}\mathbf{e}$. The derivative with respect to $\mathbf{\Psi}$ is more complicated, as the $p \times p$ matrix $\mathbf{\Psi}$ has only $p(p+1)/2$ free parameters. Following McCulloch (1982), we let $\boldsymbol{\psi} = \mathrm{vech}\,\mathbf{\Psi}$ be the $p(p+1)/2$ vector of unique elements of $\mathbf{\Psi}$. As described there, derivatives of functions with respect to $\boldsymbol{\psi}$ can be obtained as a linear transformation of derivatives with respect to $\mathbf{\Psi}$, obtained by ignoring the symmetry in $\mathbf{\Psi}$:

$$\dot{l}_{\Psi} = \frac{\partial l(\mathbf{A}, \mathbf{B}, \mathbf{\Psi} : \mathbf{y})}{\partial \mathbf{\Psi}} = -\frac{(\mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}\mathbf{e}\mathbf{e}^T\mathbf{\Sigma}^{-1})}{2}$$

$$= \frac{1}{2}\mathbf{\Sigma}^{-1/2}(\mathbf{z}\mathbf{z}^T - \mathbf{I})\mathbf{\Sigma}^{-1/2} = \frac{\mathbf{H}_z}{2},$$

$$\dot{l}_{\psi} = \frac{\partial l(\mathbf{A}, \mathbf{B}, \boldsymbol{\psi} : \mathbf{y})}{\partial \boldsymbol{\psi}} = \mathbf{G}^T\mathrm{vec}\,\dot{l}_{\Psi} = \mathbf{G}^T\mathrm{vec}\,\frac{\mathbf{H}_z}{2},$$

where $\boldsymbol{G}$ is the matrix such that $\operatorname{vec} \boldsymbol{X} = \boldsymbol{G}\operatorname{vech} \boldsymbol{X}$, as defined in Henderson and Searle (1979). Letting $\boldsymbol{a} = \operatorname{vec} \boldsymbol{A}$, $\dot{l}_{\mathrm{a}} = \operatorname{vec} \dot{l}_A$ and defining $\boldsymbol{b}$ and $\dot{l}_{\mathrm{b}}$ similarly, the expected information is

$$\mathcal{I}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\psi} : \boldsymbol{x}) = \mathrm{E}_{\mathrm{a,b},\psi} \begin{bmatrix} \dot{l}_{\mathrm{a}}\dot{l}_{\mathrm{a}}^T & \dot{l}_{\mathrm{a}}\dot{l}_{\mathrm{b}}^T & \dot{l}_{\mathrm{a}}\dot{l}_{\psi}^T \\ \dot{l}_{\mathrm{b}}\dot{l}_{\mathrm{a}}^T & \dot{l}_{\mathrm{b}}\dot{l}_{\mathrm{b}}^T & \dot{l}_{\mathrm{b}}\dot{l}_{\psi}^T \\ \dot{l}_{\psi}\dot{l}_{\mathrm{a}}^T & \dot{l}_{\psi}\dot{l}_{\mathrm{b}}^T & \dot{l}_{\psi}\dot{l}_{\psi}^T \end{bmatrix} \equiv \begin{pmatrix} \mathcal{I}_{\mathrm{aa}} & \mathcal{I}_{\mathrm{ab}} & \mathcal{I}_{\mathrm{a}\psi} \\ \mathcal{I}_{\mathrm{ab}}^T & \mathcal{I}_{\mathrm{bb}} & \mathcal{I}_{\mathrm{b}\psi} \\ \mathcal{I}_{\mathrm{a}\psi}^T & \mathcal{I}_{\mathrm{b}\psi}^T & \mathcal{I}_{\psi\psi} \end{pmatrix}.$$

The submatrices $\mathcal{I}_{\mathrm{ab}}$ and $\mathcal{I}_{\mathrm{a}\psi}$ can be expressed as expectations of mixed third moments of independent standard normal variables, and so are both zero. Calculation of $\mathcal{I}_{\mathrm{bb}}$ $\mathcal{I}_{\mathrm{b}\psi}$ and $\mathcal{I}_{\psi\psi}$ involve expectations of $(\operatorname{vec} \boldsymbol{H}_z)(\operatorname{vec} \boldsymbol{H}_z)^T$, which has expected value $(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})(\boldsymbol{I}_{p^2} + \boldsymbol{K}_{p,p})$, where $\boldsymbol{K}_{p,p}$ is the commutation matrix described in Magnus and Neudecker (1979). Straightforward calculations show that

$$\mathcal{I}_{\mathrm{aa}} = (\boldsymbol{x}\boldsymbol{x}^T) \otimes \boldsymbol{\Sigma}^{-1},$$
$$\mathcal{I}_{\mathrm{bb}} = (\boldsymbol{x}\boldsymbol{x}^T \boldsymbol{B}^T \otimes \boldsymbol{I}_p)(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})(\boldsymbol{I}_{p^2} + \boldsymbol{K}_{p,p})(\boldsymbol{B}\boldsymbol{x}\boldsymbol{x}^T \otimes \boldsymbol{I}_p),$$
$$\mathcal{I}_{\mathrm{b}\psi} = (\boldsymbol{x}\boldsymbol{x}^T \boldsymbol{B}^T \otimes \boldsymbol{I}_p)(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\boldsymbol{G},$$
$$\mathcal{I}_{\psi\psi} = \frac{1}{2}\boldsymbol{G}^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\boldsymbol{G}.$$

The expected information contained in observations to be made at $\boldsymbol{x}$-values $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ is then $\mathcal{I}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\psi} : \boldsymbol{X}) = \sum_{i=1}^{n} \mathcal{I}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\psi} : \boldsymbol{x}_i)$. Plugging the MLEs into the inverse of this matrix gives an estimate of their variance, $\hat{\mathrm{Var}}[(\hat{\boldsymbol{a}}^T, \hat{\boldsymbol{b}}^T, \hat{\boldsymbol{\psi}}^T)^T] = \mathcal{I}^{-1}(\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{\psi}} : \boldsymbol{X})$. Approximate confidence intervals for model parameters based on this variance estimate are explored in the simulation study in the next section.

3.3. Posterior approximation with the Gibbs sampler

A Bayesian analysis provides estimates and confidence intervals for arbitrary functions of the parameters, as well as a simple way of making predictive inference for future observations. Given a prior distribution $p(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\Psi})$, inference is based on the joint posterior distribution, $p(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\Psi}|\boldsymbol{Y}, \boldsymbol{X}) \propto p(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\Psi}) \times p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\Psi})$. While this posterior distribution is not available in closed-form, a Monte Carlo approximation to the joint posterior distribution of $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\Psi})$ is available via Gibbs sampling. Using the random-effects representation of the model in (3.1), the Gibbs sampler constructs a Markov chain in $\{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\Psi}, \gamma_1, \ldots, \gamma_n\}$ whose stationary distribution is the joint posterior distribution of these quantities.

