Cholesky Residuals for Assessing Normal Errors in a Linear Model with Correlated Outcomes: Technical Report

E. Andres Houseman∗ Louise Ryan†
Brent Coull‡

∗Harvard School of Public Health, ahousema@hsph.harvard.edu
†Harvard School of Public Health and Dana-Farber Cancer Institute, lryan@hsph.harvard.edu
‡Harvard School of Public Health, bcoull@hsph.harvard.edu

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

http://biostats.bepress.com/harvardbiostat/paper19

Copyright ©2004 by the authors.
Cholesky Residuals for Assessing Normal Errors in a Linear Model with Correlated Outcomes: Technical Report

E. Andres Houseman, Louise Ryan, and Brent Coull

Abstract

Despite the widespread popularity of linear models for correlated outcomes (e.g. linear mixed models and time series models), distribution diagnostic methodology remains relatively underdeveloped in this context. In this paper we present an easy-to-implement approach that lends itself to graphical displays of model fit. Our approach involves multiplying the estimated marginal residual vector by the Cholesky decomposition of the inverse of the estimated marginal variance matrix. The resulting “rotated” residuals are used to construct an empirical cumulative distribution function and pointwise standard errors. The theoretical framework, including conditions and asymptotic properties, involves technical details that are motivated by Lange and Ryan (1989), Pierce (1982), and Randles (1982). Our method appears to work well in a variety of circumstances, including models having independent units of sampling (clustered data) and models for which all observations are correlated (e.g., a single time series). Our methods can produce satisfactory results even for models that do not satisfy all of the technical conditions stated in our theory.
Cholesky Residuals for Assessing Normal Errors in a Linear Model with Correlated Outcomes: Technical Report

E. Andrés Houseman
Department of Biostatistics
Harvard School of Public Health

Louise M. Ryan
Department of Biostatistics
Harvard School of Public Health

Brent A. Coull
Department of Biostatistics
Harvard School of Public Health

July 18, 2003

Abstract

Despite the widespread popularity of linear models for correlated outcomes (e.g., linear mixed models and time series models), distribution diagnostic methodology remains relatively underdeveloped in this context. In this paper we present an easy-to-implement approach that lends itself to graphical displays of model fit. Our approach involves multiplying the estimated marginal residual vector by the Cholesky decomposition of the inverse of the estimated marginal variance matrix. The resulting “rotated” residuals are used to construct an empirical cumulative distribution function and pointwise standard errors. The theoretical framework, including conditions and asymptotic properties, involves technical details that are motivated by Lange and Ryan (1989), Pierce (1982), and Randles (1982). Our method appears to work well in a variety of circumstances, including models having independent units of sampling (clustered data) and models for which all observations are correlated (e.g., a single time series). Our methods can produce satisfactory results even for models that do not satisfy all of the technical conditions stated in our theory.

Keywords

cumulative distribution function; goodness-of-fit; linear mixed model; random effects; residual diagnostics

1 Introduction

Correlated data typically arise in public health, biomedical and environmental applications, requiring techniques such as linear mixed models and time series regression. Despite the widespread popularity of these methods, diagnostic methodology for addressing goodness-of-fit in this context remains relatively underdeveloped. Verbeke and
Molenberghs (2000) noted that the choice of diagnostic method for linear mixed models is not obvious (pages 151–152), and Agresti (2002) noted that research is needed on model checking and diagnostics for mixed models in general. Recently, Lin et al. (2002) have proposed graphical techniques for assessing the adequacy of the deterministic portion of a generalized linear mixed model, but their methods do not address the random component of model fit.

Limited methodology is available for assessing the distribution of the error term. Existing work includes Lange and Ryan (1989), who developed methods using estimated random effects from growth curve models, Fraccaro et al. (2000), who discussed residual diagnostic plots in time series regression, and Louis (1988), who discussed an approach similar to the one we describe. The latter two papers present no asymptotic theory. Several authors have recently proposed goodness-of-fit tests that apply in the mixed model setting (Hodges, 1998; Jiang, 2001), but the approaches are complex and do not lend themselves well to graphical displays.

For models with subject-specific random effects, Pinheiro and Bates (2000) advocated the use of the standardized residual formed using predictions of subject-specific means and an estimate of residual error (page 239). However, comparing fitted and observed values can be misleading, as these comparisons reflect intended shrinkage of estimates towards overall means. Coull et al. (2001) and Longford (2001) addressed model fit by simulating data sets from the maximum likelihood fit of the model and checking to see whether the observed data set was extreme relative to the reference set. This idea has also been used in spatial statistics (Ripley, 2001) and is similar to the posterior predictive checking strategy often used in Bayesian analyses (Gelman et al., 1995).
In this paper we present an approach that, like that of Pinheiro and Bates (2000), involves standardized residuals. However, our approach rotates the estimated marginal residual vector by the Cholesky decomposition of the inverse of the estimated marginal variance matrix. Asymptotic properties are established using the methods of Pierce (1982) and Randles (1982). Our approach has the advantage of enabling graphical depictions of goodness-of-fit along the lines of a quantile-quantile (Q-Q) plot. Additionally, our methods apply to a larger class of models than those studied by Lange and Ryan (1989); they apply not only to linear mixed models, but also linear models for which there are no independent units of sampling (for example, time series models).

We have focused on distribution diagnostics, but remark that the cumulative residual approach of Lin et al. (2002) can be used to check the mean structure, and graphical displays such as a draftsman’s display (Dawson et al., 1997) and PRISM (Zimmerman, 2000) are useful for diagnosing the form of the covariance.

Our paper is organized as follows. Section 2 describes the background of our proposed methodology, introducing the problem and the heuristics of our solution. Section 3 presents the technical justification for our proposed methods. Section 4 reports simulation results, Section 5 demonstrates the utility of our approach in two correlated data settings, and Section 6 states our conclusions.

2 Rotated Residuals of a Linear Model

In this section we formally describe the context of our proposed methodology and define the concept of a rotated residual. Throughout our exposition, we assume that a linear
model has been specified as follows:

\[ y = X\beta_0 + \varepsilon, \tag{1} \]

where \( \varepsilon \sim N_n(0, V(\gamma_0)) \). Here \( y \in \mathbb{R}^n \) consists of \( n \) outcomes, possibly correlated. The marginal mean of \( y \) is determined by an \( n \times p \) design matrix of covariates \( X \) and a \( p \)-dimensional parameter vector \( \beta_0 \in \mathbb{R}^p \). The marginal covariance of \( y \) is determined by the \( n \times n \) matrix \( V(\gamma_0) \) having known form depending on a \( q \)-dimensional unknown parameter \( \gamma_0 \in \mathbb{R}^q \). For notational brevity, we sometimes denote \( \theta = (\beta, \gamma) \) as an arbitrary point in the parameter space and \( \theta_0 = (\beta_0, \gamma_0) \) as the true parameter value.

For example, a random effects model with a single random intercept is equivalent to (1), where the marginal variance \( V(\gamma_0) \) is block-diagonal, each block having a compound symmetry structure. For such a model, \( \gamma_0 \in \mathbb{R}^2 \), which we might explicitly write as \( \gamma_0^T = (\sigma_0^2, \rho_0) \). The elements of the diagonal of \( V(\gamma_0) \) are \( \sigma_0^2 \), while the off-diagonal elements of the nonzero blocks are \( \sigma_0^2 \rho_0 \). A more general mixed effects model has a block-diagonal matrix with blocks \( \sigma^2(\gamma_0)I_h + Z_h\Delta(\gamma_0)Z_h^T \), where \( Z_h \) is a matrix of covariates for cluster \( h \), \( I_h \) is the identity matrix of corresponding dimension, \( \sigma^2(\gamma_0) \) is the error variance, and \( \Delta(\gamma_0) \) is the covariance matrix of the random effects. In the more general random effects model, \( \Delta(\gamma_0) \) is a matrix having known structure depending upon the parameter vector \( \gamma_0 \). Another example is time series regression, in which the residual correlation of outcomes observed at times \( s \) and \( t \) depends upon a function of \( s \) and \( t \), e.g., \( \rho^{|s-t|} \).

Model fitting for (1) is easily achieved using least squares, maximum likelihood (ML), or restricted maximum likelihood (REML). Therefore, we focus on the problem of goodness-of-fit and detection of outliers. Specifically, we are interested in whether the empirical
CDF of our proposed rotated residuals matches the CDF of a normal distribution. To this end, we define the vector of residuals $z(\beta, \gamma)$ as a function of possible parameter values $\beta$ and $\gamma$. Although the residuals $y - X\beta$ can be standardized using the fitted diagonal of $V(\gamma)$, the resulting residuals are still potentially highly correlated, even in large samples. Therefore, we propose an $n$-dimensional rotation of these residuals to yield new ones that are asymptotically uncorrelated. Let $L(\gamma)$ be the Cholesky decomposition of $V(\gamma)^{-1}$, so that $V(\gamma)^{-1} = L(\gamma)L(\gamma)^T$, and define

$$z(\beta, \gamma) = L(\gamma)^T(y - X\beta).$$

Then $z(\beta_0, \gamma_0) \sim N_n(0, I_n)$, where $I_n$ is the $n \times n$ identity matrix. If $\beta_0$ and $\gamma_0$ are known, a goodness-of-fit procedure for the appropriateness of model (1) could be established by comparing $\Phi(x)$, the standard normal CDF evaluated at $x$, to the empirical CDF evaluated at $x$, for various values of $x$. The CDF can be expressed as $F_n(\beta_0, \gamma_0; x)$, where

$$F_n(\beta, \gamma; x) = n^{-1} \sum_{i=1}^n I \left( x - P_i z(\beta, \gamma) \right),$$

$I(u) = 1$ if $u \leq 0$, $I(u) = 0$ if $u > 0$, and $P_i$ is the projection onto the $i$th component. Under the null hypothesis that $z_i(\beta_0, \gamma_0) = P_i z(\beta_0, \gamma_0) \sim N(0, 1)$ [equivalent to (1)], a pointwise confidence interval for $F_n(\beta_0, \gamma_0; x)$ is easily constructed by noting that the variance of (3), evaluated at $\theta_0$, is simply $n^{-1}\Phi(x)(1 - \Phi(x))$.

In any practical situation, $\beta_0$ and $\gamma_0$ must be estimated from the data. This complicates the construction of confidence intervals in that the variance of (3) tends to shrink when the parameters are replaced by estimates based on the data. Lange and Ryan (1989) described this phenomenon extensively in the context of constructing a Q-Q plot of random effects predicted in growth curve models. To address this problem, Lange and Ryan (1989) used the results of Randles (1982) and Pierce (1982) to prove asymptotic
normality of a similar estimator, but the results apply only when there are independent units of sampling. Our proposed method applies even when all observations are correlated. We remark that, although Randles (1982) and Pierce (1982) provide the inspiration for our proofs, the results described in these papers cannot be used here directly because they assume independent residuals.

In this section we describe the heuristics of our method, deferring technical details until Section 3. Let \( \hat{\beta} \) and \( \hat{\gamma} \) be consistent estimators for \( \beta_0 \) and \( \gamma_0 \) respectively, and denote the joint estimator as \( \hat{\theta} = (\hat{\beta}, \hat{\gamma}) \), consistent for \( \theta_0 \). In many situations, \( F_n(\hat{\beta}, \hat{\gamma}; x) \) is still asymptotically normal when model (1) is correct. However, its asymptotic variance is not \( \Phi(x)(1 - \Phi(x)) \). As described in Lange and Ryan (1989), a variance correction factor must be applied. In Section 3, we show that

\[
n^{\frac{1}{2}} \left( F_n(\hat{\beta}, \hat{\gamma}; x) - \Phi(x) \right) = Z + O_p(1),
\]

where \( Z \sim N(0, \tau^2) \),

\[
\tau^2 = \Phi(x)(1 - \Phi(x)) - \delta_0^T W \delta_0,
\]

\( W = Var(\hat{\theta}) \), and \( \delta_0 \) is a vector that can be consistently estimated from the data. We defer the precise definition of \( \delta_0 \) until Section 3, but remark that it can be estimated by \( \delta_n(\hat{\theta}, \hat{\beta}) \), where

\[
\delta_n(\theta, \theta_0) = \frac{\partial \mu_n}{\partial \theta}(\theta, \theta_0; x)
\]

and

\[
\mu_n(\theta, \theta_0; x) = \frac{1}{n} \sum_{i=1}^{n} \Phi \left( \frac{x - P_i L(\gamma)^T X (\beta_0 - \beta)}{\sqrt{P_i L(\gamma)^T V(\gamma_0) L(\gamma) P_i^T}} \right).
\]  \hspace{1cm} (4)

At first glance, the notation may appear cumbersome, but is necessary to elucidate the theory. The function \( \mu_n \) has two arguments, the first of which describes the local
behavior of $F_n(\beta, \gamma)$ near $\theta_0$ and the second of which describes the dependence on the true parameter value. The unknown vector $\delta_0$ depends on the local behavior near $\theta_0$, and is defined as the derivative (with respect to the first argument), evaluated at $\theta_0$, of the limit of $\mu_n$. Therefore, its estimate $\delta_n$ has two arguments that should be evaluated as near to $\theta_0$ as possible.

Applications are straightforward in practice. Standard applications of ML and REML include consistent estimators for $W$. Consequently, $\tau^2$ can be estimated consistently by replacing $W$ and $\delta_0$ by consistent estimates. The derivative of (4) generally has a closed-form expression, but is often complicated when $V(\gamma)$ has a complicated structure; even in such cases satisfactory results can still be obtained by numerical differentiation. For example, finite difference approximations (Dennis and Schnabel, 1996) were used for all of the simulations presented in Section 4.

3 Technical Results

In this section we present the major technical results needed to justify the proposed methods, relegating proofs to the appendices. Let $\theta_0 = (\beta_0^T, \gamma_0^T)^T \in \Theta \subset \mathbb{R}^p \times \mathbb{R}^q$, where $\Theta$ is an open subset. Assume $V_n(\gamma)$ is a positive definite symmetric matrix of order $n$, continuously differentiable in a neighborhood of $\gamma_0$, and that $L_n(\gamma)L_n(\gamma)^T = V_n(\gamma)^{-1}$. Note that $L_n(\gamma)$, a Cholesky decomposition, is also positive definite and continuously differentiable in $\gamma$. Let $y_n \sim N_n(X_n\beta_0, V_n(\gamma_0))$, where $X_n$ is bounded. When the context is clear, the subscript $n$ will be omitted from $y_n$, $X_n$, and $V_n$.

Let $z(\beta, \gamma)$ be defined as in (2) and let $z_i(\beta, \gamma) = P_i z(\beta, \gamma)$. Fixing $x \in \mathbb{R}$, define
\(F_n(\beta, \gamma; x)\) as in (3) and define

\[
\mu_n(\beta, \gamma; x) = n^{-1} \sum_{i=1}^{n} \Phi(x_i(\beta, \gamma; x)),
\]

where

\[
x_i(\beta, \gamma; x) = \frac{x - m_i(\beta, \gamma)}{s_i(\gamma)},
\]

\[m_i(\beta, \gamma) = P_i L(\gamma)^T X(\beta_0 - \beta),\]

and

\[s_i(\gamma)^2 = P_i L(\gamma)^T V(\gamma_0) L(\gamma) P_i^T.\]

Here, \(m_i\) and \(s_i^2\) are the mean and variance of \(z_i\). Note that \(z_i(\beta, \gamma) \sim N(m_i(\beta, \gamma), s_i(\gamma)^2)\), so

\[z_i^*(\beta, \gamma) = \frac{z_i(\beta, \gamma) - m_i(\beta, \gamma)}{s_i(\gamma)} \sim N(0, 1)\]

and

\[E [I (x - z_i(\beta, \gamma))] = E [I (x_i(\beta, \gamma; x) - z_i^*(\beta, \gamma))] = \Phi(x_i(\beta, \gamma; x)).\]

Finally, we write the column-vector gradient of a vector valued function \(g(\zeta)\) as \(\frac{\partial g}{\partial \zeta}\) and its transpose as \(\frac{\partial^T g}{\partial \zeta^T}\).

Theorem 1 establishes our main result, namely that the asymptotic distribution (pointwise in \(x\)) of the empirical CDF (3) is normal.

**Theorem 1 (Asymptotic Normality).** Suppose \(\hat{\theta}_n\) is a consistent and asymptotically efficient estimator of \(\theta_0\), that

\[
n^{- \frac{1}{2}} \begin{bmatrix} \hat{\theta}_n - \theta_0 \\ F_n(\beta_0, \gamma_0; x) - \Phi(x) \end{bmatrix} \quad \overset{\zeta}{\sim} \quad \begin{bmatrix} 0 \\ W_{11} & W_{21} \\ 0 & W_{12} \Phi(x)(1 - \Phi(x)) \end{bmatrix}
\]

and that

\[
R_n \equiv n^{- \frac{1}{2}} \left| F_n(\hat{\beta}_n, \hat{\gamma}_n; x) - \mu_n(\hat{\beta}_n, \hat{\gamma}_n; x) - F_n(\beta_0, \gamma_0; x) + \Phi(x) \right| \overset{P}{\to} 0.
\]
In addition, suppose there exists a continuously twice-differentiable function \( \mu^*(\beta, \gamma; x) \) such that for all \( n \),

\[
\mu_n(\beta, \gamma) = \mu^*(\beta, \gamma) + \varepsilon_n^T(\theta - \theta_0) + O(|\theta - \theta_0|^2),
\]

where \( \varepsilon_n \) is constant with respect to \( \theta \), \( \varepsilon_n \rightarrow 0 \) as \( n \rightarrow \infty \), and the boundedness of \( O(|\theta - \theta_0|^2) \) is uniform in \( n \). Then

\[
n^\frac{1}{2}[F_n(\tilde{\beta}_n, \tilde{\gamma}_n; x) - \Phi(x)] \xrightarrow{L} N(0, \tau^2),
\]

where

\[
\tau^2 = \Phi(x) - \Phi(x) - \delta(\theta_0)^T W_{11}\delta(\theta_0)
\]

and

\[
\delta(\theta) = \frac{\partial^T \mu^*}{\partial \theta}.
\]

Condition (6) is an asymptotic differentiability property that permits the replacement of a non-differentiable estimator by a probability limit that is differentiable. Condition (7) refers to a theoretical function \( \mu^* \) that is never observable in practice; \( \mu^* \) can be interpreted as a limit of \( \mu_n \), which is identically equal to \( \Phi(x) \) when evaluated at \( \theta_0 \). However, the local behavior of \( \mu_n \) affects the asymptotic variance of (3), hence the presence of \( \delta \) in (9).