Calculations are facilitated by the use of a semi-conjugate prior distribution for $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\Psi})$, in which $p(\boldsymbol{\Psi})$ is an inverse-Wishart$(\boldsymbol{\Psi}_0^{-1}, \nu_0)$ distribution having

expectation $\boldsymbol{\Psi}_0/(\nu_0 - p - 1)$ and $\boldsymbol{C} = (\boldsymbol{A}, \boldsymbol{B})$ has a matrix normal prior distribution, $\{\boldsymbol{C}|\boldsymbol{\Psi}\} \sim$ matrix normal$(\boldsymbol{C}_0, \boldsymbol{\Psi}, \boldsymbol{V}_0)$. The Gibbs sampler proceeds by iteratively sampling $\boldsymbol{C} = (\boldsymbol{A}, \boldsymbol{B})$, $\boldsymbol{\Psi}$ and $\{\gamma_1, \ldots, \gamma_n\}$ from their full conditional distributions. One iteration of a Gibbs sampler is as follows:

1. Sample $\gamma_i \sim$ normal(m_i, v_i) for each $i \in \{1, \ldots, n\}$, where

$$v_i = (1 + \boldsymbol{x}_i^T \boldsymbol{B}^T \boldsymbol{\Psi}^{-1} \boldsymbol{B} \boldsymbol{x}_i)^{-1} \; ;$$
$$m_i = v_i \boldsymbol{x}_i^T \boldsymbol{\Psi}^{-1} \boldsymbol{B} (\boldsymbol{y}_i - \boldsymbol{A} \boldsymbol{x}_i).$$

2. Sample $(\boldsymbol{C}, \boldsymbol{\Psi}) \sim p(\boldsymbol{C}, \boldsymbol{\Psi}|\boldsymbol{Y}, \boldsymbol{X}, \gamma_1, \ldots, \gamma_n)$ as follows:

   (a) sample $\boldsymbol{\Psi} \sim$ inverse-Wishart$(\boldsymbol{\Psi}_n^{-1}, \nu_0 + n)$, and

   (b) sample $\boldsymbol{C} \sim$ matrix normal$(\boldsymbol{C}_n, \boldsymbol{\Psi}, [\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma + \boldsymbol{V}_0^{-1}]^{-1})$, where

   $\boldsymbol{X}_\gamma = (\boldsymbol{X}, \boldsymbol{\Gamma} \boldsymbol{X})$, with $\boldsymbol{\Gamma} = \text{diag}(\gamma_1, \ldots, \gamma_n)$,
   $\boldsymbol{C}_n = (\boldsymbol{Y}^T \boldsymbol{X}_\gamma + \boldsymbol{C}_0 \boldsymbol{V}_0^{-1})(\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma + \boldsymbol{V}_0^{-1})^{-1}$, and
   $\boldsymbol{\Psi}_n = \boldsymbol{\Psi}_0 + (\boldsymbol{Y} - \boldsymbol{X}_\gamma \boldsymbol{C}_n)^T (\boldsymbol{Y} - \boldsymbol{X}_\gamma \boldsymbol{C}_n) + (\boldsymbol{C}_n - \boldsymbol{C}_0)^T \boldsymbol{V}_0^{-1} (\boldsymbol{C}_n - \boldsymbol{C}_0)$.

In the absence of strong prior information, default values for the prior parameters $\{\boldsymbol{C}_0, \boldsymbol{V}_0, \boldsymbol{\Psi}_0, \nu_0\}$ can be based on other considerations. In normal regression for example, Zellner (1986) suggests a "g-prior" which makes the Bayes procedure invariant to linear transformations of the design matrix $\boldsymbol{X}$. An analogous result can be obtained in the covariance regression model by selecting $\boldsymbol{C}_0 = \boldsymbol{0}$ and $\boldsymbol{V}_0$ to be block diagonal, consisting of two $q \times q$ blocks both proportional to $(\boldsymbol{X}^T \boldsymbol{X})^{-1}$, i.e. the prior precision of $\boldsymbol{C}$ is related to the precision given by the observed design matrix. Often the proportionality constant is set equal to the sample size n so that, roughly speaking, the information in the prior distribution is equivalent to that contained in one observation. Such choices lead to what Kass and Wasserman (1995) call a "unit-information" prior distribution, which weakly centers the prior distribution around an estimate based on the data. For example, setting $\nu_0 = p + 2$ and $\boldsymbol{\Psi}_0$ equal to the sample covariance matrix of $\boldsymbol{Y}$ weakly centers the prior distribution of $\boldsymbol{\Psi}$ around a "homoscedastic" sample estimate.

4. Simulation Study

In this section we present a simulation study to evaluate the MLEs obtained from the proposed covariance regression model. In addition to evaluating the ability of the model to describe heteroscedasticity, we also evaluate the effect of heteroscedasticity on the estimation of the mean function.

As is well known, the ordinary least squares (OLS) estimator of a matrix of multivariate regression coefficients has a higher mean squared error (MSE) than the generalized least squares (GLS) estimator in the presence of known

heteroscedasticity. The OLS estimator, or equivalently the MLE assuming a homoscedastic normal model, is $\hat{\boldsymbol{A}} = \boldsymbol{Y}^T \boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{X})^{-1}$ or equivalently, $\hat{\boldsymbol{a}} = \text{vec}(\hat{\boldsymbol{A}}) = [(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \otimes \boldsymbol{I}_p] \boldsymbol{y}$, where $\boldsymbol{y} = \text{vec}\,\boldsymbol{Y}$. The variability of the estimator around $\boldsymbol{a} = \text{vec}\,\boldsymbol{A}$ is given by

$$\text{Cov}\,[\hat{\boldsymbol{a}}] = [(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \otimes \boldsymbol{I}_p] \boldsymbol{\Omega} [\boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{X})^{-1} \otimes \boldsymbol{I}_p],$$

where $\boldsymbol{\Omega}$ is the $np \times np$ covariance matrix $\boldsymbol{y}$. If the rows of $\boldsymbol{Y}$ are independent with constant variance $\boldsymbol{\Sigma}$, then $\boldsymbol{\Omega} = \boldsymbol{I}_n \otimes \boldsymbol{\Sigma}$, $\text{Cov}\,[\hat{\boldsymbol{a}}]$ reduces to $(\boldsymbol{X}^T \boldsymbol{X})^{-1} \otimes \boldsymbol{\Sigma}$ and $\hat{\boldsymbol{a}}$ is the best linear unbiased estimator of $\text{vec}\,\boldsymbol{A}$ (see, for example, Mardia, Kent, and Bibby (1979, Sec. 6.6)). If the rows of $\boldsymbol{Y}$ are independent but with known non-constant covariance matrices $\{\boldsymbol{\Sigma}_i, i = 1, \ldots, n\}$ then the GLS estimator $\hat{\boldsymbol{a}}_{\text{GLS}}$ is more precise than the OLS estimator in the sense that $\text{Cov}\,[\hat{\boldsymbol{a}}] = \text{Cov}\,[\hat{\boldsymbol{a}}_{\text{GLS}}] + \boldsymbol{H}$, where $\boldsymbol{H}$ is positive definite.

In general, the exact nature of the heteroscedasticity will be unknown, but if it can be well-estimated then we expect an estimator that accounts for heteroscedasticity to be more efficient in terms of MSE. The precision of covariance regression parameter estimates $\hat{\boldsymbol{B}}$ and $\hat{\boldsymbol{\Psi}}$ can be described by the expected information matrix given in the previous section, but how this translates into improved estimation for the mean is difficult to describe with a simple formula. Instead, we examine the potential for improved estimation of $\boldsymbol{A}$ with a simulation study in the simple case of $p = q = 2$, for a variety of sample sizes and scales of the heteroscedasticity. Specifically, we generate samples of size $n \in \{50, 100, 200\}$ from the multivariate normal model with $\text{E}\,[\boldsymbol{y}|\boldsymbol{x}] = \boldsymbol{A}\boldsymbol{x}$ and $\text{Var}\,[\boldsymbol{y}|\boldsymbol{x}] = \boldsymbol{\Psi} + \boldsymbol{B}\boldsymbol{x}\boldsymbol{x}^T \boldsymbol{B}^T$, where $\boldsymbol{x}^T = (1, x)^T$, $\boldsymbol{A} = [(1, -1)^T, (-1, 1)^T]$ and

$$\boldsymbol{B} = \frac{w}{w+1} \times \boldsymbol{B}_0, \quad \boldsymbol{\Psi} = \frac{1}{w+1} \times \boldsymbol{\Psi}_0, \quad \boldsymbol{B}_0 = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}, \quad \boldsymbol{\Psi}_0 = \boldsymbol{B}_0 \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{3} \end{pmatrix} \boldsymbol{B}_0^T,$$
(4.1)

where we consider $w \in \{0, 1/3, 1, 3\}$. Note that if x is uniformly distributed on $[-1, 1]$ then the expected value of $\boldsymbol{B}_0 \boldsymbol{x}\boldsymbol{x}^T \boldsymbol{B}_0^T$ is equal to $\boldsymbol{\Psi}_0$. As a result, the average value of $\boldsymbol{\Psi} + \boldsymbol{B}\boldsymbol{x}\boldsymbol{x}^T \boldsymbol{B}^T$, averaged across uniformly distributed design points, is constant across values of w. The resulting mean and variances functions for $x \in (-1, 1)$ and $w \in \{0, 1/3, 1, 3\}$ are shown graphically in Figure 2. The means for y_1 and y_2 are decreasing and increasing, respectively, with x, whereas for $w \neq 0$ the variances are increasing and decreasing, respectively.

For each combination of n and w, 1000 datasets were generated by simulating x-values from the uniform(-1,1) distribution, then simulating $\boldsymbol{y}$ conditional on $\boldsymbol{x} = (1, x)^T$ from the model given by (4.1). The EM-algorithm described in Section 3.1 was used to obtain parameter estimates of the model parameters. In terms of summarizing results, we first evaluate the covariance regression
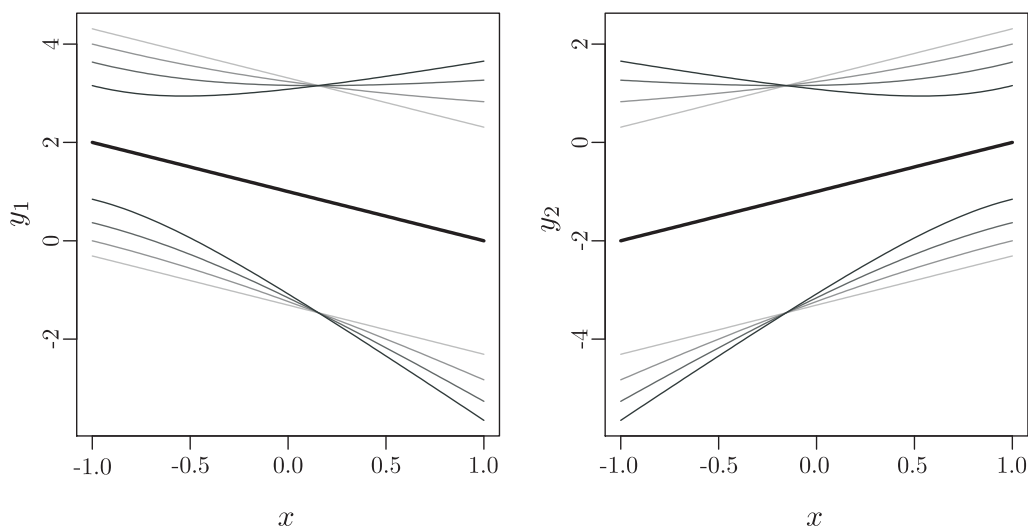
Figure 2. Population mean and variance functions for the simulation study. The black line is the mean function, and the gray lines give the mean plus and minus two standard deviations under $w \in \{0, 1/3, 1, 3\}$.

model in terms of its potential for improved estimation of the mean function. The first set of four columns of Table 1 compares the ratio of $\mathrm{E}\left[||\boldsymbol{A} - \hat{\boldsymbol{A}}_{\mathrm{OLS}}||^2\right]$ to $\mathrm{E}\left[||\boldsymbol{A} - \hat{\boldsymbol{A}}_{\mathrm{CVR}}||^2\right]$, the former being the MSE of the OLS estimate and the latter the MSE of the MLE from the covariance regression (CVR) model. Not surprisingly, when the sample size is low ($n = 50$) and there is little or no heteroscedasticity ($w \in \{0, 1/3\}$), the OLS estimator slightly outperforms the overly complex CVR estimator. However, as the sample size increases the CVR estimator improves to roughly match the OLS estimator in terms of MSE. In the presence of more substantial heteroscedasticity ($w \in \{1, 3\}$), the CVR estimator outperforms the OLS estimator for each sample size, with the MSE of the OLS estimator being around 40% higher than that of the CVR estimator for the case $w = 3$.

In practical data analysis settings it is often recommended to favor a simple model over a more complex alternative unless there is substantial evidence that the simple model fits poorly. With this in mind, we consider an estimator $\hat{\boldsymbol{A}}_{\mathrm{MS}}$ based on model selection:

1. Perform the level-α likelihood ratio test of $\mathrm{H}_0 : \boldsymbol{B} = \boldsymbol{0}$ versus $\mathrm{H}_1 : \boldsymbol{B} \neq \boldsymbol{0}$

2. Calculate $\hat{\boldsymbol{A}}_{\mathrm{MS}}$ as follows:

    (a) If H_0 is rejected, set $\hat{\boldsymbol{A}}_{\mathrm{MS}} = \hat{\boldsymbol{A}}_{\mathrm{CVR}}$;
    (b) If H_0 is accepted, set $\hat{\boldsymbol{A}}_{\mathrm{MS}} = \hat{\boldsymbol{A}}_{\mathrm{OLS}}$.