Theorem 1 establishes the theoretical variance \( \tau^2 \) of the estimator \( F_n \). However, \( \tau^2 \) must be estimated if the theorem is to be of practical use. If consistent estimators \( \hat{W}_{11} \) and \( \hat{\delta} \) of \( W_{11} \) and \( \delta(\theta_0) \) are available, then \( \tau^2 \) is consistently estimated by substituting \( \hat{W}_{11} \) for \( W_{11} \) and \( \hat{\delta} \) for \( \delta(\theta_0) \) in (9). Standard methods for obtaining \( \hat{W}_{11} \) are available using ML and REML [see, for example, McCulloch and Searle (2001) or Diggle et al. (2002)]. To justify a consistent estimator for \( \delta(\theta_0) \), we present Theorem 2, which uses

9
\( \mu_n(\beta, \gamma) \) as an approximation to \( \mu^*(\beta, \gamma) \). Since \( \mu_n(\beta, \gamma) \) implicitly depends on \( \theta_0 \), we must justify the substitution of \( \tilde{\theta}_n \) for \( \theta_0 \), as well as the differentiation of \( \mu_n(\beta, \gamma) \).

**Theorem 2 (Estimation of \( \tau^2 \)).** Let \( \delta(\theta) \) be defined as in (10) and suppose condition (7) in Theorem 1 holds. In addition, let \( \hat{\mu}_n(\theta, \theta_0) \) denote \( \mu_n(\beta, \gamma) \) considered as a function of both \( \theta = (\beta, \gamma)^T \) and the true parameter \( \theta_0 = (\beta_0, \gamma_0)^T \). Assume \( \hat{\mu}_n(\theta, \theta_0) \) is continuously twice-differentiable in the first parameter. Define

\[
\delta_n(\theta, \theta_0) = \frac{\partial^T \hat{\mu}_n}{\partial \theta^T},
\]

where the derivative is taken with respect to the first argument. Finally, suppose that for all \( n \) and for all \( \theta \) sufficiently close to \( \theta_0 \), there is an \( M > 0 \) such that

\[
|\delta_n(\theta, \theta') - \delta_n(\theta, \theta_0)| \leq M|\theta' - \theta_0|.
\]

If \( \tilde{\theta}_n \xrightarrow{p} \theta_0 \), then \( \delta_n(\tilde{\theta}_n, \tilde{\theta}_n) \xrightarrow{p} \delta(\theta_0) \).

The asymptotic normality condition (5), along with the corresponding efficiency requirement, can be established for many types of models. For clustered data, standard methods [e.g. van der Vaart (1998), Chapter 5] may be used to establish the result. In this context, Searle (1970) describes an explicit form for \( W_{11} \). For ARMA time series models, Durbin (1960) is a useful reference. Joint asymptotic normality of \( \hat{\theta}_n \) with \( F_n(\beta_0, \gamma_0; x) \) is reasonable in many applications, and can be derived using the Cramér-Wold device when \( \hat{\theta}_n \) is asymptotically linear in the sense described by van der Vaart (1998).

Conditions (6), (7), and (12) are difficult to verify directly, but amount to regularity conditions on the covariate matrix \( X \) and the variance \( V \). In essence, condition (6) requires that the covariance of \( z_i \) and \( z_j \) be negligible when observations \( i \) and \( j \) are
widely separated. This clearly occurs for clustered data models, such as longitudinal mixed models, and can be shown to hold for ARMA models. In the appendix, we state and prove theorems that verify the conditions for both of these cases. Conditions (7) and (12) require a degree of uniformity in the sequence of data collected, as Theorem 7 in the appendix suggests, and reduce to regularity conditions on $X$ and $V$.

In the next section, we report the results of several simulation studies that show that our proposed methods provide satisfactory results for many different types of models in finite sample settings.

4 Simulations

In the appendix we show that for many practical linear model designs, the empirical CDF obtained from the rotated residuals leads (pointwise) to asymptotically normal estimators having variance that is easily estimated. In the present section we demonstrate that these good statistical properties can be obtained at moderate sample sizes for various designs, even those that do not technically fit within the framework described in the appendix. Additionally, we demonstrate the sensitivity of our proposed methods when the error vector is not normal, and when the structure for the “within-cluster” correlation is misspecified.

We simulated several different cases. For each case, 1000 simulated data sets were used to summarize the behavior of our proposed method. In each case, the sample size was either $n = 50$ or $n = 200$, $X$ was an $n \times 2$ matrix with an intercept column and a column of uniform $(0, 1)$ elements, and $\beta_0^T = (10, 0.5)$. We estimated the empirical CDF at the values $x = 0$ and $x = 1$. We investigated several different correlation
structures, including clustered compound-symmetry and AR-1 models and several types of non-clustered designs described in the following subsections. We also investigated the behavior of our method when the true error vector contained heavy-tailed elements and skewed elements, and when the internal correlation structure of the clusters was misspecified.

All simulations were executed within Splus. For each simulated data set, we used maximum likelihood (through the Splus function \textit{nlminb}) to obtain \( \hat{\beta} \) and \( \hat{\gamma} \), then computed \( F_n(\hat{\beta}, \hat{\gamma}) \) and the standard error as implied by (9), using finite difference approximations (Dennis and Schnabel, 1996) to compute the derivative. For each simulation, we calculated the simulation standard deviation, the simulation confidence interval for \( E[F_n(\hat{\beta}, \hat{\gamma})] \), and coverage probabilities for a nominal 95\% confidence interval. For comparison purposes we computed a “naive” standard error estimate \( (\Phi(x)(1 - \Phi(x)))^{1/2} \), unadjusted for estimation of \( \beta \) and \( \gamma \). We also computed the t-statistic for the average of the realized \( F_n(\hat{\beta}, \hat{\gamma}) \) estimates with respect to the simulation standard error; although we do not interpret this statistic formally, we present it to compare the bias in \( F_n(\hat{\beta}, \hat{\gamma}) \) between cases for which the model was correctly specified and cases for which the model was misspecified.

### 4.1 Clustered Data

To examine the behavior of clustered designs, we considered two types of clustered designs, compound symmetry and AR-1. In each case, we used a uniform cluster size of 5, either with 10 clusters \((n = 50)\) or with 40 clusters \((n = 200)\). We note that for both types of clustered designs, \( \text{Cov}[y_{hi}, y_{h'i'}] = 0 \) when \( h \neq h' \).

For the compound symmetry case, each distinct observation within a cluster is cor-
related by the same correlation coefficient $\gamma_2$, so that the overall correlation structure is characterized by $\text{Cov}[y_{hi}, y_{hi'}] = \gamma_1 \gamma_2$ when $i \neq i'$ and $\text{Cov}[y_{hi}, y_{hi}] = \gamma_1$. For the (homogeneous) AR-1 case, the overall correlation structure is characterized by $\text{Cov}[y_{hi}, y_{hi'}] = \gamma_1 \gamma_2^{|i-i'|}$. For each of our simulations we used $\gamma_1 = 1$ and $\gamma_2$ equal to 0.1, 0.5, or 0.9.

The results are summarized in Table 1. Note that the estimator $F_n$ seems unbiased for $x = 0$ and may be just slightly biased for $x = 1$. The standard error computed by (9) matches the simulation standard deviation, whereas the naive standard error does not. Additionally, the coverage probabilities for a nominal 95% confidence interval are adequate when our standard error is used, but is generally too large when the naive standard error is used. Note also that the asymptotic behavior with respect to $n$ appears to match the theory, as the simulation standard deviations for $n = 10 \times 5$ are about twice those for $n = 40 \times 5$.

### 4.2 Non-clustered Data

To examine the behavior of non-clustered designs, we considered four types of designs for which no natural clusters exist. We considered an AR-1 time series model (as described in Theorem 5), a spatial model, and a crossed effects model we refer to as $CS \times CS$. We also considered a model having a complicated correlation structure; denoted $AR-1 \times CS$, it describes clusters of data whose random intercepts have an autoregressive structure.

The covariance structure of the AR-1 time series is described above in Theorem 5. For each simulated case we held the scale parameter $\gamma_1$ equal to 1, but allowed the temporal autocorrelation parameter $\gamma_2$ to vary among 0.1, 0.5, and 0.9.
Table 1: Simulations for Correctly Specified Clustered Data

\( \phi(x) = 0.500 \)

<table>
<thead>
<tr>
<th>Cluster Struct</th>
<th>( \gamma_2 )</th>
<th>( n )</th>
<th>( F_n )</th>
<th>SE</th>
<th>Sim</th>
<th>SE</th>
<th>Ours</th>
<th>Naive</th>
<th>( t_{sim} )</th>
<th>95% Sim CI Low</th>
<th>95% Sim CI Hi</th>
<th>95% Coverage Ours</th>
<th>95% Coverage Naive</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR-1</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.499</td>
<td>0.042</td>
<td>0.043</td>
<td>0.071</td>
<td>-0.73</td>
<td>0.498</td>
<td>0.502</td>
<td>0.972</td>
<td>1.000</td>
<td>0.964</td>
<td>0.999</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.1</td>
<td>40 x 5</td>
<td>0.500</td>
<td>0.021</td>
<td>0.021</td>
<td>0.035</td>
<td>-0.55</td>
<td>0.498</td>
<td>0.501</td>
<td>0.964</td>
<td>0.999</td>
<td>0.966</td>
<td>0.999</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.501</td>
<td>0.042</td>
<td>0.045</td>
<td>0.071</td>
<td>1.07</td>
<td>0.499</td>
<td>0.504</td>
<td>0.966</td>
<td>0.999</td>
<td>0.955</td>
<td>0.981</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.5</td>
<td>40 x 5</td>
<td>0.498</td>
<td>0.023</td>
<td>0.022</td>
<td>0.035</td>
<td>-2.09</td>
<td>0.497</td>
<td>0.500</td>
<td>0.925</td>
<td>0.994</td>
<td>0.955</td>
<td>0.981</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.9</td>
<td>10 x 5</td>
<td>0.501</td>
<td>0.055</td>
<td>0.054</td>
<td>0.071</td>
<td>0.74</td>
<td>0.498</td>
<td>0.505</td>
<td>0.955</td>
<td>0.981</td>
<td>0.948</td>
<td>0.993</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.9</td>
<td>40 x 5</td>
<td>0.501</td>
<td>0.027</td>
<td>0.028</td>
<td>0.035</td>
<td>0.76</td>
<td>0.499</td>
<td>0.502</td>
<td>0.948</td>
<td>0.993</td>
<td>0.968</td>
<td>0.998</td>
</tr>
<tr>
<td>CS</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.501</td>
<td>0.043</td>
<td>0.043</td>
<td>0.071</td>
<td>0.47</td>
<td>0.498</td>
<td>0.503</td>
<td>0.968</td>
<td>0.998</td>
<td>0.944</td>
<td>0.998</td>
</tr>
<tr>
<td>CS</td>
<td>0.1</td>
<td>40 x 5</td>
<td>0.500</td>
<td>0.022</td>
<td>0.022</td>
<td>0.035</td>
<td>0.46</td>
<td>0.499</td>
<td>0.502</td>
<td>0.944</td>
<td>0.998</td>
<td>0.944</td>
<td>0.998</td>
</tr>
<tr>
<td>CS</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.500</td>
<td>0.050</td>
<td>0.049</td>
<td>0.071</td>
<td>0.17</td>
<td>0.497</td>
<td>0.503</td>
<td>0.944</td>
<td>0.988</td>
<td>0.931</td>
<td>0.992</td>
</tr>
<tr>
<td>CS</td>
<td>0.5</td>
<td>40 x 5</td>
<td>0.499</td>
<td>0.025</td>
<td>0.025</td>
<td>0.035</td>
<td>-1.64</td>
<td>0.497</td>
<td>0.500</td>
<td>0.931</td>
<td>0.992</td>
<td>0.933</td>
<td>0.968</td>
</tr>
<tr>
<td>CS</td>
<td>0.9</td>
<td>10 x 5</td>
<td>0.499</td>
<td>0.062</td>
<td>0.059</td>
<td>0.071</td>
<td>-0.33</td>
<td>0.495</td>
<td>0.503</td>
<td>0.933</td>
<td>0.968</td>
<td>0.942</td>
<td>0.977</td>
</tr>
<tr>
<td>CS</td>
<td>0.9</td>
<td>40 x 5</td>
<td>0.500</td>
<td>0.031</td>
<td>0.030</td>
<td>0.035</td>
<td>-0.20</td>
<td>0.498</td>
<td>0.502</td>
<td>0.942</td>
<td>0.977</td>
<td>0.948</td>
<td>0.998</td>
</tr>
</tbody>
</table>

\( \phi(x) = 0.841 \)

<table>
<thead>
<tr>
<th>Cluster Struct</th>
<th>( \gamma_2 )</th>
<th>( n )</th>
<th>( F_n )</th>
<th>SE</th>
<th>Sim</th>
<th>SE</th>
<th>Ours</th>
<th>Naive</th>
<th>( t_{sim} )</th>
<th>95% Sim CI Low</th>
<th>95% Sim CI Hi</th>
<th>95% Coverage Ours</th>
<th>95% Coverage Naive</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR-1</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.839</td>
<td>0.030</td>
<td>0.030</td>
<td>0.052</td>
<td>-2.86</td>
<td>0.837</td>
<td>0.840</td>
<td>0.948</td>
<td>0.998</td>
<td>0.951</td>
<td>1.000</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.1</td>
<td>40 x 5</td>
<td>0.841</td>
<td>0.015</td>
<td>0.015</td>
<td>0.026</td>
<td>-1.37</td>
<td>0.840</td>
<td>0.842</td>
<td>0.951</td>
<td>1.000</td>
<td>0.957</td>
<td>0.993</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.839</td>
<td>0.032</td>
<td>0.031</td>
<td>0.052</td>
<td>-1.97</td>
<td>0.837</td>
<td>0.841</td>
<td>0.957</td>
<td>0.993</td>
<td>0.941</td>
<td>0.999</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.5</td>
<td>40 x 5</td>
<td>0.841</td>
<td>0.016</td>
<td>0.016</td>
<td>0.026</td>
<td>-1.19</td>
<td>0.840</td>
<td>0.842</td>
<td>0.941</td>
<td>0.999</td>
<td>0.935</td>
<td>0.976</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.9</td>
<td>10 x 5</td>
<td>0.837</td>
<td>0.038</td>
<td>0.037</td>
<td>0.052</td>
<td>-3.91</td>
<td>0.834</td>
<td>0.839</td>
<td>0.935</td>
<td>0.976</td>
<td>0.945</td>
<td>0.992</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.9</td>
<td>40 x 5</td>
<td>0.840</td>
<td>0.019</td>
<td>0.018</td>
<td>0.026</td>
<td>-2.73</td>
<td>0.839</td>
<td>0.841</td>
<td>0.945</td>
<td>0.992</td>
<td>0.936</td>
<td>0.992</td>
</tr>
<tr>
<td>CS</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.840</td>
<td>0.031</td>
<td>0.031</td>
<td>0.052</td>
<td>-1.65</td>
<td>0.838</td>
<td>0.842</td>
<td>0.936</td>
<td>0.992</td>
<td>0.947</td>
<td>0.998</td>
</tr>
<tr>
<td>CS</td>
<td>0.1</td>
<td>40 x 5</td>
<td>0.840</td>
<td>0.015</td>
<td>0.015</td>
<td>0.026</td>
<td>-3.43</td>
<td>0.839</td>
<td>0.841</td>
<td>0.947</td>
<td>0.998</td>
<td>0.953</td>
<td>0.982</td>
</tr>
<tr>
<td>CS</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.837</td>
<td>0.035</td>
<td>0.034</td>
<td>0.052</td>
<td>-3.61</td>
<td>0.835</td>
<td>0.840</td>
<td>0.953</td>
<td>0.982</td>
<td>0.952</td>
<td>0.996</td>
</tr>
<tr>
<td>CS</td>
<td>0.5</td>
<td>40 x 5</td>
<td>0.840</td>
<td>0.017</td>
<td>0.017</td>
<td>0.026</td>
<td>-2.42</td>
<td>0.839</td>
<td>0.841</td>
<td>0.952</td>
<td>0.996</td>
<td>0.950</td>
<td>0.982</td>
</tr>
<tr>
<td>CS</td>
<td>0.9</td>
<td>10 x 5</td>
<td>0.837</td>
<td>0.038</td>
<td>0.039</td>
<td>0.052</td>
<td>-3.89</td>
<td>0.834</td>
<td>0.839</td>
<td>0.950</td>
<td>0.982</td>
<td>0.947</td>
<td>0.991</td>
</tr>
<tr>
<td>CS</td>
<td>0.9</td>
<td>40 x 5</td>
<td>0.840</td>
<td>0.020</td>
<td>0.020</td>
<td>0.026</td>
<td>-2.62</td>
<td>0.838</td>
<td>0.841</td>
<td>0.947</td>
<td>0.991</td>
<td>0.957</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Simulations for clustered normally distributed data whose clusters have first-order autoregressive (AR-1) or compound symmetry (CS) correlation structures. The correlation parameter is indicated by \( \gamma_2 \). Each simulated data set had 10 or 40 clusters, each cluster having 5 observations each.
As an example of a spatial design, we considered the spatial-exponential model

\[
\text{Cov}[y_i, y_{i'}] = \gamma_1 \exp\left(-|u_i - u_{i'}|/\gamma_2\right),
\]

where \(u_i \in \mathbb{R}^2\) represents the point in space at which \(y_i\) was observed. For simulation purposes, the spatial coordinates \(u_i\) were generated as standard bivariate-normal random variables. For each simulated case we held the scale parameter \(\gamma_1\) equal to 1, but allowed the spatial autocorrelation parameter \(\gamma_2\) to vary among 0.45, 1.0, and 9.5. We remark that for the spatial-exponential model, the parameter \(\gamma_2\) cannot be interpreted directly as a correlation, and that our choices for \(\gamma_2\) correspond approximately to correlations of 0.1, 0.37, and 0.9 for points one unit in space apart.