Table 1. MSE comparison and power from the simulation study. The sample size is n and the magnitude of the covariance effects is w. The first set of columns gives the ratio of the MSE of the OLS estimator to that from the covariance regression model; the second set of columns gives the estimated power of the likelihood ratio test for heteroscedasticity; the third set of columns gives the relative MSE of the model selected estimator.

| | relative MSE | | | | power | | | | relative MSE | | | |
| | w | | | | w | | | | w | | | |
n	0	1/3	1	3	0	1/3	1	3	0	1/3	1	3
50	0.92	0.93	1.01	1.36	0.083	0.106	0.550	0.993	0.98	0.98	0.98	1.36
100	0.96	0.97	1.06	1.42	0.056	0.121	0.855	1.000	1.00	1.00	1.05	1.42
200	0.99	0.99	1.06	1.41	0.057	0.154	0.996	1.000	1.00	1.00	1.06	1.41

Table 2. Observed coverage of 95% Wald confidence intervals, for the case $w = 1$.

n	$b_{1,1}$	$b_{1,2}$	$b_{2,1}$	$b_{2,2}$	$\psi_{1,1}$	$\psi_{1,2}$	$\psi_{2,2}$
50	0.89	0.88	0.90	0.89	0.88	0.94	0.87
100	0.92	0.92	0.93	0.93	0.93	0.96	0.93
200	0.94	0.95	0.94	0.93	0.95	0.97	0.96

The asymptotic null distribution of the -2 log-likelihood ratio statistic is a χ^2 distribution with $p \times q$ degrees of freedom. The second set of four columns in Table 1 describes the estimated finite-sample level and power of this test when $\alpha = 0.05$. The level of the test can be obtained from the first column of the set, as $w = 0$ corresponds to the null hypothesis being true. The level is somewhat liberal when $n = 50$, but is closer to the nominal level for the larger sample sizes (note that power estimates here are subject to Monte Carlo error, and that 95% Wald intervals for the actual levels contain 0.05 for both $n = 100$ and $n = 200$). As expected, the power of the test increases as either the sample size or the amount of heteroscedasticity increase. The MSE of $\hat{A}_{\mathrm{OLS}}$ relative to $\hat{A}_{\mathrm{MS}}$, given in the third set of four columns, shows that the model selected estimate $\hat{A}_{\mathrm{MS}}$ performs quite well, having essentially the same MSE as the OLS estimate when there is little or no heteroscedasticity, but having the same MSE as the CVR estimate in the presence of more substantial heteroscedasticity.

Beyond improved estimation of the regression matrix $\boldsymbol{A}$, the covariance regression model can be used to describe patterns of non-constant covariance in the data. If the likelihood ratio test described above rejects the constant covariance model, it will often be of interest to obtain point estimates and confidence intervals for $\boldsymbol{B}$ and $\boldsymbol{\Psi}$. In terms of point estimates, recall that the sign of $\boldsymbol{B}$ is not identifiable, with $\boldsymbol{B}$ and $-\boldsymbol{B}$ corresponding to the same covariance function. To facilitate a description of the simulation results, estimates of $\boldsymbol{B}$ were
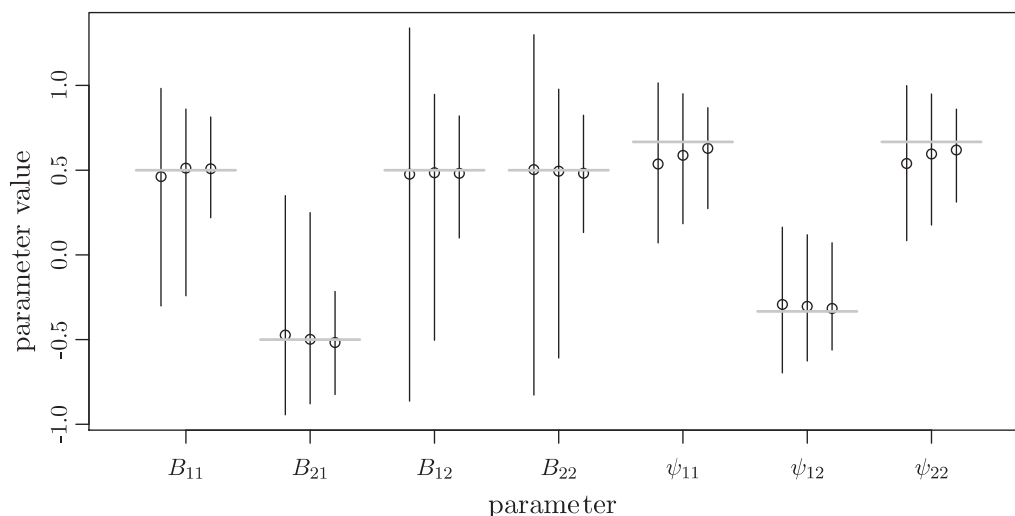
Figure 3. Sampling distribution quantiles of the covariance regression parameter estimates for the case $w = 1$ and $n \in \{50, 100, 200\}$. Horizontal gray lines are the true parameter values, and vertical lines and dots give the 2.5, 50, and 97.5 percentiles of the sampling distributions for each parameter and sample size, with sample size increasing from left to right for each group of three lines.

processed as follows: Given a parameter value $\check{B}$ from the EM algorithm, the value of $\hat{B}$ was taken to be either $\check{B}$ or $-\check{B}$ depending on which was closer to $B = [(1, -1)^T (1, 1)^T]$.

In the interest of brevity we present detailed results only for the case $w = 1$, as results for other values of w follow similar patterns. Figure 3 shows 2.5%, 50%, and 97.5% quantiles of the empirical distribution of the 1000 $\hat{B}$ and $\hat{\Psi}$-values for the case $w = 1$. Although skewed, the sampling distributions of the point estimates are generally centered around their correct values, becoming more concentrated around the truth as the sample size increases. The skew of the sampling distributions diminishes as the log-likelihood becomes more quadratic with increasing sample size.

Regarding confidence intervals, as described in Section 3.3, an asymptotic approximation to the variance-covariance matrix of $\hat{B}$ and $\hat{\Psi}$ can be obtained by plugging the values of the MLEs into the inverse of the expected information matrix. Approximate confidence intervals for individual parameters can then be constructed with Wald intervals. For example, an approximate 95% confidence interval for $b_{j,k}$ would be $\hat{b}_{j,k} \pm 1.96 \times \text{se}(\hat{b}_{j,k})$, where the standard error $\text{se}(\hat{b}_{j,k})$ is the approximation of the standard deviation of $\hat{b}_{j,k}$ based on the expected information matrix. Table 2 presents empirical coverage probabilities from the

simulation study for the case $w = 1$ (results for other non-zero values of w are similar). The intervals are generally a bit too narrow for the low sample size case $n = 50$, although the coverage rates become closer to the nominal level as the sample size increases.

4.1. Multiple regressors

The proposed covariance regression model may be of particular use when the covariance depends on several explanatory variables in a simple way. For example, consider the case of one continuous regressor x_1 and two binary regressors x_2 and x_3. There are four covariance functions of x_1 in this case, one for each combination of x_2 and x_3. As in the case of mean regression, a useful parsimonious model might assume that the differences between the groups can be parameterized in a relatively simple manner. For example, consider the random effects representation of a covariance regression model with additive effects:

$$y_i = Ax_i + \gamma_i \times Bx_i + \epsilon_i$$
$$Bx_i = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + b_3 x_{i,3},$$

where b_0, b_1, b_2, b_3 are four $p \times 1$ column vectors of B. In particular, suppose $Ax_i = (1, -1)^T + (-1, 1)^T x_{i,1}$, $\text{Cov}[\epsilon_i] = \Psi_0/(w+1)$ where Ψ_0 is as in the first simulation study and

$$B = \frac{w}{w+1} \begin{pmatrix} 1 & 1 & \frac{1}{2} & 1 \\ -1 & 1 & -\frac{1}{2} & -1 \end{pmatrix}.$$

Note that the "baseline" case of $x_2 = x_3 = 0$ corresponds to the covariance function in the previous simulation study, and the effects of non-zero values of x_2 or x_3 are additive on the scale of the random effect γ_i. The four covariance functions of x_1 are plotted in Figure 4 for the case $w = 1/3$.

As in the previous study, we generated 1000 datasets for each value of $w \in \{1/3, 1, 3\}$ with a sample size of $n = 50$ for each of the four groups. We estimated the parameters in the covariance regression model as before using the EM algorithm, and compared the results to those obtained using the kernel estimator described in Yin et al. (2010). This latter approach requires a user-specified kernel bandwidth, which we obtain by cross-validation separately for each simulated dataset.

We compare each estimated covariance function $\hat{\Sigma}_x$ to the truth Σ_x with a discrepancy function given by

$$g(\hat{\Sigma}_x : \Sigma_x) = \sum_{x_1 \in \mathcal{X}} \sum_{x_2=0}^{1} \sum_{x_3=0}^{1} \left(\log |\hat{\Sigma}_x| + \text{tr}(\hat{\Sigma}_x^{-1} \Sigma_x) \right),$$
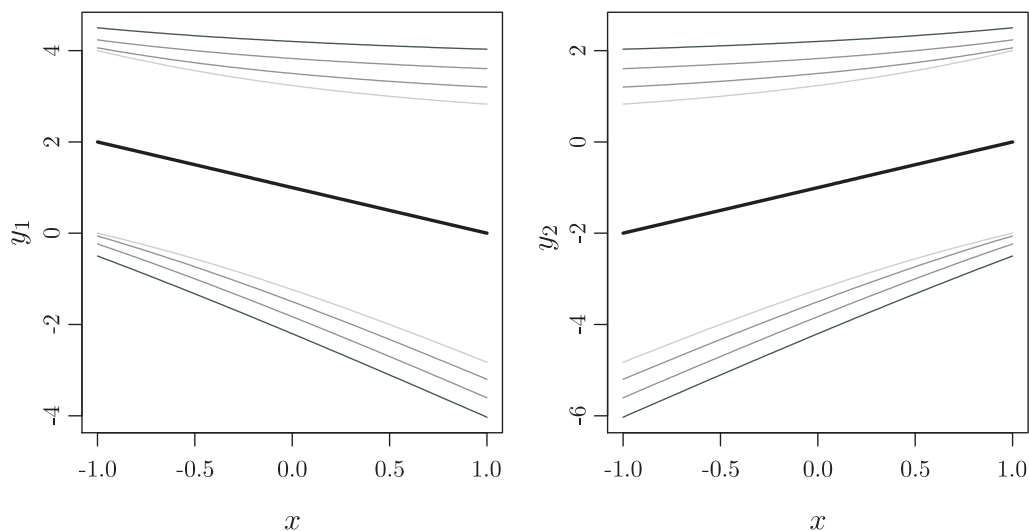
Figure 4. Population mean and variance functions for the second simulation study. The black line is the mean function, and the gray lines give the mean plus and minus two standard deviations under $w = 1/3$. Moving out from the center, the gray lines correspond to $(x_2, x_3) = (0, 0)$, $(1, 0)$, $(0, 1)$ and $(1, 1)$.

where $\mathcal{X}$ is a set of 10 equally-spaced x_1-values between -1 and 1. Note that this discrepancy is minimized by the true covariance function. For the case $w = 1/3$, where the heteroscedasticity is a minimum, the CVR estimator had a lower value of the function g than the kernel density estimator in 73.2% of the simulations. For the $w = 1$ and $w = 3$ cases, the CVR estimator had a lower g-value in 98.5% and 99.5% of the simulations, respectively, with the average difference in g between the two estimators increasing with increasing w. However, the point here is not that the kernel estimator is deficient, but that the kernel estimator cannot take advantage of situations in which the covariance functions across groups are similar in some easily parameterizable way.

5. Higher Rank Models

The model given by (2.1) restricts the difference between $\boldsymbol{\Sigma}_{\mathrm{x}}$ and the baseline matrix $\boldsymbol{\Psi}$ to be a rank-one matrix. To allow for higher-rank deviations, consider the following extension of the random-effects representation in (2.4):

$$\boldsymbol{y} = \boldsymbol{\mu}_{\mathrm{x}} + \gamma \times \boldsymbol{B}\boldsymbol{x} + \phi \times \boldsymbol{C}\boldsymbol{x} + \boldsymbol{\epsilon}, \tag{5.1}$$

where γ and ϕ are mean-zero variance-one random variables, uncorrelated with

each other and with $\boldsymbol{\epsilon}$. Under this model, the covariance of $\boldsymbol{y}$ is

$$\boldsymbol{\Sigma}_{\mathrm{x}} = \boldsymbol{\Psi} + \boldsymbol{B}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{B}^T + \boldsymbol{C}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{C}^T.$$

This model allows the deviation of $\boldsymbol{\Sigma}_{\mathrm{x}}$ from the baseline $\boldsymbol{\Psi}$ to be of rank 2. Additionally, we can interpret the second random effect ϕ as allowing an additional, independent source of heteroscedasticity for the set of the p response variables. Whereas the rank-1 model essentially requires that extreme residuals for one element of $\boldsymbol{y}$ co-occur with extreme residuals of the other elements, the rank-2 model allows for more flexibility, and can allow for heteroscedasticity across individual elements of $\boldsymbol{y}$ without requiring extreme residuals for all of the elements. Further flexibility can be gained by adding additional random effects, allowing the difference between $\boldsymbol{\Sigma}_{\mathrm{x}}$ and the baseline $\boldsymbol{\Psi}$ to be of any desired rank up to and including p.

Identifiability: For a rank-r model with $r > 1$, consider a random-effects representation given by $\boldsymbol{y}_i - \boldsymbol{\mu}_{\mathrm{x}_i} = \sum \gamma_{i,k} \times \boldsymbol{B}^{(k)}\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$. Let $\boldsymbol{B}_1 = (\boldsymbol{b}_1^{(1)}, \ldots, \boldsymbol{b}_1^{(r)})$ be the $p \times r$ matrix defined by the first columns of $\boldsymbol{B}^{(1)}, \ldots, \boldsymbol{B}^{(r)}$, and define $\{\boldsymbol{B}_j : k = 1, \ldots, q\}$ similarly. The model can then be expressed as

$$\boldsymbol{y}_i - \boldsymbol{\mu}_{\mathrm{x}_i} = \sum_{k=1}^{q} x_k \boldsymbol{B}_k \boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i.$$