The crossed effects model \(CS \times CS\) corresponds to the random effects model

\[
y_{ij} = \mu_{ij} + a_i + b_j + \varepsilon_{ij},
\]

where \(\varepsilon_{ij} \sim N(0, \gamma_1)\), \(a_i \sim N(0, \gamma_2)\), and \(b_j \sim N(0, \gamma_3)\), each error and random effect being drawn independently. In our simulations, we allowed \(i\) to range from 1 to \(k\) and \(j\) to range from 1 to \(n/k\), with each combination of \(i\) and \(j\) occurring once. We considered the cases \(k = 10\) and \(k = 50\) (for \(n = 200\)), combined with four choices of \(\gamma^T_0 = (\gamma_1, \gamma_2, \gamma_3)\):

\[
\gamma^T_0 = (1.0, 1.0, 1.0), \quad \gamma^T_0 = (1.0, 1.0, 0.5), \quad \gamma^T_0 = (1.0, 0.5, 1.0) \quad \text{and} \quad \gamma^T_0 = (1.0, 0.5, 0.5).
\]

The \(AR-1 \times CS\) model corresponds to the random effects model

\[
y_{ij} = \mu_{ij} + a_i + \varepsilon_{ij},
\]

where \(\varepsilon_{ij} \sim N(0, \sigma)\), and \(a_i \sim N(0, \sigma_a)\). As in the \(CS \times CS\) model, each error is drawn independently. However, the random effects \(a_i\) are not independent, but rather follow an
AR-1 time series. This random effects model leads to a covariance structure

\[ \text{Cov}[y_{ij}, y_{i'j'}] = \begin{cases} 
\gamma_1 + \gamma_2 & \text{if } i = i' \text{ and } j = j' \\
\gamma_2 |i-j|^{\gamma_3} & \text{otherwise}
\end{cases} \]

where \( \gamma_2 = \sigma_a \) and \( \gamma_3 \) represents a temporal autocorrelation parameter. Such a model may be useful when, for example, both spatial and temporal correlation are present and the spatial associations are minimal. The marginal variance of the model (with slightly different parametrization) can be expressed as the Kronecker product of two matrices, one which describes an AR-1 variance structure and the other which describes a CS variance structure. For our simulation, we considered \( n = 200 \), allowing \( i \) to range between 1 and 8 and \( j \) to range between 1 and 25. For each simulated case we held the scale parameter \( \gamma_1 \) equal to 1, but allowed \( \gamma_2 \) to be 1.0 or 0.5 and \( \gamma_3 \) to be 0.1 or 0.5.

Table 2 provides simulation results for time series and spatial models. Note that the estimator \( F_n \) seems reasonably unbiased for \( x = 0 \) and may be slightly biased for \( x = 1 \). As for clustered data, the standard error computed by (9) matches the simulation standard deviation, whereas the naive standard error does not. Our coverage probabilities for a nominal 95% confidence interval may be just slightly too large, but are still noticeably smaller than those obtained by using the naive standard error. Finally, the asymptotic behavior with respect to \( n \) appears to match that predicted by theory.

Results for \( CS \times CS \) and \( AR-1 \times CS \), summarized in Table 3, are similarly well-behaved. We remark that it is unclear that (18) holds for such models, nor is it clear that maximum likelihood leads to an estimator \( \hat{\theta}_n \) that has the stated asymptotic properties. Nevertheless, our method still produces good results.
Table 2: Simulations for Time Series and Spatial Models

\[ x = 0, \Phi(x) = 0.500 \]

<table>
<thead>
<tr>
<th>Struct</th>
<th>(\gamma_2)</th>
<th>(n)</th>
<th>(F_n)</th>
<th>SE</th>
<th>Ours</th>
<th>Naive</th>
<th>(t_{sim})</th>
<th>95% Sim CI Low</th>
<th>95% Sim CI Hi</th>
<th>95% Coverage Ours</th>
<th>95% Coverage Naive</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR-1</td>
<td>0.1</td>
<td>50</td>
<td>0.499</td>
<td>0.042</td>
<td>0.043</td>
<td>0.071</td>
<td>-0.37</td>
<td>0.497</td>
<td>0.902</td>
<td>0.964</td>
<td>0.999</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.1</td>
<td>200</td>
<td>0.500</td>
<td>0.021</td>
<td>0.021</td>
<td>0.035</td>
<td>0.50</td>
<td>0.499</td>
<td>0.502</td>
<td>0.962</td>
<td>0.999</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.5</td>
<td>50</td>
<td>0.500</td>
<td>0.042</td>
<td>0.043</td>
<td>0.071</td>
<td>0.16</td>
<td>0.498</td>
<td>0.503</td>
<td>0.968</td>
<td>0.999</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.5</td>
<td>200</td>
<td>0.499</td>
<td>0.022</td>
<td>0.021</td>
<td>0.035</td>
<td>-2.06</td>
<td>0.497</td>
<td>0.500</td>
<td>0.945</td>
<td>0.996</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.9</td>
<td>50</td>
<td>0.502</td>
<td>0.047</td>
<td>0.046</td>
<td>0.071</td>
<td>1.01</td>
<td>0.499</td>
<td>0.504</td>
<td>0.954</td>
<td>0.995</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.9</td>
<td>200</td>
<td>0.498</td>
<td>0.023</td>
<td>0.022</td>
<td>0.035</td>
<td>-2.98</td>
<td>0.496</td>
<td>0.499</td>
<td>0.944</td>
<td>0.998</td>
</tr>
<tr>
<td>Sp Exp</td>
<td>0.45</td>
<td>50</td>
<td>0.497</td>
<td>0.050</td>
<td>0.052</td>
<td>0.070</td>
<td>-1.82</td>
<td>0.494</td>
<td>0.500</td>
<td>0.952</td>
<td>0.988</td>
</tr>
<tr>
<td>Sp Exp</td>
<td>0.45</td>
<td>200</td>
<td>0.497</td>
<td>0.032</td>
<td>0.030</td>
<td>0.035</td>
<td>-3.03</td>
<td>0.495</td>
<td>0.499</td>
<td>0.943</td>
<td>0.968</td>
</tr>
<tr>
<td>Sp Exp</td>
<td>1</td>
<td>50</td>
<td>0.500</td>
<td>0.058</td>
<td>0.058</td>
<td>0.070</td>
<td>0.09</td>
<td>0.497</td>
<td>0.504</td>
<td>0.948</td>
<td>0.983</td>
</tr>
<tr>
<td>Sp Exp</td>
<td>1</td>
<td>200</td>
<td>0.501</td>
<td>0.031</td>
<td>0.032</td>
<td>0.035</td>
<td>1.08</td>
<td>0.499</td>
<td>0.503</td>
<td>0.957</td>
<td>0.969</td>
</tr>
<tr>
<td>Sp Exp</td>
<td>9.5</td>
<td>50</td>
<td>0.501</td>
<td>0.061</td>
<td>0.063</td>
<td>0.070</td>
<td>0.77</td>
<td>0.498</td>
<td>0.505</td>
<td>0.956</td>
<td>0.967</td>
</tr>
<tr>
<td>Sp Exp</td>
<td>9.5</td>
<td>200</td>
<td>0.499</td>
<td>0.035</td>
<td>0.034</td>
<td>0.035</td>
<td>-1.10</td>
<td>0.497</td>
<td>0.501</td>
<td>0.951</td>
<td>0.954</td>
</tr>
</tbody>
</table>

\[ x = 1, \Phi(x) = 0.841 \]

<table>
<thead>
<tr>
<th>Struct</th>
<th>(\gamma_2)</th>
<th>(n)</th>
<th>(F_n)</th>
<th>SE</th>
<th>Ours</th>
<th>Naive</th>
<th>(t_{sim})</th>
<th>95% Sim CI Low</th>
<th>95% Sim CI Hi</th>
<th>95% Coverage Ours</th>
<th>95% Coverage Naive</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR-1</td>
<td>0.1</td>
<td>50</td>
<td>0.839</td>
<td>0.029</td>
<td>0.030</td>
<td>0.052</td>
<td>-2.84</td>
<td>0.837</td>
<td>0.841</td>
<td>0.951</td>
<td>0.996</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.1</td>
<td>200</td>
<td>0.841</td>
<td>0.015</td>
<td>0.015</td>
<td>0.026</td>
<td>-0.50</td>
<td>0.840</td>
<td>0.842</td>
<td>0.947</td>
<td>1.000</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.5</td>
<td>50</td>
<td>0.840</td>
<td>0.030</td>
<td>0.030</td>
<td>0.052</td>
<td>-1.90</td>
<td>0.838</td>
<td>0.841</td>
<td>0.948</td>
<td>0.996</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.5</td>
<td>200</td>
<td>0.841</td>
<td>0.015</td>
<td>0.015</td>
<td>0.026</td>
<td>-0.96</td>
<td>0.840</td>
<td>0.842</td>
<td>0.951</td>
<td>0.999</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.9</td>
<td>50</td>
<td>0.838</td>
<td>0.034</td>
<td>0.032</td>
<td>0.052</td>
<td>-3.51</td>
<td>0.836</td>
<td>0.840</td>
<td>0.951</td>
<td>0.988</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.9</td>
<td>200</td>
<td>0.841</td>
<td>0.016</td>
<td>0.016</td>
<td>0.026</td>
<td>-0.82</td>
<td>0.840</td>
<td>0.842</td>
<td>0.935</td>
<td>0.999</td>
</tr>
<tr>
<td>Sp Exp</td>
<td>0.45</td>
<td>50</td>
<td>0.840</td>
<td>0.035</td>
<td>0.035</td>
<td>0.051</td>
<td>-1.65</td>
<td>0.837</td>
<td>0.842</td>
<td>0.962</td>
<td>0.988</td>
</tr>
<tr>
<td>Sp Exp</td>
<td>0.45</td>
<td>200</td>
<td>0.840</td>
<td>0.020</td>
<td>0.020</td>
<td>0.026</td>
<td>-2.55</td>
<td>0.839</td>
<td>0.841</td>
<td>0.952</td>
<td>0.994</td>
</tr>
<tr>
<td>Sp Exp</td>
<td>1</td>
<td>50</td>
<td>0.839</td>
<td>0.039</td>
<td>0.038</td>
<td>0.051</td>
<td>-2.09</td>
<td>0.836</td>
<td>0.841</td>
<td>0.936</td>
<td>0.971</td>
</tr>
<tr>
<td>Sp Exp</td>
<td>1</td>
<td>200</td>
<td>0.842</td>
<td>0.020</td>
<td>0.021</td>
<td>0.026</td>
<td>0.24</td>
<td>0.840</td>
<td>0.843</td>
<td>0.950</td>
<td>0.984</td>
</tr>
<tr>
<td>Sp Exp</td>
<td>9.5</td>
<td>50</td>
<td>0.839</td>
<td>0.041</td>
<td>0.041</td>
<td>0.051</td>
<td>-1.77</td>
<td>0.837</td>
<td>0.842</td>
<td>0.955</td>
<td>0.956</td>
</tr>
<tr>
<td>Sp Exp</td>
<td>9.5</td>
<td>200</td>
<td>0.841</td>
<td>0.022</td>
<td>0.022</td>
<td>0.026</td>
<td>-1.17</td>
<td>0.839</td>
<td>0.842</td>
<td>0.952</td>
<td>0.984</td>
</tr>
</tbody>
</table>

Simulations for normally distributed data having first order autoregressive time-series structure or spatial exponential structure. For each structure type, the correlation parameter, described in the text, is indicated by \(\gamma_2\). Each simulated data set had 50 or 200 observations.
Table 3: Simulations for Highly Correlated Errors

CS × CS structure, \( x = 0, \Phi(x) = 0.500 \)

<table>
<thead>
<tr>
<th>( \gamma_2 )</th>
<th>( \gamma_3 )</th>
<th>( n )</th>
<th>( \bar{F}_n )</th>
<th>( \text{SE} )</th>
<th>( t_{\text{sim}} )</th>
<th>( 95% \text{ Sim CI} )</th>
<th>( 95% \text{ Coverage} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>1.0</td>
<td>20 \times 10</td>
<td>0.500</td>
<td>0.027</td>
<td>0.027</td>
<td>0.035</td>
<td>0.53</td>
</tr>
<tr>
<td>1.0</td>
<td>0.5</td>
<td>50 \times 4</td>
<td>0.500</td>
<td>0.028</td>
<td>0.028</td>
<td>0.035</td>
<td>-0.32</td>
</tr>
<tr>
<td>1.0</td>
<td>0.5</td>
<td>20 \times 10</td>
<td>0.500</td>
<td>0.026</td>
<td>0.027</td>
<td>0.035</td>
<td>0.60</td>
</tr>
<tr>
<td>0.5</td>
<td>1.0</td>
<td>20 \times 10</td>
<td>0.501</td>
<td>0.026</td>
<td>0.027</td>
<td>0.035</td>
<td>0.98</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>20 \times 10</td>
<td>0.501</td>
<td>0.025</td>
<td>0.026</td>
<td>0.035</td>
<td>1.21</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>50 \times 4</td>
<td>0.500</td>
<td>0.026</td>
<td>0.026</td>
<td>0.035</td>
<td>-0.53</td>
</tr>
</tbody>
</table>

CS × CS structure, \( x = 1, \Phi(x) = 0.841 \)

<table>
<thead>
<tr>
<th>( \gamma_2 )</th>
<th>( \gamma_3 )</th>
<th>( n )</th>
<th>( \bar{F}_n )</th>
<th>( \text{SE} )</th>
<th>( t_{\text{sim}} )</th>
<th>( 95% \text{ Sim CI} )</th>
<th>( 95% \text{ Coverage} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>1.0</td>
<td>20 \times 10</td>
<td>0.840</td>
<td>0.018</td>
<td>0.018</td>
<td>0.026</td>
<td>-1.79</td>
</tr>
<tr>
<td>1.0</td>
<td>1.0</td>
<td>50 \times 4</td>
<td>0.839</td>
<td>0.018</td>
<td>0.018</td>
<td>0.026</td>
<td>-3.25</td>
</tr>
<tr>
<td>1.0</td>
<td>0.5</td>
<td>20 \times 10</td>
<td>0.840</td>
<td>0.018</td>
<td>0.018</td>
<td>0.026</td>
<td>-2.14</td>
</tr>
<tr>
<td>1.0</td>
<td>0.5</td>
<td>50 \times 4</td>
<td>0.840</td>
<td>0.018</td>
<td>0.018</td>
<td>0.026</td>
<td>-1.67</td>
</tr>
<tr>
<td>0.5</td>
<td>1.0</td>
<td>20 \times 10</td>
<td>0.840</td>
<td>0.019</td>
<td>0.019</td>
<td>0.026</td>
<td>-1.77</td>
</tr>
<tr>
<td>0.5</td>
<td>1.0</td>
<td>50 \times 4</td>
<td>0.840</td>
<td>0.017</td>
<td>0.017</td>
<td>0.026</td>
<td>-2.27</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>50 \times 4</td>
<td>0.840</td>
<td>0.018</td>
<td>0.018</td>
<td>0.026</td>
<td>-2.29</td>
</tr>
</tbody>
</table>

AR-1 × CS Structure, \( x = 0, \Phi(x) = 0.500 \)

<table>
<thead>
<tr>
<th>( \gamma_2 )</th>
<th>( \gamma_3 )</th>
<th>( n )</th>
<th>( \bar{F}_n )</th>
<th>( \text{SE} )</th>
<th>( t_{\text{sim}} )</th>
<th>( 95% \text{ Sim CI} )</th>
<th>( 95% \text{ Coverage} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>0.1</td>
<td>25 \times 8</td>
<td>0.501</td>
<td>0.023</td>
<td>0.024</td>
<td>0.035</td>
<td>1.35</td>
</tr>
<tr>
<td>1.0</td>
<td>0.5</td>
<td>25 \times 8</td>
<td>0.501</td>
<td>0.024</td>
<td>0.024</td>
<td>0.035</td>
<td>0.74</td>
</tr>
<tr>
<td>0.5</td>
<td>0.1</td>
<td>25 \times 8</td>
<td>0.500</td>
<td>0.026</td>
<td>0.026</td>
<td>0.035</td>
<td>0.38</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>25 \times 8</td>
<td>0.501</td>
<td>0.026</td>
<td>0.026</td>
<td>0.035</td>
<td>1.57</td>
</tr>
<tr>
<td>5.0</td>
<td>0.1</td>
<td>25 \times 8</td>
<td>0.500</td>
<td>0.029</td>
<td>0.030</td>
<td>0.035</td>
<td>0.05</td>
</tr>
<tr>
<td>5.0</td>
<td>0.5</td>
<td>25 \times 8</td>
<td>0.500</td>
<td>0.030</td>
<td>0.030</td>
<td>0.035</td>
<td>0.46</td>
</tr>
</tbody>
</table>