Now suppose that $\boldsymbol{\gamma}_i$ is allowed to have a covariance matrix $\boldsymbol{\Phi}$ not necessarily equal to the identity. The above representation shows that the model given by $\{\boldsymbol{B}_1, \ldots, \boldsymbol{B}_k, \boldsymbol{\Phi}\}$ is equivalent to the one given by $\{\boldsymbol{B}_1\boldsymbol{\Phi}^{1/2}, \ldots, \boldsymbol{B}_k\boldsymbol{\Phi}^{1/2}, \boldsymbol{I}\}$, and so without loss of generality it can be assumed that $\boldsymbol{\Phi} = \boldsymbol{I}$, i.e. the random effects are independent with unit variance. In this case, note that $\mathrm{Var}\,[\boldsymbol{\gamma}_i] = \mathrm{Var}\,[\boldsymbol{H}\boldsymbol{\gamma}_i]$ where $\boldsymbol{H}$ is any $r \times r$ orthonormal matrix. This implies that the covariance function $\boldsymbol{\Sigma}_{\mathrm{x}}$ given by $\{\boldsymbol{B}_1, \ldots, \boldsymbol{B}_k, \boldsymbol{I}\}$ is equal to the one given by $\{\boldsymbol{B}_1\boldsymbol{H}, \ldots, \boldsymbol{B}_k\boldsymbol{H}, \boldsymbol{I}\}$ for any orthonormal $\boldsymbol{H}$, and so the parameters in the higher rank model are not completely identifiable. One possible identifiability constraint is to restrict $\boldsymbol{B}_1 = \{\boldsymbol{b}_1^{(1)}, \ldots, \boldsymbol{b}_1^{(r)}\}$, the matrix of first columns of $\boldsymbol{B}^{(1)}, \ldots, \boldsymbol{B}^{(r)}$, to have orthogonal columns.

Estimation: The random-effects representation for a rank-r covariance regression model is

$$\begin{aligned}\boldsymbol{y}_i &= \boldsymbol{\mu}_{\mathrm{x}_i} + \sum_{k=1}^{r} \gamma_{i,k} \times \boldsymbol{B}^{(k)}\boldsymbol{x}_i + \boldsymbol{\epsilon}_i \\ &= \boldsymbol{\mu}_{\mathrm{x}_i} + \tilde{\boldsymbol{B}}(\boldsymbol{\gamma}_i \otimes \boldsymbol{x}_i) + \boldsymbol{\epsilon}_i \ , \ \text{where } \tilde{\boldsymbol{B}} = (\boldsymbol{B}^{(1)}, \ldots, \boldsymbol{B}^{(r)}).\end{aligned}$$
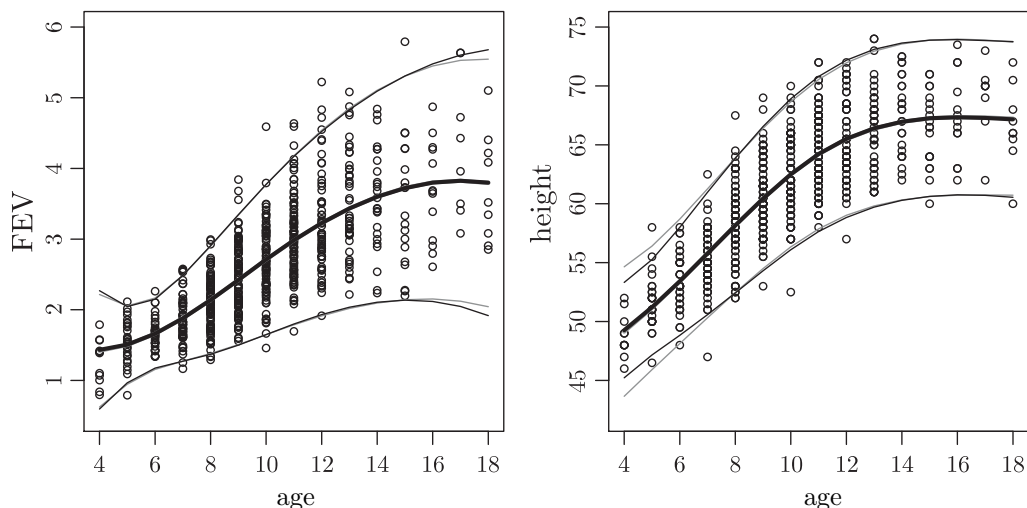
Figure 5. FEV and height data, as a function of age. The lines correspond to the mean functions plus and minus two standard deviations, as estimated by rank 1 and rank 2 covariance regression models, in gray and black, respectively.

Estimation for this model can proceed with a small modification of the Gibbs sampling algorithm given in Section 3, in which $\boldsymbol{B}^{(k)}$ and $\{\gamma_{i,k}, i = 1, \ldots, n\}$ are updated for each $k \in \{1, \ldots, r\}$ separately. An EM-algorithm is also available for estimation of this general rank model. The main modification to the algorithm presented in Section 3.1 is that the conditional distribution of each $\boldsymbol{\gamma}_i$ is multivariate normal, which leads to a more complex E-step in the procedure, while the M-step is equivalent to a multivariate least squares regression estimation, as before. We note that, in our experience, convergence of the EM-algorithm for ranks greater than 1 can be slow, due to the identifiability issue described earlier.

6. Lung Function and Height Data

To illustrate the use of the covariance regression model we analyze data on forced expiratory volume (FEV) in liters and height in inches of 654 Boston youths (Rosner (2000)). One feature of these data is the general increase in the variance of these variables with age, as shown in Figure 5. As the mean responses for these two variables are also increasing with age, one possible modeling strategy is to apply a variance stabilizing transformation to the data. In general, such transformations presume a particular mean-variance relationship, and choosing an appropriate transformation can be prone to much subjectivity. As an alternative, a covariance regression model allows heteroscedasticity to be modeled separately from the mean, and also allows for modeling on the original
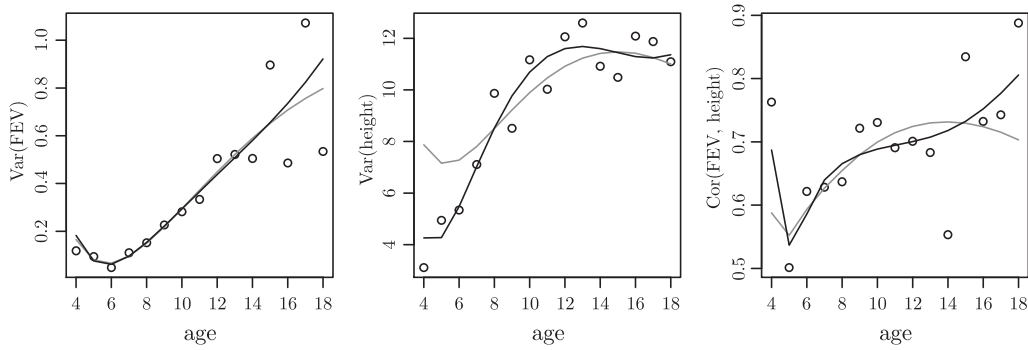
Figure 6. Sample variances and correlations as a function of age, along with rank 1 and 2 covariance regression fits in gray and black lines, respectively.

scale of the data.

6.1. Maximum likelihood estimation

Ages for the 654 subjects ranged from 3 to 19 years, although there were only two 3-year-olds and three 19-year-olds. We combine the data from children of ages 3 and 19 with those of the 4 and 18-year-olds, respectively, giving a sample size of at least 8 in each age category.