AR-1 × CS Structure, \( x = 1, \Phi(x) = 0.841 \)

<table>
<thead>
<tr>
<th>( \gamma_2 )</th>
<th>( \gamma_3 )</th>
<th>( n )</th>
<th>( \bar{F}_n )</th>
<th>( \text{SE} )</th>
<th>( t_{\text{sim}} )</th>
<th>( 95% \text{ Sim CI} )</th>
<th>( 95% \text{ Coverage} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>0.1</td>
<td>25 \times 8</td>
<td>0.840</td>
<td>0.017</td>
<td>0.017</td>
<td>0.026</td>
<td>-1.64</td>
</tr>
<tr>
<td>1.0</td>
<td>0.5</td>
<td>25 \times 8</td>
<td>0.840</td>
<td>0.016</td>
<td>0.017</td>
<td>0.026</td>
<td>-2.09</td>
</tr>
<tr>
<td>0.5</td>
<td>0.1</td>
<td>25 \times 8</td>
<td>0.841</td>
<td>0.018</td>
<td>0.018</td>
<td>0.026</td>
<td>0.72</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>25 \times 8</td>
<td>0.840</td>
<td>0.017</td>
<td>0.017</td>
<td>0.026</td>
<td>-1.56</td>
</tr>
<tr>
<td>5.0</td>
<td>0.1</td>
<td>25 \times 8</td>
<td>0.840</td>
<td>0.021</td>
<td>0.030</td>
<td>0.026</td>
<td>2.67</td>
</tr>
<tr>
<td>5.0</td>
<td>0.5</td>
<td>25 \times 8</td>
<td>0.841</td>
<td>0.019</td>
<td>0.030</td>
<td>0.026</td>
<td>-0.08</td>
</tr>
</tbody>
</table>

Simulations for normally distributed data having CS × CS or AR-1 × CS correlation structures. The structures, along with definitions of \( \gamma_2 \) and \( \gamma_3 \), are described in the text.
4.3 Clustered Data Contaminated by Non-Normal Distributions

To examine the sensitivity of our method to non-normal errors, we considered the two types of clustered designs described in Section 4.1 for $\gamma_2$ equal to 0.1 or 0.5. However, instead of using a normal distribution for each element of the error vector, we randomly contaminated the errors with samples from a heavy tailed ($t_3$) distribution or with samples from a skewed (standardized $\chi^2_3$) distribution. Contamination levels were set to 10\%, 25\%, and 50\%. Note that we contaminated the $z \sim N_n(0, I_n)$ vector before multiplying it by the Cholesky matrix of the true variance $V(\gamma_0)$, so that the resulting residual vector would have zero mean and approximately true variance $V(\gamma_0)$.

Results are summarized in Tables 4, 5, 6, and 7. The results suggest that our methods are somewhat sensitive to deviations from normality. $F_n$ can be substantially biased when the residual vector $z$ does not have a normal distribution, leading to the detection of non-normal residual vectors. Additionally, coverage probabilities of confidence intervals are lower than the nominal level, which also suggests sensitivity to non-normal errors. We remark that in general the coverage probabilities for our confidence intervals are lower than those produced by naive confidence intervals, which suggests that rotated residuals are more sensitive than standardized residuals. Finally, we note that for heavy tailed distributions, the bias appears strongest at $x = 1$, whereas for skewed distributions, the bias appears strongest at $x = 0$. This is further evidence that the method behaves as it should.
Table 4: Simulations for Clustered AR-1 Data Contaminated by a Heavy-Tailed Distribution

$x = 0, \Phi(x) = 0.500$

<table>
<thead>
<tr>
<th>Frac</th>
<th>$\gamma_2$</th>
<th>n</th>
<th>$F_n$</th>
<th>Sim SD</th>
<th>SE Ours</th>
<th>SE Naïve</th>
<th>$t_{sim}$</th>
<th>95% Sim CI Low</th>
<th>95% Sim CI Hi</th>
<th>95% Coverage Ours</th>
<th>95% Coverage Naïve</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.500</td>
<td>0.047</td>
<td>0.043</td>
<td>0.070</td>
<td>-0.31</td>
<td>0.407</td>
<td>0.502</td>
<td>0.951</td>
<td>0.997</td>
</tr>
<tr>
<td>0.10</td>
<td>0.5</td>
<td>40 x 5</td>
<td>0.500</td>
<td>0.023</td>
<td>0.021</td>
<td>0.035</td>
<td>-0.18</td>
<td>0.498</td>
<td>0.501</td>
<td>0.934</td>
<td>0.994</td>
</tr>
<tr>
<td>0.10</td>
<td>0.5</td>
<td>40 x 5</td>
<td>0.501</td>
<td>0.050</td>
<td>0.045</td>
<td>0.070</td>
<td>0.38</td>
<td>0.497</td>
<td>0.504</td>
<td>0.929</td>
<td>0.985</td>
</tr>
<tr>
<td>0.10</td>
<td>0.5</td>
<td>40 x 5</td>
<td>0.500</td>
<td>0.024</td>
<td>0.022</td>
<td>0.035</td>
<td>0.18</td>
<td>0.499</td>
<td>0.502</td>
<td>0.938</td>
<td>0.993</td>
</tr>
<tr>
<td>0.25</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.502</td>
<td>0.049</td>
<td>0.043</td>
<td>0.070</td>
<td>1.02</td>
<td>0.499</td>
<td>0.505</td>
<td>0.933</td>
<td>0.989</td>
</tr>
<tr>
<td>0.25</td>
<td>0.1</td>
<td>40 x 5</td>
<td>0.499</td>
<td>0.025</td>
<td>0.021</td>
<td>0.035</td>
<td>-1.32</td>
<td>0.497</td>
<td>0.501</td>
<td>0.911</td>
<td>0.989</td>
</tr>
<tr>
<td>0.25</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.499</td>
<td>0.053</td>
<td>0.045</td>
<td>0.070</td>
<td>-0.80</td>
<td>0.495</td>
<td>0.502</td>
<td>0.914</td>
<td>0.983</td>
</tr>
<tr>
<td>0.50</td>
<td>0.1</td>
<td>40 x 5</td>
<td>0.500</td>
<td>0.027</td>
<td>0.022</td>
<td>0.035</td>
<td>1.42</td>
<td>0.500</td>
<td>0.503</td>
<td>0.888</td>
<td>0.985</td>
</tr>
<tr>
<td>0.50</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.500</td>
<td>0.055</td>
<td>0.043</td>
<td>0.070</td>
<td>-0.28</td>
<td>0.496</td>
<td>0.503</td>
<td>0.895</td>
<td>0.975</td>
</tr>
<tr>
<td>0.50</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.489</td>
<td>0.031</td>
<td>0.022</td>
<td>0.035</td>
<td>-0.08</td>
<td>0.489</td>
<td>0.502</td>
<td>0.844</td>
<td>0.969</td>
</tr>
<tr>
<td>0.50</td>
<td>0.5</td>
<td>40 x 5</td>
<td>0.500</td>
<td>0.056</td>
<td>0.045</td>
<td>0.070</td>
<td>0.17</td>
<td>0.407</td>
<td>0.500</td>
<td>0.894</td>
<td>0.982</td>
</tr>
</tbody>
</table>

Simulations for non-normal data whose clusters have first-order autoregressive (AR-1) correlation structures. In each case, a $N(0, I)$ vector was randomly contaminated by a heavy-tailed distribution ($t_3$), with 10%, 25%, or 50% contamination, indicated in the column entitled Frac. The resulting vector was rotated to have CS structure, with correlation parameter indicated by $\gamma_2$, and used as the marginal error in a linear model. Each simulated data set had 10 or 40 clusters, each cluster having 5 observations each.
Table 5: Simulations for Clustered AR-1 Data Contaminated by a Skewed Distribution

\[ x = 0, \Phi(x) = 0.500 \]

<table>
<thead>
<tr>
<th>Frac</th>
<th>( \gamma_2 )</th>
<th>( n )</th>
<th>( F_n )</th>
<th>Sim SE</th>
<th>Ours Naive</th>
<th>( t_{sim} )</th>
<th>95% Sim CI Low</th>
<th>Hi</th>
<th>95% Coverage Ours Naive</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.500</td>
<td>0.042</td>
<td>0.043</td>
<td>0.070</td>
<td>7.09</td>
<td>0.507</td>
<td>0.512       0.968  0.999</td>
</tr>
<tr>
<td>0.10</td>
<td>0.1</td>
<td>40 x 5</td>
<td>0.511</td>
<td>0.022</td>
<td>0.021</td>
<td>0.035</td>
<td>15.62</td>
<td>0.509</td>
<td>0.512       0.919  0.997</td>
</tr>
<tr>
<td>0.10</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.611</td>
<td>0.044</td>
<td>0.045</td>
<td>0.070</td>
<td>7.65</td>
<td>0.508</td>
<td>0.513       0.959  0.999</td>
</tr>
<tr>
<td>0.10</td>
<td>0.5</td>
<td>40 x 5</td>
<td>0.510</td>
<td>0.022</td>
<td>0.022</td>
<td>0.035</td>
<td>14.65</td>
<td>0.509</td>
<td>0.512       0.914  0.993</td>
</tr>
<tr>
<td>0.25</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.523</td>
<td>0.042</td>
<td>0.043</td>
<td>0.070</td>
<td>17.04</td>
<td>0.520</td>
<td>0.525       0.947  0.996</td>
</tr>
<tr>
<td>0.25</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.527</td>
<td>0.022</td>
<td>0.021</td>
<td>0.035</td>
<td>38.94</td>
<td>0.526</td>
<td>0.529       0.759  0.967</td>
</tr>
<tr>
<td>0.25</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.525</td>
<td>0.045</td>
<td>0.045</td>
<td>0.070</td>
<td>17.35</td>
<td>0.522</td>
<td>0.527       0.918  0.992</td>
</tr>
<tr>
<td>0.25</td>
<td>0.5</td>
<td>40 x 5</td>
<td>0.526</td>
<td>0.022</td>
<td>0.022</td>
<td>0.035</td>
<td>37.93</td>
<td>0.525</td>
<td>0.528       0.768  0.971</td>
</tr>
<tr>
<td>0.50</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.554</td>
<td>0.043</td>
<td>0.043</td>
<td>0.070</td>
<td>40.18</td>
<td>0.552</td>
<td>0.557       0.795  0.967</td>
</tr>
<tr>
<td>0.50</td>
<td>0.1</td>
<td>40 x 5</td>
<td>0.554</td>
<td>0.022</td>
<td>0.021</td>
<td>0.035</td>
<td>76.63</td>
<td>0.553</td>
<td>0.556       0.308  0.705</td>
</tr>
<tr>
<td>0.50</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.550</td>
<td>0.043</td>
<td>0.045</td>
<td>0.070</td>
<td>36.50</td>
<td>0.547</td>
<td>0.553       0.820  0.973</td>
</tr>
<tr>
<td>0.50</td>
<td>0.5</td>
<td>40 x 5</td>
<td>0.552</td>
<td>0.022</td>
<td>0.022</td>
<td>0.035</td>
<td>76.19</td>
<td>0.551</td>
<td>0.553       0.328  0.757</td>
</tr>
</tbody>
</table>

\[ x = 1, \Phi(x) = 0.841 \]

<table>
<thead>
<tr>
<th>Frac</th>
<th>( \gamma_2 )</th>
<th>( n )</th>
<th>( F_n )</th>
<th>Sim SE</th>
<th>Ours Naive</th>
<th>( t_{sim} )</th>
<th>95% Sim CI Low</th>
<th>Hi</th>
<th>95% Coverage Ours Naive</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.839</td>
<td>0.031</td>
<td>0.030</td>
<td>0.052</td>
<td>-2.06</td>
<td>0.837</td>
<td>0.841       0.934  0.985</td>
</tr>
<tr>
<td>0.10</td>
<td>0.1</td>
<td>40 x 5</td>
<td>0.842</td>
<td>0.016</td>
<td>0.015</td>
<td>0.026</td>
<td>1.44</td>
<td>0.841</td>
<td>0.843       0.941  0.996</td>
</tr>
<tr>
<td>0.10</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.840</td>
<td>0.031</td>
<td>0.031</td>
<td>0.051</td>
<td>-1.38</td>
<td>0.838</td>
<td>0.842       0.958  0.987</td>
</tr>
<tr>
<td>0.10</td>
<td>0.5</td>
<td>40 x 5</td>
<td>0.842</td>
<td>0.017</td>
<td>0.016</td>
<td>0.026</td>
<td>1.36</td>
<td>0.841</td>
<td>0.843       0.927  0.998</td>
</tr>
<tr>
<td>0.25</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.841</td>
<td>0.031</td>
<td>0.030</td>
<td>0.051</td>
<td>-0.49</td>
<td>0.839</td>
<td>0.843       0.941  0.984</td>
</tr>
<tr>
<td>0.25</td>
<td>0.1</td>
<td>40 x 5</td>
<td>0.841</td>
<td>0.016</td>
<td>0.015</td>
<td>0.026</td>
<td>5.90</td>
<td>0.843</td>
<td>0.845       0.937  0.993</td>
</tr>
<tr>
<td>0.25</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.841</td>
<td>0.033</td>
<td>0.031</td>
<td>0.051</td>
<td>-0.30</td>
<td>0.839</td>
<td>0.843       0.939  0.986</td>
</tr>
<tr>
<td>0.25</td>
<td>0.5</td>
<td>40 x 5</td>
<td>0.845</td>
<td>0.017</td>
<td>0.016</td>
<td>0.026</td>
<td>6.10</td>
<td>0.844</td>
<td>0.846       0.927  0.994</td>
</tr>
<tr>
<td>0.50</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.849</td>
<td>0.034</td>
<td>0.030</td>
<td>0.050</td>
<td>6.94</td>
<td>0.847</td>
<td>0.851       0.916  0.960</td>
</tr>
<tr>
<td>0.50</td>
<td>0.1</td>
<td>40 x 5</td>
<td>0.848</td>
<td>0.018</td>
<td>0.015</td>
<td>0.025</td>
<td>12.15</td>
<td>0.847</td>
<td>0.849       0.885  0.991</td>
</tr>
<tr>
<td>0.50</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.844</td>
<td>0.034</td>
<td>0.031</td>
<td>0.051</td>
<td>2.47</td>
<td>0.842</td>
<td>0.846       0.941  0.975</td>
</tr>
<tr>
<td>0.50</td>
<td>0.5</td>
<td>40 x 5</td>
<td>0.848</td>
<td>0.017</td>
<td>0.016</td>
<td>0.025</td>
<td>13.24</td>
<td>0.847</td>
<td>0.850       0.916  0.986</td>
</tr>
</tbody>
</table>

Simulations for non-normal data whose clusters have first-order autoregressive (AR-1) correlation structures. In each case, a \( N(0, I) \) vector was randomly contaminated by a skewed distribution (standardized \( \chi^2_3 \)), with 10\%, 25\%, or 50\% contamination, indicated by the column entitled \( Frac \). The resulting vector was rotated to have CS structure, with correlation parameter indicated by \( \gamma_2 \), and used as the marginal error in a linear model. Each simulated data set had 10 or 40 clusters, each cluster having 5 observations each.
Table 6: Simulations for Clustered Compound Symmetry Data Contaminated by a Heavy-Tailed Distribution

For $x = 0$, $\Phi(x) = 0.500$

<table>
<thead>
<tr>
<th>Frac</th>
<th>$\gamma_2$</th>
<th>$n$</th>
<th>Sim $F_n$</th>
<th>SE Ours</th>
<th>Naive $t_{sim}$</th>
<th>95% Sim CI Low</th>
<th>95% Sim CI Hi</th>
<th>95% Coverage Ours</th>
<th>95% Coverage Naive</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.501</td>
<td>0.049</td>
<td>0.043</td>
<td>0.070</td>
<td>0.72</td>
<td>0.498</td>
<td>0.504</td>
</tr>
<tr>
<td>0.10</td>
<td>0.1</td>
<td>40 x 5</td>
<td>0.500</td>
<td>0.024</td>
<td>0.022</td>
<td>0.035</td>
<td>0.00</td>
<td>0.499</td>
<td>0.501</td>
</tr>
<tr>
<td>0.10</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.499</td>
<td>0.052</td>
<td>0.049</td>
<td>0.070</td>
<td>-0.33</td>
<td>0.496</td>
<td>0.503</td>
</tr>
<tr>
<td>0.10</td>
<td>0.5</td>
<td>40 x 5</td>
<td>0.500</td>
<td>0.026</td>
<td>0.025</td>
<td>0.035</td>
<td>0.19</td>
<td>0.499</td>
<td>0.502</td>
</tr>
<tr>
<td>0.25</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.499</td>
<td>0.051</td>
<td>0.043</td>
<td>0.070</td>
<td>-0.67</td>
<td>0.496</td>
<td>0.502</td>
</tr>
<tr>
<td>0.25</td>
<td>0.1</td>
<td>40 x 5</td>
<td>0.500</td>
<td>0.027</td>
<td>0.022</td>
<td>0.035</td>
<td>0.57</td>
<td>0.499</td>
<td>0.502</td>
</tr>
<tr>
<td>0.25</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.499</td>
<td>0.054</td>
<td>0.049</td>
<td>0.070</td>
<td>-0.54</td>
<td>0.496</td>
<td>0.502</td>
</tr>
<tr>
<td>0.25</td>
<td>0.5</td>
<td>40 x 5</td>
<td>0.502</td>
<td>0.029</td>
<td>0.025</td>
<td>0.035</td>
<td>2.08</td>
<td>0.500</td>
<td>0.504</td>
</tr>
<tr>
<td>0.50</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.501</td>
<td>0.057</td>
<td>0.043</td>
<td>0.070</td>
<td>0.65</td>
<td>0.498</td>
<td>0.505</td>
</tr>
<tr>
<td>0.50</td>
<td>0.1</td>
<td>40 x 5</td>
<td>0.500</td>
<td>0.031</td>
<td>0.022</td>
<td>0.035</td>
<td>-0.68</td>
<td>0.498</td>
<td>0.502</td>
</tr>
<tr>
<td>0.50</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.500</td>
<td>0.059</td>
<td>0.049</td>
<td>0.070</td>
<td>-0.02</td>
<td>0.496</td>
<td>0.504</td>
</tr>
<tr>
<td>0.50</td>
<td>0.5</td>
<td>40 x 5</td>
<td>0.500</td>
<td>0.030</td>
<td>0.025</td>
<td>0.035</td>
<td>0.10</td>
<td>0.498</td>
<td>0.502</td>
</tr>
</tbody>
</table>