As seen in Figure 5, average FEV and height are somewhat nonlinear in age. We model the mean functions of FEV and height as cubic splines with knots at ages 4, 11 and 18, so that that $\mathrm{E}\left[\boldsymbol{y}_i|\mathrm{age}_i\right] = \boldsymbol{A}\boldsymbol{w}_i$, where $\boldsymbol{y}_i^T = (\mathrm{FEV}_i, \mathrm{height}_i)$ and $\boldsymbol{w}_i$ is a vector of length five determined by age_i and the spline basis. For the regressor in the variance function we use $\boldsymbol{x}_i = (1, \mathrm{age}_i^{1/2}, \mathrm{age}_i)^T$. Note that including $\mathrm{age}^{1/2}$ as a regressor results in linear age terms being in the model. We also fit both rank 1 and rank 2 models to these data, and compare their relative fits:

Rank 1 model: $\mathrm{Cov}\left[\boldsymbol{y}_i|\mathrm{age}_i\right] = \boldsymbol{\Psi} + \boldsymbol{B}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{B}^T$.

Rank 2 model: $\mathrm{Cov}\left[\boldsymbol{y}_i|\mathrm{age}_i\right] = \boldsymbol{\Psi} + \boldsymbol{B}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{B}^T + \boldsymbol{C}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{C}^T$.

Parameter estimates from these two models are incorporated into Figure 5. The MLEs of the mean functions for the rank 1 and 2 models, given by thick gray and black lines respectively, are indistinguishable. There are some visible differences in the estimated variance functions, represented in Figure 5 by curves at the mean $\pm$ 2 times the estimated standard deviation of FEV and height as a function of age. A more detailed comparison of the estimated variance functions for the two models is given in Figure 6. The estimated variance functions for FEV match the sample variance function very well for both models, although the

Table 3. Age-specific coverage rates for the 90% homoscedastic predictive ellipse and the 90% heteroscedastic (covariance regression) predictive ellipse.

							age group								
	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
sample size	11	28	37	54	85	94	81	90	57	43	25	19	13	8	9
homoscedastic	1	0.96	0.97	0.96	0.96	0.95	0.95	0.88	0.75	0.81	0.76	0.74	0.92	0.75	0.78
heteroscedastic	1	0.86	0.92	0.89	0.88	0.93	0.95	0.91	0.89	0.91	0.88	0.89	0.92	0.88	0.89

second plot in the figure indicates some lack of fit for the variance function for height by the rank 1 model at the younger ages.

Another means of evaluating this lack of fit is with a comparison of maximized log-likelihoods, which are -1927.809 and -1922.433 for the rank 1 and rank 2 models respectively. As discussed in Section 5 the first columns of $\boldsymbol{B}$ and $\boldsymbol{C}$ are not separately identifiable and may be transformed to be orthogonal without changing the model fit. As such, the difference in the number of parameters between the rank 1 and rank 2 models is four. A likelihood ratio test comparing the rank 1 and rank 2 models gives a p-value of 0.0295, based on a χ_4^2 null distribution, suggesting moderate evidence against the rank 1 model in favor of the rank 2 model.

6.2. Prediction regions

One potential application of the covariance regression model is to make prediction regions for multivariate observations. Erroneously assuming a covariance matrix to be constant in $\boldsymbol{x}$ could give a prediction region with correct coverage rates for an entire population, but incorrect rates for specific values of $\boldsymbol{x}$, and incorrect rates for populations having distributions of $\boldsymbol{x}$-values that are different from that of the data. For the FEV data, an approximate 90% prediction ellipse for $\boldsymbol{y}$ for each age can be obtained from the set

$$\{\boldsymbol{y} : (\boldsymbol{y} - \hat{\boldsymbol{\mu}}_{\text{age}})^T \hat{\boldsymbol{\Sigma}}_{\text{age}}^{-1}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_{\text{age}}) < \chi_{.9,2}^2\},$$

where $\hat{\boldsymbol{\mu}}_{\text{age}} = \hat{\boldsymbol{A}}\boldsymbol{w}$, $\hat{\boldsymbol{\Sigma}}_{\text{age}} = \hat{\boldsymbol{\Psi}} + \hat{\boldsymbol{B}}\boldsymbol{x}\boldsymbol{x}^T\hat{\boldsymbol{B}}^T$ and $\boldsymbol{w}$ and $\boldsymbol{x}$ are vector-valued functions of age as described above.

Ellipses corresponding to the fit from the rank 2 model are displayed graphically in Figure 7, along with the data and an analogous predictive ellipse obtained from the homoscedastic model. Averaged across observations from all age groups, the homo- and heteroscedastic ellipses contain 90.1% and 90.8% of the observed data respectively, both percentages being very close to the nominal coverage rate of 90%. However, as can be seen from Table 3, the homoscedastic ellipse generally overcovers the observed data for the younger age groups, and undercovers for the older groups. In contrast, the flexibility of the covariance regression model
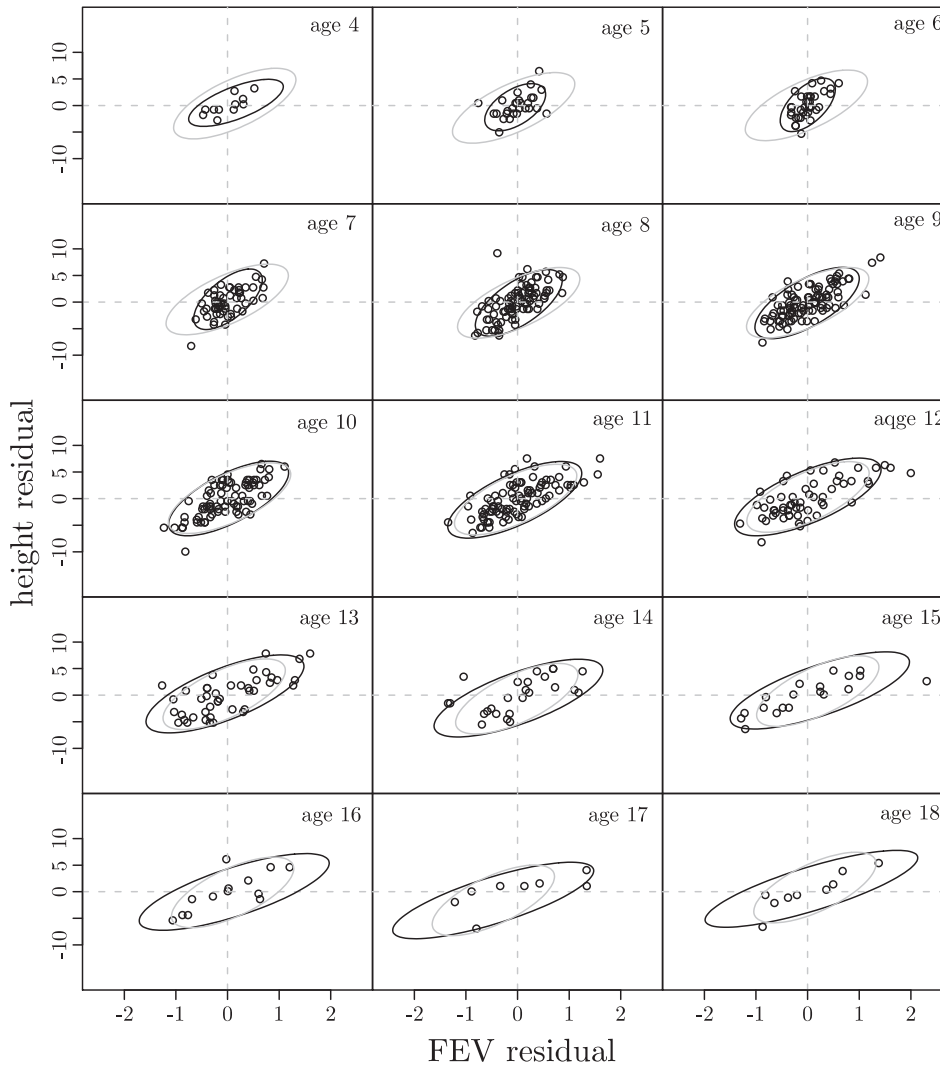
Figure 7. Observed data and approximate 90% predictive ellipsoids for each age. The black ellipsoids correspond to the covariance regression model, and the gray to the homoscedastic multivariate normal model.