For $x = 1$, $\Phi(x) = 0.841$

<table>
<thead>
<tr>
<th>Frac</th>
<th>$\gamma_2$</th>
<th>$n$</th>
<th>Sim $F_n$</th>
<th>SE Ours</th>
<th>Naive $t_{sim}$</th>
<th>95% Sim CI Low</th>
<th>95% Sim CI Hi</th>
<th>95% Coverage Ours</th>
<th>95% Coverage Naive</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.850</td>
<td>0.037</td>
<td>0.031</td>
<td>0.050</td>
<td>6.96</td>
<td>0.847</td>
<td>0.852</td>
</tr>
<tr>
<td>0.10</td>
<td>0.1</td>
<td>40 x 5</td>
<td>0.854</td>
<td>0.023</td>
<td>0.015</td>
<td>0.025</td>
<td>16.49</td>
<td>0.852</td>
<td>0.855</td>
</tr>
<tr>
<td>0.10</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.849</td>
<td>0.039</td>
<td>0.034</td>
<td>0.050</td>
<td>5.83</td>
<td>0.846</td>
<td>0.851</td>
</tr>
<tr>
<td>0.10</td>
<td>0.5</td>
<td>40 x 5</td>
<td>0.854</td>
<td>0.023</td>
<td>0.017</td>
<td>0.025</td>
<td>17.84</td>
<td>0.853</td>
<td>0.856</td>
</tr>
<tr>
<td>0.25</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.859</td>
<td>0.039</td>
<td>0.031</td>
<td>0.048</td>
<td>14.61</td>
<td>0.857</td>
<td>0.862</td>
</tr>
<tr>
<td>0.25</td>
<td>0.1</td>
<td>40 x 5</td>
<td>0.871</td>
<td>0.028</td>
<td>0.015</td>
<td>0.023</td>
<td>33.61</td>
<td>0.869</td>
<td>0.873</td>
</tr>
<tr>
<td>0.25</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.856</td>
<td>0.041</td>
<td>0.034</td>
<td>0.049</td>
<td>11.28</td>
<td>0.853</td>
<td>0.859</td>
</tr>
<tr>
<td>0.25</td>
<td>0.5</td>
<td>40 x 5</td>
<td>0.867</td>
<td>0.027</td>
<td>0.017</td>
<td>0.024</td>
<td>30.17</td>
<td>0.866</td>
<td>0.869</td>
</tr>
<tr>
<td>0.50</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.872</td>
<td>0.041</td>
<td>0.031</td>
<td>0.046</td>
<td>24.13</td>
<td>0.870</td>
<td>0.875</td>
</tr>
<tr>
<td>0.50</td>
<td>0.1</td>
<td>40 x 5</td>
<td>0.887</td>
<td>0.028</td>
<td>0.015</td>
<td>0.022</td>
<td>50.65</td>
<td>0.885</td>
<td>0.888</td>
</tr>
<tr>
<td>0.50</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.872</td>
<td>0.042</td>
<td>0.033</td>
<td>0.046</td>
<td>23.05</td>
<td>0.869</td>
<td>0.874</td>
</tr>
<tr>
<td>0.50</td>
<td>0.5</td>
<td>40 x 5</td>
<td>0.886</td>
<td>0.028</td>
<td>0.017</td>
<td>0.022</td>
<td>50.45</td>
<td>0.884</td>
<td>0.888</td>
</tr>
</tbody>
</table>

Simulations for non-normal data whose clusters have compound symmetry (CS) correlation structures. In each case, a $N(0, I)$ vector was randomly contaminated by a heavy-tailed distribution ($t_\gamma$), with 10%, 25%, or 50% contamination, indicated by the column entitled Frac. The resulting vector was rotated to have CS structure, with correlation parameter indicated by $\gamma_2$, and used as the marginal error in a linear model. Each simulated data set had 10 or 40 clusters, each cluster having 5 observations each.
Table 7: Simulations for Clustered Compound Symmetry Data Contaminated by a Skewed Distribution

\[ x = 0, \Phi(x) = 0.500 \]

<table>
<thead>
<tr>
<th>Frac</th>
<th>( \gamma_2 )</th>
<th>( n )</th>
<th>Sim</th>
<th>SE</th>
<th>95% Sim CI</th>
<th>95% Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.511</td>
<td>0.043</td>
<td>0.043</td>
<td>0.070</td>
</tr>
<tr>
<td>0.10</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.511</td>
<td>0.021</td>
<td>0.022</td>
<td>0.035</td>
</tr>
<tr>
<td>0.10</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.512</td>
<td>0.049</td>
<td>0.049</td>
<td>0.070</td>
</tr>
<tr>
<td>0.10</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.511</td>
<td>0.025</td>
<td>0.025</td>
<td>0.035</td>
</tr>
<tr>
<td>0.25</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.522</td>
<td>0.044</td>
<td>0.043</td>
<td>0.070</td>
</tr>
<tr>
<td>0.25</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.527</td>
<td>0.022</td>
<td>0.022</td>
<td>0.035</td>
</tr>
<tr>
<td>0.25</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.524</td>
<td>0.048</td>
<td>0.049</td>
<td>0.070</td>
</tr>
<tr>
<td>0.25</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.528</td>
<td>0.025</td>
<td>0.025</td>
<td>0.035</td>
</tr>
<tr>
<td>0.50</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.551</td>
<td>0.044</td>
<td>0.043</td>
<td>0.070</td>
</tr>
<tr>
<td>0.50</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.553</td>
<td>0.022</td>
<td>0.022</td>
<td>0.035</td>
</tr>
<tr>
<td>0.50</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.551</td>
<td>0.051</td>
<td>0.049</td>
<td>0.070</td>
</tr>
<tr>
<td>0.50</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.553</td>
<td>0.024</td>
<td>0.025</td>
<td>0.035</td>
</tr>
</tbody>
</table>

\[ x = 1, \Phi(x) = 0.841 \]

<table>
<thead>
<tr>
<th>Frac</th>
<th>( \gamma_2 )</th>
<th>( n )</th>
<th>Sim</th>
<th>SE</th>
<th>95% Sim CI</th>
<th>95% Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.838</td>
<td>0.031</td>
<td>0.031</td>
<td>0.052</td>
</tr>
<tr>
<td>0.10</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.843</td>
<td>0.016</td>
<td>0.015</td>
<td>0.026</td>
</tr>
<tr>
<td>0.10</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.838</td>
<td>0.035</td>
<td>0.034</td>
<td>0.052</td>
</tr>
<tr>
<td>0.10</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.842</td>
<td>0.017</td>
<td>0.017</td>
<td>0.026</td>
</tr>
<tr>
<td>0.25</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.841</td>
<td>0.031</td>
<td>0.031</td>
<td>0.051</td>
</tr>
<tr>
<td>0.25</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.844</td>
<td>0.017</td>
<td>0.015</td>
<td>0.026</td>
</tr>
<tr>
<td>0.25</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.839</td>
<td>0.035</td>
<td>0.034</td>
<td>0.052</td>
</tr>
<tr>
<td>0.25</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.846</td>
<td>0.018</td>
<td>0.017</td>
<td>0.025</td>
</tr>
<tr>
<td>0.50</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.846</td>
<td>0.034</td>
<td>0.031</td>
<td>0.051</td>
</tr>
<tr>
<td>0.50</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.848</td>
<td>0.018</td>
<td>0.015</td>
<td>0.025</td>
</tr>
<tr>
<td>0.50</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.845</td>
<td>0.035</td>
<td>0.034</td>
<td>0.051</td>
</tr>
<tr>
<td>0.50</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.849</td>
<td>0.019</td>
<td>0.017</td>
<td>0.025</td>
</tr>
</tbody>
</table>

Simulations for non-normal data whose clusters have compound symmetry (CS) correlation structures. In each case, a \( N(0, I) \) vector was randomly contaminated by a skewed distribution (standardized \( \chi^2 \)), with 10%, 25%, or 50% contamination, indicated by the column entitled \( \text{Frac} \). The resulting vector was rotated to have CS structure, with correlation parameter indicated by \( \gamma_2 \), and used as the marginal error in a linear model. Each simulated data set had 10 or 40 clusters, each cluster having 5 observations each.
4.4 Clustered Data with Misspecified Correlation Structure within Clusters

To examine the sensitivity of our method to incorrectly specified covariance structures, we considered the two types of clustered designs described in Section 4.1. However, when the true covariance structure was compound-symmetry, we used an AR-1 model to analyze the data; and when the true covariance structure was AR-1 we used a compound-symmetry model to analyze the data.

Results are summarized in Table 8. We note that our method does not appear to be very sensitive to misspecified covariance structure, as $F_n$ seems unbiased and the coverage probabilities match their nominal levels. Since the residual vector $z$ is still normally distributed (with non-diagonal variance) when $L(\gamma_0)^T$ is incorrect, it may be that the empirical CDF is an inappropriate method for detecting goodness-of-fit with respect to covariance structure. We speculate that using a projection matrix other than the canonical $P_t$ in the construction of (4) [and $m_t(\beta, \gamma)$ and $s_t(\gamma)$] may increase sensitivity to covariance misspecification, and we plan to pursue this in the future.

4.5 Residuals of Linear Models with Known Parameters

The poor behavior of the “naive” standard errors, unadjusted for estimation of $\theta$, is to be expected. However, it is not immediately obvious whether this is the case only because the estimation of $\theta$ is not taken into account, or because the very failure to rotate the residuals leads to inaccurate standard errors. To address this question, we investigated the behavior of rotated and standardized (unrotated) residuals when all parameters are known. That is, using the known marginal error, we compared the empirical CDF using the rotated vector (known to have $N_n(0, I_n)$ distribution) against the empirical CDF using
Table 8: Simulations for Clustered Data with Misspecified Correlation Structure Within Clusters

**CS missedpecified as AR-1, \( x = 0, \Phi(x) = 0.500 \)**

<table>
<thead>
<tr>
<th>( \gamma_2 )</th>
<th>( n )</th>
<th>Sim</th>
<th>SE</th>
<th>Ours</th>
<th>Naive</th>
<th>( t_{\text{sim}} )</th>
<th>95% Sim CI Low</th>
<th>95% Sim CI Hi</th>
<th>95% Coverage Ours</th>
<th>95% Coverage Naive</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>10 \times 5</td>
<td>0.499</td>
<td>0.043</td>
<td>0.043</td>
<td>0.070</td>
<td>-0.82</td>
<td>0.496</td>
<td>0.502</td>
<td>0.977</td>
<td>1.000</td>
</tr>
<tr>
<td>0.1</td>
<td>40 \times 5</td>
<td>0.501</td>
<td>0.021</td>
<td>0.021</td>
<td>0.035</td>
<td>0.93</td>
<td>0.499</td>
<td>0.502</td>
<td>0.962</td>
<td>1.000</td>
</tr>
<tr>
<td>0.5</td>
<td>10 \times 5</td>
<td>0.499</td>
<td>0.044</td>
<td>0.045</td>
<td>0.070</td>
<td>-0.67</td>
<td>0.496</td>
<td>0.502</td>
<td>0.957</td>
<td>0.997</td>
</tr>
<tr>
<td>0.5</td>
<td>40 \times 5</td>
<td>0.499</td>
<td>0.021</td>
<td>0.022</td>
<td>0.035</td>
<td>-0.87</td>
<td>0.498</td>
<td>0.501</td>
<td>0.956</td>
<td>0.999</td>
</tr>
<tr>
<td>0.9</td>
<td>10 \times 5</td>
<td>0.499</td>
<td>0.047</td>
<td>0.054</td>
<td>0.070</td>
<td>-0.80</td>
<td>0.496</td>
<td>0.502</td>
<td>0.976</td>
<td>0.993</td>
</tr>
<tr>
<td>0.9</td>
<td>40 \times 5</td>
<td>0.501</td>
<td>0.023</td>
<td>0.028</td>
<td>0.035</td>
<td>0.94</td>
<td>0.499</td>
<td>0.502</td>
<td>0.985</td>
<td>0.997</td>
</tr>
</tbody>
</table>

**CS missedpecified as AR-1, \( x = 1, \Phi(x) = 0.841 \)**

<table>
<thead>
<tr>
<th>( \gamma_2 )</th>
<th>( n )</th>
<th>Sim</th>
<th>SE</th>
<th>Ours</th>
<th>Naive</th>
<th>( t_{\text{sim}} )</th>
<th>95% Sim CI Low</th>
<th>95% Sim CI Hi</th>
<th>95% Coverage Ours</th>
<th>95% Coverage Naive</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>10 \times 5</td>
<td>0.837</td>
<td>0.030</td>
<td>0.030</td>
<td>0.052</td>
<td>-4.14</td>
<td>0.836</td>
<td>0.839</td>
<td>0.944</td>
<td>0.992</td>
</tr>
<tr>
<td>0.1</td>
<td>40 \times 5</td>
<td>0.841</td>
<td>0.015</td>
<td>0.015</td>
<td>0.026</td>
<td>-0.26</td>
<td>0.840</td>
<td>0.842</td>
<td>0.954</td>
<td>0.999</td>
</tr>
<tr>
<td>0.5</td>
<td>10 \times 5</td>
<td>0.840</td>
<td>0.032</td>
<td>0.031</td>
<td>0.052</td>
<td>-1.55</td>
<td>0.838</td>
<td>0.842</td>
<td>0.951</td>
<td>0.994</td>
</tr>
<tr>
<td>0.5</td>
<td>40 \times 5</td>
<td>0.841</td>
<td>0.015</td>
<td>0.016</td>
<td>0.026</td>
<td>-1.03</td>
<td>0.840</td>
<td>0.842</td>
<td>0.958</td>
<td>1.000</td>
</tr>
<tr>
<td>0.9</td>
<td>10 \times 5</td>
<td>0.835</td>
<td>0.032</td>
<td>0.036</td>
<td>0.052</td>
<td>-6.15</td>
<td>0.833</td>
<td>0.837</td>
<td>0.973</td>
<td>0.991</td>
</tr>
<tr>
<td>0.9</td>
<td>40 \times 5</td>
<td>0.839</td>
<td>0.017</td>
<td>0.018</td>
<td>0.026</td>
<td>-3.55</td>
<td>0.838</td>
<td>0.841</td>
<td>0.969</td>
<td>0.997</td>
</tr>
</tbody>
</table>

**AR-1 missedpecified as CS, \( x = 0, \Phi(x) = 0.500 \)**

<table>
<thead>
<tr>
<th>( \gamma_2 )</th>
<th>( n )</th>
<th>Sim</th>
<th>SE</th>
<th>Ours</th>
<th>Naive</th>
<th>( t_{\text{sim}} )</th>
<th>95% Sim CI Low</th>
<th>95% Sim CI Hi</th>
<th>95% Coverage Ours</th>
<th>95% Coverage Naive</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>10 \times 5</td>
<td>0.499</td>
<td>0.043</td>
<td>0.043</td>
<td>0.035</td>
<td>1.02</td>
<td>0.499</td>
<td>0.503</td>
<td>0.988</td>
<td>0.939</td>
</tr>
<tr>
<td>0.1</td>
<td>40 \times 5</td>
<td>0.499</td>
<td>0.019</td>
<td>0.021</td>
<td>0.035</td>
<td>-0.95</td>
<td>0.498</td>
<td>0.501</td>
<td>0.978</td>
<td>0.999</td>
</tr>
<tr>
<td>0.5</td>
<td>10 \times 5</td>
<td>0.500</td>
<td>0.042</td>
<td>0.043</td>
<td>0.035</td>
<td>-0.08</td>
<td>0.497</td>
<td>0.503</td>
<td>0.961</td>
<td>0.906</td>
</tr>
<tr>
<td>0.5</td>
<td>40 \times 5</td>
<td>0.501</td>
<td>0.021</td>
<td>0.021</td>
<td>0.035</td>
<td>0.75</td>
<td>0.499</td>
<td>0.502</td>
<td>0.961</td>
<td>0.998</td>
</tr>
<tr>
<td>0.9</td>
<td>10 \times 5</td>
<td>0.504</td>
<td>0.061</td>
<td>0.052</td>
<td>0.035</td>
<td>1.84</td>
<td>0.500</td>
<td>0.507</td>
<td>0.908</td>
<td>0.750</td>
</tr>
</tbody>
</table>

**AR-1 missedpecified as CS, \( x = 1, \Phi(x) = 0.841 \)**

<table>
<thead>
<tr>
<th>( \gamma_2 )</th>
<th>( n )</th>
<th>Sim</th>
<th>SE</th>
<th>Ours</th>
<th>Naive</th>
<th>( t_{\text{sim}} )</th>
<th>95% Sim CI Low</th>
<th>95% Sim CI Hi</th>
<th>95% Coverage Ours</th>
<th>95% Coverage Naive</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>10 \times 5</td>
<td>0.839</td>
<td>0.021</td>
<td>0.020</td>
<td>0.026</td>
<td>-7.09</td>
<td>0.832</td>
<td>0.836</td>
<td>0.911</td>
<td>0.878</td>
</tr>
<tr>
<td>0.1</td>
<td>40 \times 5</td>
<td>0.839</td>
<td>0.017</td>
<td>0.015</td>
<td>0.026</td>
<td>-5.02</td>
<td>0.838</td>
<td>0.840</td>
<td>0.923</td>
<td>1.000</td>
</tr>
<tr>
<td>0.5</td>
<td>10 \times 5</td>
<td>0.839</td>
<td>0.030</td>
<td>0.030</td>
<td>0.026</td>
<td>-2.23</td>
<td>0.837</td>
<td>0.841</td>
<td>0.940</td>
<td>0.910</td>
</tr>
<tr>
<td>0.5</td>
<td>40 \times 5</td>
<td>0.841</td>
<td>0.015</td>
<td>0.015</td>
<td>0.026</td>
<td>-0.34</td>
<td>0.840</td>
<td>0.842</td>
<td>0.970</td>
<td>0.999</td>
</tr>
<tr>
<td>0.9</td>
<td>10 \times 5</td>
<td>0.840</td>
<td>0.040</td>
<td>0.035</td>
<td>0.026</td>
<td>-0.95</td>
<td>0.838</td>
<td>0.843</td>
<td>0.926</td>
<td>0.796</td>
</tr>
<tr>
<td>0.9</td>
<td>40 \times 5</td>
<td>0.843</td>
<td>0.021</td>
<td>0.018</td>
<td>0.026</td>
<td>1.93</td>
<td>0.841</td>
<td>0.844</td>
<td>0.901</td>
<td>0.984</td>
</tr>
</tbody>
</table>

Simulations for clustered normally distributed data whose clusters have first-order autoregressive (AR-1) or compound symmetry (CS) correlation structures, but whose structure is misspecified. The correlation parameter for the correct structure is indicated by \( \gamma_2 \). Each simulated data set had 10 or 40 clusters, each cluster having 5 observations each.
the standardized vector (known to have $N_n(0, R_n)$ distribution for some $n \times n$ matrix $R_n$). We conducted simulations using the covariance structures described in Section 4.1.

Results appear in Table 9. Note that both types of residuals lead to unbiased estimates of $P(z_i \leq x)$, but that the standard error computed as $(\Phi(x)(1 - \Phi(x))/n)^{-1/2}$ matches the simulation standard deviation of the rotated residuals only. In general, the standardized (unrotated) residuals had a larger standard deviation. This simulation confirms that the failure to rotate the estimated residuals results in poorer estimation of the empirical CDF.

4.6 Discussion

The simulations we have conducted suggest that our proposed method enjoys good properties for a wide variety of linear models, including those that may not strictly fit within the framework of the theory we have presented in Section 3. Additionally, our method displays adequate sensitivity to non-normal errors, but does not display much sensitivity to mis-specified covariance structures. We speculate that the method may be modified slightly to produce increased sensitivity to mis-specified covariance, and recommend this as a subject for future research.

5 Applications

In this section we present two applications of our proposed methodology. The first demonstrates the utility of distribution diagnostics in a time series setting, in which the results of our diagnostic justify the simplification of an analysis. In the second application, we present a mean regression that produces scientifically couterintuitive results, and use

http://biostats.bepress.com/harvardbiostat/paper19
Table 9: Simulations for Clustered Data with Known Parameters

\[
x = 0, \, \Phi(x) = 0.500
\]

<table>
<thead>
<tr>
<th>Struct</th>
<th>(\gamma_1)</th>
<th>(n)</th>
<th>Rotated (z \sim N_n(0, I_n))</th>
<th>Standardized (z \sim N_n(0, R_n))</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.499 0.072</td>
<td>0.498 0.080</td>
<td>0.071</td>
</tr>
<tr>
<td>CS</td>
<td>0.1</td>
<td>40 x 5</td>
<td>0.498 0.035</td>
<td>0.498 0.040</td>
<td>0.035</td>
</tr>
<tr>
<td>CS</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.500 0.069</td>
<td>0.498 0.102</td>
<td>0.071</td>
</tr>
<tr>
<td>CS</td>
<td>0.5</td>
<td>40 x 5</td>
<td>0.499 0.036</td>
<td>0.500 0.054</td>
<td>0.035</td>
</tr>
<tr>
<td>CS</td>
<td>0.9</td>
<td>10 x 5</td>
<td>0.502 0.071</td>
<td>0.500 0.141</td>
<td>0.071</td>
</tr>
<tr>
<td>CS</td>
<td>0.9</td>
<td>40 x 5</td>
<td>0.499 0.037</td>
<td>0.497 0.068</td>
<td>0.035</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.497 0.074</td>
<td>0.496 0.072</td>
<td>0.071</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.1</td>
<td>40 x 5</td>
<td>0.500 0.037</td>
<td>0.500 0.039</td>
<td>0.035</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.499 0.072</td>
<td>0.500 0.094</td>
<td>0.071</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.5</td>
<td>40 x 5</td>
<td>0.498 0.034</td>
<td>0.500 0.048</td>
<td>0.035</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.9</td>
<td>10 x 5</td>
<td>0.501 0.074</td>
<td>0.500 0.129</td>
<td>0.071</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.9</td>
<td>40 x 5</td>
<td>0.498 0.036</td>
<td>0.497 0.067</td>
<td>0.035</td>
</tr>
</tbody>
</table>

\[
x = 1, \, \Phi(x) = 0.841
\]

<table>
<thead>
<tr>
<th>Struct</th>
<th>(\gamma_1)</th>
<th>(n)</th>
<th>Rotated (z \sim N_n(0, I_n))</th>
<th>Standardized (z \sim N_n(0, R_n))</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.840 0.052</td>
<td>0.839 0.057</td>
<td>0.052</td>
</tr>
<tr>
<td>CS</td>
<td>0.1</td>
<td>40 x 5</td>
<td>0.840 0.027</td>
<td>0.840 0.029</td>
<td>0.026</td>
</tr>
<tr>
<td>CS</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.843 0.051</td>
<td>0.842 0.076</td>
<td>0.052</td>
</tr>
<tr>
<td>CS</td>
<td>0.5</td>
<td>40 x 5</td>
<td>0.841 0.026</td>
<td>0.841 0.036</td>
<td>0.026</td>
</tr>
<tr>
<td>CS</td>
<td>0.9</td>
<td>10 x 5</td>
<td>0.841 0.052</td>
<td>0.836 0.141</td>
<td>0.052</td>
</tr>
<tr>
<td>CS</td>
<td>0.9</td>
<td>40 x 5</td>
<td>0.840 0.026</td>
<td>0.840 0.068</td>
<td>0.026</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.1</td>
<td>10 x 5</td>
<td>0.837 0.052</td>
<td>0.837 0.055</td>
<td>0.052</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.1</td>
<td>40 x 5</td>
<td>0.842 0.026</td>
<td>0.842 0.028</td>
<td>0.026</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.5</td>
<td>10 x 5</td>
<td>0.841 0.051</td>
<td>0.842 0.067</td>
<td>0.052</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.5</td>
<td>40 x 5</td>
<td>0.841 0.025</td>
<td>0.841 0.031</td>
<td>0.026</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.9</td>
<td>10 x 5</td>
<td>0.840 0.053</td>
<td>0.842 0.093</td>
<td>0.052</td>
</tr>
<tr>
<td>AR-1</td>
<td>0.9</td>
<td>40 x 5</td>
<td>0.842 0.026</td>
<td>0.839 0.046</td>
<td>0.026</td>
</tr>
</tbody>
</table>

Simulations for clustered normally distributed data whose clusters have first-order autoregressive (AR-1) or compound symmetry (CS) correlation structures. The correlation parameter is indicated by \(\gamma_2\). Each simulated data set had 10 or 40 clusters, each cluster having 5 observations each. In these cases, the regression and variance parameters are known. For the Rotated method, the residuals are transformed to \(N(0, I_n)\) by applying the Cholesky decomposition of the known variance structure, while for the Standardized method, the residuals are standardized componentwise only, resulting in residuals that are \(N(0, R_n)\) for some nondiagonal matrix \(R_n\). The estimated standard error (SE) is computed as \(\Phi(x)(1 - \Phi(x))/n^{-1/2}\). Although both methods correctly estimate \(F_n(0)\) and \(F_n(1)\), the simulation standard deviation resulting from the Standardized method does not match the estimated standard error.
the Q-Q plot to guide us to a more appropriate analysis.

5.1 Pollen Counts

Stark et al. (1997) and Brumback et al. (2000) described pollen counts associated with meteorologic data. Between 1991 and 1994, pollen was collected seven days a week during the pollen season in Kalamazoo, MI. Because the outcome consisted of counts, the authors used Poisson regression techniques to establish a model for forecasting pollen levels. The time series nature of the problem complicated the analysis and led to a computationally intensive procedure that needed to be fit iteratively each day. While these papers present a predictive model based on Poisson regression, we use the 1991 data to illustrate the behavior of our methods when they are applied to non-normal time series data.

We briefly present two analyses. Both employ the model described by Theorem 5 \((\bar{p} = 1\) and \(\bar{q} = 0\)), with a design matrix composed of an intercept and six covariates: \textit{rain} (1 if there were at least 3 hours of steady rain or brief but intense rain, 0 otherwise), \textit{day} (day in season), \(\ln(\text{day})\), \textit{wind} (wind speed in knots), \textit{temp trend} (a smooth curve fit to daily temperatures in °F) and \textit{temp resid} (residuals of \textit{temp trend}). The choice of covariates is motivated by Stark et al. (1997). In the first analysis we use the raw pollen counts as the outcome. In the second analysis, we use the square root of the pollen counts, since the square root is the variance-stabilizing transformation for the Poisson distribution. Parameter estimates appear in Table 10.

Figure 1 illustrates the quantile-quantile plots for each analysis, using the rotated residuals. It is clear from the plots that failing to transform the pollen counts results in a residual vector that deviates substantially from normality. On the other hand, the square root transformation may be adequate to transform the residuals to normality.
Table 10: Pollen Count Model Estimates

<table>
<thead>
<tr>
<th>Param</th>
<th>No Transformation</th>
<th></th>
<th>Square Root Transformation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est</td>
<td>SE</td>
<td>Est</td>
<td>SE</td>
</tr>
<tr>
<td>Intercept</td>
<td>-921.9</td>
<td>346.6</td>
<td>-53.66</td>
<td>16.8</td>
</tr>
<tr>
<td>rain</td>
<td>12.57</td>
<td>13.39</td>
<td>0.84</td>
<td>0.70</td>
</tr>
<tr>
<td>day</td>
<td>2.11</td>
<td>2.15</td>
<td>0.05</td>
<td>0.10</td>
</tr>
<tr>
<td>log(numday)</td>
<td>16.91</td>
<td>26.97</td>
<td>2.30</td>
<td>1.33</td>
</tr>
<tr>
<td>wind</td>
<td>1.92</td>
<td>1.41</td>
<td>0.18</td>
<td>0.07</td>
</tr>
<tr>
<td>temp trend</td>
<td>12.02</td>
<td>4.71</td>
<td>0.70</td>
<td>0.23</td>
</tr>
<tr>
<td>temp resid</td>
<td>0.83</td>
<td>0.96</td>
<td>0.08</td>
<td>0.05</td>
</tr>
<tr>
<td>Scale (γ₁)</td>
<td>2032</td>
<td>632</td>
<td>7.37</td>
<td>1.53</td>
</tr>
<tr>
<td>Autocorrelation (γ₂)</td>
<td>0.60</td>
<td>0.08</td>
<td>0.57</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Regression model for the 1991 pollen count data presented in Stark et al. (1997) and Brumback et al. (2000), with first-order autoregressive (AR-1) errors assumed. The text describes the covariates.

Thus, our residual diagnostic justifies the use of square-root transformed residuals for a reasonably simple time-series predictive model, which has the advantages of parameter interpretability and ease of computation over the Poisson transitional regression model.

5.2 Dissolved Oxygen in Boston Harbor

Another example of environmental time series data is unpublished data from the Massachusetts Water Resources Authority (MWRA). Among numerous other indicators of water quality, dissolved oxygen (DO) was measured at several stations in Boston Harbor between mid-1995 and 1999. Prior to 1998, only primary sewage treatment was used, and sewage was discharged near Nut Island in the harbor. In 1998, the Nut Island facility was closed and sewage was treated at Deer Island, a secondary treatment facility. Interest focuses on whether the closure of the Nut Island facility affected water quality in the harbor, as measured by DO and TDN. Analysis is complicated by the spatio-temporal associations between observations. Additionally, it is unclear whether there are transformations that can convert the outcomes to variables that are plausibly normal.
Figure 1: Normal Quantile-Quantile Plots for Pollen Data

No Transformation

Square Root Transformation

Quantiles of Standard Normal
A plausible model for analyzing the data is

\[ y_{ht} = x_{ht}^T \beta + a_t + \varepsilon_{ht}, \]  

(13)

where \( y_{ht} \) is the outcome (DO or \( \log_{10} \) DO) measured at the fixed location \( h \) at time \( t \) (discretized into weeks), \( x_{ht} \) is a design vector to be described shortly, \( \beta \) is a fixed parameter, \( a_t \sim N(0, \sigma_a^2) \) are iid random intercepts common for each time point, and \( \varepsilon_{ht} \sim N(0, \varepsilon^2) \) are iid error terms. Times are discretized to weeks to provide sufficient data for estimating a random effect. The design matrix includes an intercept term, water temperature and salinity measured at location \( h \) at time \( t \), and indicators for fixed location effects. In addition, it contains periodic terms of the form \( \sin(2\pi t'/D) \) and \( \cos(2\pi t'/365.24) \), where \( t' \) is continuous time measured in days. The periodic terms address the clear seasonal pattern observed in time series plots of the data, and serve to model time-dependence in terms of a mean predictor. The random intercepts \( \{a_t\} \) model residual correlation among outcomes collected at the same time but at different locations. We remark that it is possible to choose a more complicated spatial model for the correlation of observations observed at different locations, but since there were only a few locations dispersed in a relatively small area, we believe (13) is adequate as a first-order approximation. It is also possible to replace \( a_t \) by a more complicated random effect \( Z_{ht}b_t \), where \( Z_{ht} \) consists of a subset of the terms in \( x_{ht} \). We investigated these alternatives and found that they do not materially affect the diagnostic conclusions. The design vector \( x_{ht} \) also includes a term representing the effect of closure (0 before the closure of Nut Island and 1 afterwards) and interactions between the location and the closure effects. The effects of interest are the components of \( \beta \) corresponding to the closure effects.
Figure 2 displays the diagnostic plot for $\log_{10}$ DO. Although the residuals appear symmetric, the tails are clearly very heavy compared with a normal distribution. It would be difficult to assess how heavy-tailed the true distribution could be without the pointwise bands provided by our theory. The plot for untransformed DO (not shown) reveals a more skewed distribution, deviating even more clearly from normality.

Simple alternatives are available when normal assumptions fail. An alternative semiparametric approach to modeling the mean is to use generalized estimating equations [GEEs, Diggle et al. (2002)]. In the GEE approach, we employ the following generalization of (13):

$$y_{ht} = x_{ht}^T \beta + \varepsilon_{ht}, \quad (14)$$

where $\varepsilon_{ht} \sim G$, $G$ is an unspecified distribution having zero mean, and $\varepsilon_{hs}$ and $\varepsilon_{ht}$ are independent when $s \neq t$. Note that the class of models described by (13) is properly contained within the class of models described by (14), since (14) does not require errors within the same time period to be independent. If the goal is to estimate central tendency, yet another alternative is median regression. Such an approach uses expression (14), but relaxes assumptions involving $\varepsilon_{ht}$; it requires only that $\varepsilon_{ht} \sim G_{ht}$, where the median of $\varepsilon_{ht}$ is zero and $\varepsilon_{hs}$ and $\varepsilon_{ht}$ are independent when $s \neq t$. With these restrictions, $G_{ht}$ is allowed to differ for each observation. For median regression, there are numerous approaches to estimation [e.g., Hendricks and Koenker (1992) and Jung (1996)], but we choose the approach described in Chen et al. (2002). The latter approach employs a resampling technique for computing standard errors that is valid for clustered data but does not require subsampling the clusters. It is analogous to a GEE on the median with a working independence assumption. We implemented the approach described in Chen
et al. (2002) using the R software package with a simple application of the function `qr` (in the module `quantreg`).

Table 11 presents coefficient estimates for log-transformed values analyzed by three final models: linear mixed effects (LME) as described by (13); the corresponding GEE with a working independence assumption; and the corresponding median regression. From each model we removed high-order periodic terms that did not appear to be significant. For DO, the fixed effects are mostly negative. Because higher values of DO indicate better water quality, the parameter estimates for the closure effect might cause concern. In particular, the LME model shows a significantly negative effect at Station 135 $(−0.021 − 0.008 = −0.029$, with a Wald-based $P$-value of 0.004). The GEE and median models produce similar results. However, Wald tests for the collective closure effect differ between the mean and median models. Wald $P$-values for the collective closure effect are 0.047 for the LME model, 0.0009 for the GEE model, and 0.244 for the median regression model. One possible reason for the qualitative difference between these results is that mean regression is more efficient than quantile regression when the residuals are normally distributed. However, Figure 2 clearly depicts heavy-tailed residuals, so that mean regression may be less efficient. In fact, using the determinant as a measure of size, median regression appears more efficient than the GEE for estimating the closure effects: the covariance of the closure terms in the LME model has a determinant of $e^{-67.57}$, and the corresponding determinants for the GEE and median models are $e^{-64.87}$ and $e^{-66.51}$, respectively. Given that the GEE is slightly less powerful than the median regression model, the difference in conclusion is striking.