allows the confidence ellipsoids to change size and shape as a function of age, and thus match the nominal coverage rate fairly closely across the different ages.

7. Discussion

This article has presented a model for a covariance matrix $\text{Cov}[\boldsymbol{y}|\boldsymbol{x}] = \boldsymbol{\Sigma}_{\text{x}}$ as a function of an explanatory variable $\boldsymbol{x}$. We have presented a geometric interpretation in terms of curves along the boundary of a translated positive

definite cone, and have provided a random-effects representation that facilitates parameter estimation. This covariance regression model goes beyond what can be provided by variance stabilizing transformations, which serve to reduce the relationship between the mean and the variance. Unlike models or methods which accommodate heteroscedasticity in the form of a mean-variance relationship, the covariance regression model allows for the mean function $\boldsymbol{\mu}_x$ to be separately parameterized from the variance function $\boldsymbol{\Sigma}_x$.

The covariance regression model accommodates explanatory variables of all types, including categorical variables. This could be useful in the analysis of multivariate data sampled from a large number of groups, such as groups defined by the cross-classification of several categorical variables. For example, it may be desirable to estimate a separate covariance matrix for each combination of age group, education level, race, and religion in a given population. The number of observations for each combination of explanatory variables may be quite small, making it impractical to estimate a separate covariance matrix for each group. One strategy, taken by Flury (1984) and Pourahmadi, Daniels, and Park (2007), is to assume that a particular feature of the covariance matrices (principal components, correlation matrix, Cholesky decomposition) is common across groups. A simple alternative to assuming that certain features are exactly preserved across groups would be a covariance regression model, allowing a parsimonious but flexible representation of the heteroscedasticity across the groups.

While neither the covariance regression model nor its random effects representation in Section 2 assume normally distributed errors, normality was assumed for parameter estimation in Section 3. Accommodating other types of error distributions is feasible and straightforward to implement in some cases. For example, heavy-tailed error distributions can be accommodated with a multivariate t model, in which the error term can be written as a multivariate normal random variable multiplied by a χ^2 random variable. Estimates based upon this data-augmented representation can then be made using the EM algorithm or the Gibbs sampler (see, for example, Gelman et al. (2004, Chap. 17)).

Like mean regression, a challenge for covariance regression modeling is variable selection, i.e. the choice of an appropriate set of explanatory variables. One possibility is to use selection criteria such as AIC or BIC, although non-identifiability of some parameters in the higher-rank models requires a careful accounting of the dimension of the model. Another possibility may be to use Bayesian procedures, either by MCMC approximations to Bayes factors, or by explicitly formulating a prior distribution to allow some coefficients to be zero with non-zero probability.

Replication code and data for the analyses in this article are available at the first author's website: `http://www.stat.washington.edu/~hoff`.

Acknowledgement

References

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. (With discussion). *J. Roy. Statist. Soc. Ser. B* **26**, 211-252.

Carroll, R. J. (1982). Adapting for heteroscedasticity in linear models. *Ann. Statist.* **10**, 1224-1233.

Carroll, R. J., Ruppert, D. and Holt, Jr. R. N. (1982). Some aspects of estimation in heteroscedastic linear models. In *Statistical decision theory and related topics*, III, **1** (West Lafayette, Ind., 1981), 231-241. New York: Academic Press.

Chiu, T. Y. M., Leonard, T. and Tsui, K.-W. (1996). The matrix-logarithmic covariance model. *J. Amer. Statist. Assoc.* **91**, 198-210.

Engle, R. F. and Kroner, K. F. (1995). Multivariate simultaneous generalized arch. *Econometric Theory* **11**, 122-150.

Flury, B. N. (1984). Common principal components in k groups. *J. Amer. Statist. Assoc.* **79**, 892-898.

Fong, P. W., Li, W. K. and An, H.-Z. (2006). A simple multivariate ARCH model specified by random coefficients. *Comput. Statist. Data Anal.* **51**, 1779-1802.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004). *Bayesian Data Analysis*. 2nd edition. Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL.

Henderson, H. V. and Searle, S. R. (1979). Vec and vech operators for matrices, with some uses in Jacobians and multivariate statistics. *Canad. J. Statist.* **7**, 65-81.

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Amer. Statist. Assoc.* **90**, 928-934.

Magnus, J. R. and Neudecker, H. (1979). The commutation matrix: some properties and applications. *Ann. Statist.* **7**, 381-394.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis*. London: Academic Press [Harcourt Brace Jovanovich Publishers]. Probability and Mathematical Statistics: A Series of Monographs and Textbooks.

McCulloch, C. E. (1982). Symmetric matrix derivatives with applications. *J. Amer. Statist. Assoc.* **77**, 679-682.

Müller, H.-G. and Stadtmüller, U. (1987). Estimation of heteroscedasticity in regression analysis. *Ann. Statist.* **15**, 610-625.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika* **86**, 677-690.

Pourahmadi, M., Daniels, M. J. and Park, T. (2007). Simultaneous modelling of the Cholesky decomposition of several covariance matrices. *J. Multivariate Anal.* **98**, 568-587.

Rosner, B. (2000). *Fundamentals of Biostatistics*. Duxbury Press.

Rutemiller, H. C. and Bowers, D. A. (1968). Estimation in a heteroscedastic regression model. *J. Amer. Statist. Assoc.* **63**, 552-557.

Scott, M. and Handcock, M. (2001). Covariance models for latent structure in longitudinal data. *Sociological Methodology* **31**, 265-303.

Smyth, G. K. (1989). Generalized linear models with varying dispersion. *J. Roy. Statist. Soc. Ser. B* **51**, 47-60.

Yin, J., Geng, Z., Li, R. and Wang, H. (2010). Nonparametric covariance model. *Statist. Sinica* **20**, 469-479.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with *g*-prior distributions. In *Bayesian Inference and Decision Techniques*, Volume 6 of *Stud. Bayesian Econometrics Statist.*, 233-243. Amsterdam: North-Holland.

Departments of Statistics and Biostatistics, University of Washington, Seattle, WA 98195-4322, U.S.A.

E-mail: pdhoff@uw.edu

Department of Statistics, Penn State University, University Park, PA, 16802, U.S.A.

E-mail: xiaoyue@psu.edu