In summary, the mean regression (LME and GEE) results are somewhat inconsistent
with an overall improvement in water quality after the closure of the Nut Island facility. On the other hand, the median regression results are consistent with a slight overall improvement. Our distribution diagnostics suggest that the true error distribution is symmetric, but has heavy tails relative to the normal distribution. Consequently, the median regression model may be more appropriate in this context.

6 Discussion and Conclusions

Our proposed methodology makes use of rotated residuals, formed as the product of the Cholesky decomposition of the estimated marginal variance with the estimated residual vector, to construct an empirical CDF and pointwise standard errors. It thereby accommodates natural graphical display techniques.
Table 11: Estimates of Nut Island Closure Effects for $\log_{10}$ Dissolved Oxygen

<table>
<thead>
<tr>
<th></th>
<th>Mean Regression (LME)</th>
<th></th>
<th>Mean Regression (GEE)</th>
<th></th>
<th>Median Regression</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est.</td>
<td>SE</td>
<td>$P$-value</td>
<td>Est.</td>
<td>SE</td>
<td>$P$-value</td>
</tr>
<tr>
<td>(Intercept)</td>
<td>0.944</td>
<td>0.006</td>
<td>$&lt;0.0001$</td>
<td>0.939</td>
<td>0.007</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td>Temp. Resid.</td>
<td>-0.003</td>
<td>0.001</td>
<td>$&lt;0.0001$</td>
<td>-0.004</td>
<td>0.001</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td>Sal. Resid.</td>
<td>-0.005</td>
<td>0.001</td>
<td>$&lt;0.0001$</td>
<td>-0.001</td>
<td>0.002</td>
<td>0.7519</td>
</tr>
<tr>
<td>Sin</td>
<td>0.073</td>
<td>0.005</td>
<td>$&lt;0.0001$</td>
<td>0.073</td>
<td>0.003</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td>Cos</td>
<td>0.032</td>
<td>0.005</td>
<td>$&lt;0.0001$</td>
<td>0.033</td>
<td>0.003</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td>Station 44</td>
<td>0.011</td>
<td>0.007</td>
<td>0.1016</td>
<td>0.018</td>
<td>0.011</td>
<td>0.1057</td>
</tr>
<tr>
<td>Station 65</td>
<td>0.014</td>
<td>0.006</td>
<td>0.0134</td>
<td>0.022</td>
<td>0.013</td>
<td>0.0864</td>
</tr>
<tr>
<td>Station 106</td>
<td>0.015</td>
<td>0.005</td>
<td>0.0039</td>
<td>0.016</td>
<td>0.009</td>
<td>0.0829</td>
</tr>
<tr>
<td>Station 135</td>
<td>0.030</td>
<td>0.006</td>
<td>$&lt;0.0001$</td>
<td>0.036</td>
<td>0.011</td>
<td>0.0003</td>
</tr>
<tr>
<td>Station 136</td>
<td>0.024</td>
<td>0.006</td>
<td>$&lt;0.0001$</td>
<td>0.034</td>
<td>0.011</td>
<td>0.0016</td>
</tr>
<tr>
<td>Station 141</td>
<td>0.022</td>
<td>0.005</td>
<td>$&lt;0.0001$</td>
<td>0.023</td>
<td>0.008</td>
<td>0.0054</td>
</tr>
<tr>
<td>Nut Is. Closure</td>
<td>-0.008</td>
<td>0.010</td>
<td>0.4145</td>
<td>-0.001</td>
<td>0.009</td>
<td>0.8885</td>
</tr>
<tr>
<td>St. 44 x Closure</td>
<td>0.006</td>
<td>0.011</td>
<td>0.5829</td>
<td>-0.001</td>
<td>0.017</td>
<td>0.9552</td>
</tr>
<tr>
<td>St. 65 x Closure</td>
<td>-0.004</td>
<td>0.009</td>
<td>0.6530</td>
<td>-0.012</td>
<td>0.015</td>
<td>0.4371</td>
</tr>
<tr>
<td>St. 106 x Closure</td>
<td>-0.009</td>
<td>0.009</td>
<td>0.2857</td>
<td>-0.010</td>
<td>0.012</td>
<td>0.4245</td>
</tr>
<tr>
<td>St. 135 x Closure</td>
<td>-0.021</td>
<td>0.009</td>
<td>0.0241</td>
<td>-0.031</td>
<td>0.013</td>
<td>0.0218</td>
</tr>
<tr>
<td>St. 136 x Closure</td>
<td>-0.017</td>
<td>0.010</td>
<td>0.0790</td>
<td>-0.027</td>
<td>0.014</td>
<td>0.0519</td>
</tr>
<tr>
<td>St. 141 x Closure</td>
<td>-0.010</td>
<td>0.009</td>
<td>0.2477</td>
<td>-0.011</td>
<td>0.011</td>
<td>0.3440</td>
</tr>
<tr>
<td><strong>Closure Effect</strong></td>
<td></td>
<td></td>
<td><strong>(7 d.f.)</strong></td>
<td><strong>P = 0.0467</strong></td>
<td><strong>P = 0.0009</strong></td>
<td><strong>P = 0.2440</strong></td>
</tr>
</tbody>
</table>

Fixed effects estimates for Boston Harbor dissolved oxygen (mg/L) data collected by MWRA from mid-1995 through 1999. *Nut Island Closure* refers to the 1998 closure of the primary Nut Island treatment facility. Reference station is Station 77. Trigonometric terms have a period of one year. Linear mixed effects (LME) model includes a random intercept term grouping together data collected within the same 7-day period. Generalized estimating equations (GEE) and median regression models are based on the same 7-day clustering scheme. Higher order periodic terms were not significant for LME and median models. Collective closure effect is based on a Wald $\chi^2$ statistic.
While the empirical CDF and corresponding Q-Q plot are adequate for independent and identically distributed errors, our method works for a large class of correlated errors. Our method is quite different from Lange and Ryan (1989), whose diagnostic targets the random effects in a longitudinal mixed model. It is also more general than such approaches as Fraccaro et al. (2000) and Louis (1988), in that we present asymptotic theory. To our knowledge, this has has not been achieved previously. The simulations we have conducted suggest that our proposed method enjoys good properties in a variety of circumstances. It appears to work well for models having independent units of sampling (clustered data) and for many models for which all observations are correlated.

One limitation is that the conditions required for the asymptotic theory to hold can be challenging to demonstrate. We have proven it for clustered data (with some weak regularity conditions) and for ARMA errors. We have not shown that it holds in crossed random effects models or conditionally autoregressive models. However, simulations demonstrate that it may perform well even in these circumstances. The general steps for proving the conditions for asymptotic normality involve, essentially, demonstrating exponential decay of correlation over widely separated observations. The proof of Theorem 5 (in the Appendix) is a template for the type of proof required. Unfortunately, it is difficult to state a general theorem, as the effect of the Cholesky matrix in the quantity $m_i$ must be considered, and in the proof of Theorem 5 we were able to exploit convenient algebraic relationships.

Another limitation is one that is present for any omnibus testing procedure: one type of lack-of-fit may mask another, so the procedure may not be powerful in certain situations. This problem has been discussed by authors such as Verbeke and Molenberghs (2000),
although we believe such concerns are less relevant in a time-series or spatial setting. Transformed residuals are not appropriate in every case. For example, in cases where the normality of the random effects are an issue, the approach of Lange and Ryan (1989) is clearly superior. When the diagnostic is indicative of a problem, it may be unclear what the cause is; however, the Q-Q plot is an old graphical technique, and much of the intuition in that setting applies to our method. For example, the dissolved oxygen example suggests that a heavy-tailed distribution is easily recognized.

It is possible to generalize Theorem 1 in many directions. The efficiency of \( \hat{\theta}_n \) is not critical, although the form of the variance \( \tau^2 \) (clearly obtained within the proof of Theorem 1) becomes more complex when \( \hat{\theta}_n \) is less than efficient. To examine a different distribution \( F \), one need replace \( \Phi \) with \( F \); however, the algebra of normal distributions will no longer apply, and interpretation of results may be less clear.

Our two examples suggest possible remedial actions when normal assumptions fail: transformations may correct a skewed distribution, while median regression may serve when the distribution is symmetric but heavy tailed. Complicated hierarchical models may be an alternative. For example, a t-error model may be used in the dissolved oxygen example.

Several issues warrant further study. For example, our method addresses only pointwise confidence intervals and does not provide a rigorous method for constructing confidence bands over a region of the empirical CDF. To this end, it would be desirable to establish stochastic convergence of the CDF \( F_n(\hat{\beta}, \hat{\gamma}; x) \) for all \( x \) in a subinterval of \( \mathbb{R} \). Such a result may be obtained using the elegant theory of empirical processes. For example, the results in Chapter 19 of van der Vaart (1998) may be used to prove the stochastic
version of Theorem 1, but would require a stronger condition than \((6)\). In particular, the convergence of \((6)\) should be uniform in \(x\). Even with appropriate theoretical results, estimation of critical bands around the empirical CDF may be difficult to achieve, and may require simulation techniques. The problem of constructing the confidence bands is particularly difficult for cases where \(V(\gamma)\) is not block diagonal.

Another question is whether our proposed estimator may be modified slightly to achieve greater sensitivity in detecting particular types of deviation from null model assumptions. In additional simulations, we investigated the method’s behavior when the covariance structure of clusters was mis-specified. In general, the method does not display much sensitivity to mis-specified covariance structures. We speculate that generalizations of the projection matrix \(P_i\) in the construction of \(m_i(\beta, \gamma)\) and \(s_i(\gamma)\) may be used to enhance sensitivity. For example, we suspect that the weighted normal plot introduced by Lange and Ryan (1989), constructed to be sensitive to departures from the assumed random effects distribution, can be derived as a special case of our theory. It would be of interest to explore this connection, as well as other related constructions that might be sensitive to specific types of departures.

Finally, extensions of our theory to the generalized linear mixed model setting would be also be of interest. For example, a graphical method for detection overdispersion in Poisson models would be of great interest. Efforts towards this goal could be directed in developing generalizations to rotated “working residuals” (McCullagh and Nelder, 1989).
References


Appendix

In the following appendix we provide formal proofs of the theorems given in the main text. We also present some additional theorems to address special cases. Notation and definitions are the same as in the main text, with one minor exception: since $x$ is assumed fixed, we drop it from the notation. We also introduce some additional conventions as follows. We denote the correlation between $z_i(\beta, \gamma)$ and $z_i(\beta_0, \gamma_0)$ as $r_i(\gamma)$, easily shown to be

$$r_i(\gamma) = \text{Corr} [z_i(\beta, \gamma), z_i(\beta_0, \gamma_0)] = \frac{P_i L(\gamma)^T V(\gamma_0) L(\gamma_0) P_i^T}{s_i(\gamma)}. \tag{15}$$

In addition, we define the difference

$$a_i(\beta, \gamma; \beta', \gamma', x) = I (x - z_i(\beta, \gamma)) - I (x - z_i(\beta', \gamma')). \tag{16}$$

Note that $E[a_i(\beta, \gamma; \beta', \gamma', x)] = \Phi (x_i(\beta, \gamma; x)) - \Phi (x_i(\beta', \gamma'; x))$, that $a_i(\beta, \gamma; \beta', \gamma', x) = 1$ when $z_i(\beta, \gamma) < z_i(\beta', \gamma')$, that $a_i(\beta, \gamma; \beta', \gamma', x) = -1$ when $z_i(\beta', \gamma') < z_i(\beta, \gamma)$, and that $a_i(\beta, \gamma; \beta', \gamma', x) = 0$ otherwise.

We require additional notation as follows. We denote the absolute value and vector norm identically as $| \cdot |$, and reserve $\| \cdot \|$ for the $\mathcal{L}^2$ matrix norm; that is, if $A$ is a $k \times d$ matrix, then $\|A\| = \min \{a \in \mathbb{R} : |Ax| \leq a|x| \ \forall x \in \mathbb{R}^d\}$. We denote the derivative of a matrix $U$ with respect to a scalar $\zeta$ as $d_\zeta U$. We often require boundedness of $\|d_\zeta U\|$ for all elements of some multidimensional parameter vector $\zeta = (\zeta_1, ..., \zeta_k)^T$; in such a case, we state this condition informally as a boundedness condition on $\|d_\zeta U\|$.

In order to establish desired results, it is necessary to consider the combined limiting behavior both of the parameter estimator and of the residuals. As a consequence, it is necessary to distinguish a fixed $\theta$ close to $\theta_0$ apart from an estimator $\widetilde{\theta}_n$ which may yield
\(\theta\) as a value. Therefore, quantities that converge in \(O(|\theta - \theta_0|)\) or \(O(|\theta - \theta_0|^2)\) must be understood to converge uniformly with respect to \(n\), so that convergence is preserved when we replace \(\theta\) by \(\hat{\theta}_n\). In general, when we write \(f_n(\theta) = O(g(\theta))\), we mean that there exists an \(M > 0\), independent of \(n\), such that \(|f_n(\theta)| \leq M g(\theta)\) for all \(n\). This notation extends standard \(O\)-notation [e.g., described in Serfling (1980), Section 1.1.2] to address our uniform convergence requirements.

Theorem 1, stated in Section 3, establishes the desired result. However, it relies on an asymptotic differentiability property that is difficult to establish for correlated residuals \(z_t\). Theorem 3 provides a sufficient condition (18), but this condition is itself difficult to verify. Theorem 4 establishes easily verifiable conditions for block-diagonal variance matrices, Theorem 5 establishes a result for a time series model, and Theorem 6 establishes a result for variances based on Kronecker products. Finally, Theorem 7 justifies (7) and suggests conditions required for a justification of (12).

**Proof of Theorem 1.** Since

\[
\mu^*(\beta_0, \gamma_0) = \mu_n(\beta_0, \gamma_0) = \frac{1}{n} \sum_{i=1}^{n} \Phi(x) = \Phi(x),
\]

we have

\[
\mu_n(\hat{\beta}_n, \hat{\gamma}_n) - \Phi(x) = \mu_n(\hat{\beta}_n, \hat{\gamma}_n) - \mu^*(\hat{\beta}_n, \hat{\gamma}_n) + \mu^*(\hat{\beta}_n, \hat{\gamma}_n) - \mu^*(\beta_0, \gamma_0) + \mu^*(\beta_0, \gamma_0) - \Phi(x)
\]

\[
= \varepsilon_n^T(\hat{\theta}_n - \theta_0) + O(|\hat{\theta}_n - \theta_0|^2) + \mu^*(\hat{\beta}_n, \hat{\gamma}_n) - \mu^*(\beta_0, \gamma_0)
\]

\[
= \delta(\theta_0)^T(\hat{\theta}_n - \theta_0) + \varepsilon_n^T(\hat{\theta}_n - \theta_0) + O(|\hat{\theta}_n - \theta_0|^2),
\]

where the last equality follows by Taylor expansion. By Slutsky’s Theorem, \(n^{\frac{1}{2}} \varepsilon_n^T(\hat{\theta}_n - \theta_0)\) and \(n^{\frac{1}{2}}|\hat{\theta}_n - \theta_0|^2\) both converge in probability to zero. Consequently,

\[
n^{\frac{1}{2}}|\mu_n(\hat{\beta}_n, \hat{\gamma}_n) - \Phi(x)| = n^{\frac{1}{2}}\delta(\theta_0)^T(\hat{\theta}_n - \theta_0) + n^{\frac{1}{2}}\varepsilon_n^T(\hat{\theta}_n - \theta_0) + O(n^{\frac{1}{2}}|\hat{\theta}_n - \theta_0|^2)
\]
\[
\begin{align*}
&= n^{\frac{1}{2}} \delta(\theta_0)^T(\tilde{\theta}_n - \theta_0) + o_p(1).
\end{align*}
\]

Now,
\[
\begin{align*}
n^{\frac{1}{2}}[F_n(\tilde{\beta}_n, \tilde{\gamma}_n) - \Phi(x)] &= n^{\frac{1}{2}}[F_n(\tilde{\beta}_n, \tilde{\gamma}_n) - \mu_n(\tilde{\beta}_n, \tilde{\gamma}_n) - F_n(\beta_0, \gamma_0) + \Phi(x) + \\
&\quad \mu_n(\tilde{\beta}_n, \tilde{\gamma}_n) + F_n(\beta_0, \gamma_0) - \Phi(x) - \Phi(x)] \\
&= o_p(1) + n^{\frac{1}{2}}[\mu_n(\tilde{\beta}_n, \tilde{\gamma}_n) - \Phi(x)] + n^{\frac{1}{2}}[F_n(\beta_0, \gamma_0) - \Phi(x)] \\
&= n^{\frac{1}{2}}[F_n(\beta_0, \gamma_0) - \Phi(x)] + n^{\frac{1}{2}}\delta(\theta_0)^T(\tilde{\theta}_n - \theta_0) + o_p(1).
\end{align*}
\]

This result shows that \( n^{\frac{1}{2}}[F_n(\tilde{\beta}_n, \tilde{\gamma}_n) - \Phi(x)] \) converges in law to a normal distribution, but is not directly useful for determining the form of \( \tau^2 \) in (9). Instead, we note that
\[
\begin{align*}
n^{\frac{1}{2}}[F_n(\beta_0, \gamma_0) - \Phi(x)] &= n^{\frac{1}{2}}[F_n(\tilde{\beta}_n, \tilde{\gamma}_n) - \Phi(x)] - n^{\frac{1}{2}}\delta(\theta_0)^T(\tilde{\theta}_n - \theta_0) + o_p(1),
\end{align*}
\]
from which we calculate the following relationship between the asymptotic dispersion parameters:

\[
\Phi(x)(1 - \Phi(x)) = \tau^2 + \delta(\theta_0)^T W_{11} \delta(\theta_0) - 2 Cov[D, \delta],
\]

where \( D \sim N_1(0, \tau^2) \) is the distribution limit of \( n^{\frac{1}{2}}[F_n(\tilde{\beta}_n, \tilde{\gamma}_n) - \Phi(x)] \) and \( \delta \sim N_{p+q}(0, W_{11}) \) is the distribution limit of \( n^{\frac{1}{2}}(\tilde{\theta}_n - \theta_0) \). The fact that \( \tilde{\theta}_n \) is asymptotically efficient and that \( D \) has constant expectation leads to the conclusion that \( Cov[D, \delta] = 0 \). Heuristically, this result, proven formally in Pierce (1982), is the asymptotic version of the fact that a random variable with constant expectation is uncorrelated with a MVUE. In particular, we note that the result
\[
\begin{align*}
n^{\frac{1}{2}}[F_n(\tilde{\beta}_n, \tilde{\gamma}_n) - \Phi(x)] &= n^{\frac{1}{2}}[F_n(\beta_0, \gamma_0) - \Phi(x)] + n^{\frac{1}{2}}\delta(\theta_0)^T(\tilde{\theta}_n - \theta_0) + o_p(1)
\end{align*}
\]
is condition (1.2) of Pierce (1982). Together with condition (5), which is condition (1.1) of Pierce (1982), and the asymptotic efficiency of \( \tilde{\theta}_n \), our result follows directly from Pierce (1982). \( \square \)
Proof of Theorem 2. Let

\[\zeta_n(\theta) = \tilde{\mu}_n(\theta, \theta_0) - \mu^*(\theta) = \varepsilon_n^T(\theta - \theta_0) + O(|\theta - \theta_0|^2),\]

(17)

where the second equality is true by condition (7). Then \(\zeta_n\) is continuously twice-differentiable and, by Taylor’s Theorem,

\[\zeta_n(\theta) = \zeta_n(\theta_0) + \frac{\partial \zeta_n}{\partial \theta} \bigg|_{\theta=\theta_0} (\theta - \theta_0) + O(|\theta - \theta_0|^2).\]

Since \(\tilde{\mu}_n(\theta_0, \theta_0) = \mu^*(\theta_0)\), this expression, combined with (17), shows that

\[\frac{\partial \zeta_n}{\partial \theta} \bigg|_{\theta=\theta_0} = \varepsilon_n,\]

and therefore that

\[\frac{\partial \zeta_n}{\partial \theta} = \varepsilon_n + O(|\theta - \theta_0|).\]

In other words, \(\delta_n(\theta, \theta_0) - \delta(\theta_0) = \varepsilon_n + O(|\theta - \theta_0|)\), where convergence of the latter term is uniform in \(n\). Condition (12) implies that \(\delta_n(\theta, \theta') - \delta_n(\theta, \theta_0) = O(|\theta' - \theta_0|)\), uniformly in \(n\) for \(\theta\) sufficiently close to \(\theta_0\). Therefore,

\[
\delta_n(\tilde{\theta}_n, \tilde{\theta}_n) - \delta(\theta_0) = \delta_n(\tilde{\theta}_n, \tilde{\theta}_n) - \delta_n(\tilde{\theta}_n, \theta_0) + \delta_n(\tilde{\theta}_n, \theta_0) - \delta(\theta_0)
\]

\[= O(\|\tilde{\theta}_n - \theta_0\|) + \varepsilon_n + O(\|\tilde{\theta}_n - \theta_0\|),\]

\[= \varepsilon_n + O(\|\tilde{\theta}_n - \theta_0\|),\]

\[= o_p(1).\]

The last equality follows from the consistency of \(\tilde{\theta}_n\). \(\square\)

Randles (1982) provided general conditions from which the condition (6) follows naturally, but unfortunately they do not apply in the present case. Randles’ result would require our residuals \(z_i(\beta, \gamma)\) to be independent, but this is true only when \(\theta = \theta_0\). Therefore, we develop alternative arguments to establish (6). Theorem 3 is a general statement,
but condition (18), which applies to the indicator residuals $a_i$ (not to $z_i$), is also difficult

to verify. This condition turns out to hold for a large class of interesting models, and we

present some special cases.

**Theorem 3 (Asymptotic Differentiability).** Let $\hat{\theta}_n$ be a consistent estimator of $\theta_0$.

If

$$n^{-1} \sum_{i \neq 1}^{n} \left| Cov \left[ a_i(\beta, \gamma; \beta_0, \gamma_0, x), a_i(\beta, \gamma; \beta_0, \gamma_0, x) \right] \right| = O \left( |\theta - \theta_0| \right), \quad (18)$$

where the boundedness of $O \left( |\theta - \theta_0| \right)$ is understood to be uniform in $n$, then

$$R_n = n^{\frac{1}{2}} \left| F_n(\hat{\beta}_n, \hat{\gamma}_n; x) - F_n(\beta_0, \gamma_0; x) + \Phi(x) \right| \overset{P}{\rightarrow} 0.$$ 

**Proof.** Let

$$R_n^*(\theta) = n^{\frac{1}{2}} \left( F_n(\beta, \gamma) - F_n(\beta_0, \gamma_0) + \Phi(x) \right),$$

so that $R_n = |R_n^*(\hat{\theta}_n)|$. For the moment, fix $\theta = (\beta, \gamma) \in \Theta$. Note that

$$F_n(\beta, \gamma) - F_n(\beta_0, \gamma_0) = n^{-1} \sum_{i=1}^{n} a_i(\beta, \gamma; \beta_0, \gamma_0)$$

and

$$E \left[ F_n(\beta, \gamma) - F_n(\beta_0, \gamma_0) \right] = n^{-1} \sum_{i=1}^{n} E \left[ a_i(\beta, \gamma; \beta_0, \gamma_0) \right] = \mu_n(\beta, \gamma) - \Phi(x).$$

Therefore $E[|R_n^*(\theta)|] = 0$. Also, by (18),

$$Var \left[ R_n^*(\theta) \right] = \sum_{i=1}^{n} Var \left[ \sum_{i'=1}^{n} a_i(\beta, \gamma; \beta_0, \gamma_0) \right] \leq n^{-1} \sum_{i=1}^{n} \sum_{i'=1}^{n} \left| Cov \left[ a_i(\beta, \gamma; \beta_0, \gamma_0), a_i(\beta, \gamma; \beta_0, \gamma_0) \right] \right|.$$

Thus $Var \left[ R_n^*(\theta) \right] = O \left( |\theta - \theta_0| \right)$. Together with the Chebychev inequality, this implies

that there exists an $M > 0$ such that

$$P \left( R_n^*(\theta) > \varepsilon \right) \leq \frac{Var \left[ R_n^*(\theta) \right]}{\varepsilon^2} \leq \frac{M}{\varepsilon^2} \left| \theta - \theta_0 \right|.$$
Now let $\pi_n$ be the probability measure for the random variable $\hat{\theta}_n$. For $\delta > 0$, it is possible to select $n_0$ large enough so that $\pi_n \left\{ |\theta - \theta_0| > \varepsilon^2 \delta / M \right\} = P \left( |\hat{\theta}_n - \theta_0| > \varepsilon^2 \delta / M \right) \leq \delta$ whenever $n \geq n_0$. For such $n,$

\[
P \left( R_n > \varepsilon, |\hat{\theta}_n - \theta_0| \leq \frac{\varepsilon^2 \delta}{M} \right) = \int_{\Theta} 1 \left\{ |\theta - \theta_0| \leq \frac{\varepsilon^2 \delta}{M} \right\} P(R_n^*(\theta) > \varepsilon |\theta) d\pi_n(\theta)
\leq \int_{\Theta} 1 \left\{ |\theta - \theta_0| \leq \frac{\varepsilon^2 \delta}{M} \right\} \frac{M}{\varepsilon^2} |\theta - \theta_0| d\pi_n(\theta)
\leq \frac{M \varepsilon^2 \delta}{\varepsilon^2} \int_{\Theta} 1 \left\{ |\theta - \theta_0| \leq \frac{\varepsilon^2 \delta}{M} \right\} d\pi_n(\theta)
\leq \delta
\]

and

\[
P \left( R_n > \varepsilon, |\hat{\theta}_n - \theta_0| > \frac{\varepsilon^2 \delta}{M} \right) \leq P \left( |\hat{\theta}_n - \theta_0| > \frac{\varepsilon^2 \delta}{M} \right) \leq \delta.
\]

Therefore, $P(R_n > \varepsilon)$ can be made arbitrarily small by choosing $n$ large enough. \(\Box\)

In practice, condition (18) may be difficult to verify directly. Theorem 4 addresses this condition for a large class of block-diagonal (clustered) designs, while Theorems 5 and 6 address some specific models whose marginal covariance is not block-diagonal. The proof of Theorem 4 depends upon the following relationship:

\[
\text{Var} [a_i(\beta, \gamma; \beta_0, \gamma_0, x)] = O(|\theta - \theta_0|),
\]

which bounds the variance and covariance of the $a_i$ quantities defined in (16).

The proof of (19) is complicated by the fact that the variance of $a_i(\beta, \gamma; \beta_0, \gamma_0, x)$ is not differentiable at $\theta_0$ because its behavior depends upon whether or not $\gamma = \gamma_0$; consequently, standard Taylor's Theorem arguments do not apply. We outline the heuristics of an alternative argument as follows. The variance and covariance of the $a_i$ quantities can be decomposed into probabilities of events involving both $z_i(\beta, \gamma)$ and $z_i(\beta_0, \gamma_0)$. For
example, \( a_i(\beta, \gamma; \beta_0, \gamma_0, x) = -1 \) precisely when \( z_i(\beta_0, \gamma_0) \leq x < z_i(\beta, \gamma) \). As \( \theta \to \theta_0 \), the correlation \( r_i(\gamma) \) of \( z_i(\beta, \gamma) \) and \( z_i(\beta_0, \gamma_0) \) approaches 1. This leads to singularity in the joint distribution of \( z_i(\beta, \gamma) \) and \( z_i(\beta_0, \gamma_0) \), and differentiation of \( \text{Var}[a_i] \) becomes impossible. However, though the variance and covariance of the \( a_i \) quantities lead to expressions that have no derivative at \( \theta_0 \), the derivatives near \( \theta_0 \) are bounded, and the maximum of the derivatives as \( \theta \to \theta_0 \) from any direction suffices for the bound implied in (19). We state the result, condition (19), as Lemma 1.

**Lemma 1.** Let \( a_i(\beta, \gamma; \beta', \gamma') \) be defined as in (16). If there is a neighborhood of \( \gamma_0 \) over which \( \|V(\gamma)\| \) is bounded and bounded away from zero, and over which \( \|d, L\| \) is bounded, then there is a neighborhood of \( \theta_0 \) and an \( M_n > 0 \) such that \( \text{Var}[a_i(\beta, \gamma, \beta_0, \gamma_0)] \leq M_n |\theta - \theta_0| \). If the gradients of \( m_i(\beta, \gamma) \), \( s_i(\gamma) \), and \( \cos^{-1} r_i(\gamma) \) are uniformly bounded in \( n \), then there is a uniform bound \( M = M_n > 0 \) for all \( n \).

To prove (18) for non-clustered designs a similar condition is required: \( \text{Cov}[a_i, a_j] \) for \( i \neq j \) must be similarly bounded and must diminish at a sufficient rate. Lemma 2 is useful in this context.

**Lemma 2.** Let \( a_i(\beta, \gamma; \beta', \gamma') \) be defined as in (16), and let \( \sigma_{ij} \) be defined as follows:

\[
\sigma_{ij}(\beta, \gamma) = \begin{bmatrix}
\sigma_{ij0} \\
\sigma_{ij1} \\
\sigma_{ij2}
\end{bmatrix} = \begin{bmatrix}
\text{Cov}[z_i(\beta, \gamma), z_j(\beta, \gamma)] \\
\text{Cov}[z_i(\beta, \gamma), z_j(\beta_0, \gamma_0)] \\
\text{Cov}[z_i(\beta_0, \gamma_0), z_j(\beta, \gamma)]
\end{bmatrix}.
\]

Then there is a neighborhood of \( \theta_0 \) and an \( M_n > 0 \) such that

\[
\text{Cov}[a_i(\beta, \gamma, \beta_0, \gamma_0), a_j(\beta, \gamma, \beta_0, \gamma_0)] \leq M_n |\sigma_{ij}|.
\]

If the gradients of \( m_i(\beta, \gamma) \), \( s_i(\gamma) \), and \( \cos^{-1} r_i(\gamma) \) are uniformly bounded in \( n \), then there is a uniform bound \( M = M_n > 0 \) for all \( n \).
To prove Lemmas 1 and 2, we require additional results, stated as Lemmas 3 through 6. They are stated and proved below the proofs of Lemmas 1 and 2.

**Proof of Lemma 1.** Let \( A_i(\beta, \gamma) \) and \( B_i(\beta, \gamma) \) be as in Lemma 5, and assume for the moment that \( \beta \) and \( \gamma \) are fixed. Write \( a_i = a_i(\beta, \gamma, \gamma_0) \), \( A_i = A_i(\beta, \gamma) \), and \( B_i = B_i(\beta, \gamma) \). Note that \( A_i \) and \( B_i \) are probabilities of mutually exclusive events, and that \( a_i^2 = 1 \) when either event holds. Computing the variance directly,

\[
Var[a_i] = E[a_i^2] - E[a_i]^2
\]

\[
= (A_i + B_i) - (A_i - B_i)^2
\]

\[
= A_i(1 - A_i) + B_i(1 - B_i) + 2A_iB_i
\]

\[
\leq A_i + B_i + 2\max(A_i, B_i)
\]

\[
\leq 4M_n|\theta - \theta_0|,
\]

where \( M_n \) is the bound determined in Lemma 5. If the gradients are uniformly bounded, as in Lemma 5, the bound for \( Var[a_i] \) is uniform for all \( n \).

**Proof of Lemma 2.** Let \( A_i(\beta, \gamma) \) and \( B_i(\beta, \gamma) \) be as in Lemma 5, and let \( AA_{ij}(\beta, \gamma) \), \( AB_{ij}(\beta, \gamma) \), \( BA_{ij}(\beta, \gamma) \), and \( BB_{ij}(\beta, \gamma) \) be as in Lemma 6. Assume for the moment that \( \beta \) and \( \gamma \) are fixed. Suppressing the functional notation, we compute the covariance directly:

\[
Cov[a_i, a_j] = E[a_i a_j] - E[a_i]E[a_j]
\]

\[
= (AA_{ij} - AB_{ij} - BA_{ij} + BB_{ij}) - (A_i - B_i)(A_j - B_j)
\]

\[
= AA_{ij} - A_iA_j - AB_{ij} + A_iB_j - BA_{ij} + B_iA_j + BB_{ij} - B_iB_j.
\]

The result now follows from Lemma 6.
Lemma 3. Let $e_i \in \mathbb{R}^d$ denote the unit vector in the $i$th direction, and define $g : \mathbb{R}^d - \{e_i\} \to \mathbb{R}$ by
\[
g(u) = \cos^{-1}\left(\frac{e_i^T u}{|u|}\right).
\]
Then $|\nabla_u g| \to 1$ as $u \to e_i$.

Proof. A straightforward computation using the chain rule shows that
\[
\frac{\partial g}{\partial u_j} = \frac{u_i u_j - e_i^T e_j |u|^2}{|u|^2 \sqrt{|u|^2 - u_i^2}},
\]
Therefore,
\[
\sum_{j \neq i} \left|\frac{\partial g}{\partial u_j}\right|^2 = \sum_{j \neq i} \frac{|u_i u_j|^2}{|u|^4 (|u|^2 - u_i^2)} = \frac{u_i^2}{|u|^4 (|u|^2 - u_i^2)} \sum_{j \neq i} u_j^2 = \frac{u_i^2 (|u|^2 - u_i^2)}{|u|^4 (|u|^2 - u_i^2)} = \frac{u_i^2}{|u|^4}.
\]
Also,
\[
\left|\frac{\partial g}{\partial u_i}\right|^2 = \frac{|u_i^2 - |u|^2|^2}{|u|^4 (|u|^2 - u_i^2)} = |u|^{-4} (|u|^2 - u_i^2).
\]
Therefore,
\[
|\nabla_u g|^2 = \sum_{j=1}^n \left|\frac{\partial g}{\partial u_j}\right|^2 = |u|^{-4} (u_i^2 + |u|^2 - u_i^2) = |u|^{-2} \to 1
\]
as $u \to e_i$. \hfill \square

Lemma 4. Let $r_i(\gamma)$ be defined as in (15). If there is a neighborhood of $\gamma_0$ over which $\|V(\gamma)\|$ is bounded and bounded away from zero, and over which $\|d_\gamma L\|$ is bounded, then $\cos^{-1} r_i(\gamma)$ has bounded gradient near $\gamma_0$. Additionally, if the bounds for $V$ and $d_\gamma L$ are uniform in $n$, then so is the bound for $\cos^{-1} r_i(\gamma)$.

Proof. Define $\varphi_i : (\mathbb{R}^n - \{0\}) \times \mathbb{R}^+ \to \mathbb{R}$ by
\[
\varphi_i(u, \alpha) = \cos^{-1}\left(\frac{\alpha P_i u}{|\alpha u|}\right) = \cos^{-1}\left(\frac{P_i u}{|u|}\right).
\]
Then
\[
\cos^{-1} r_i(\gamma) = \varphi_i \left(L(\gamma_0)^{-1} L(\gamma) P_i^T, (P_i L(\gamma) V(\gamma_0) L(\gamma)^T P_i^T)^{-1/2}\right) = \varphi_i(u(\gamma), \alpha(\gamma)),
\]
where \( u : \mathbb{R}^q \to \mathbb{R}^n \) and \( \alpha : \mathbb{R}^q \to \mathbb{R}^+ \) are continuously differentiable near \( \gamma_0 \). Therefore,

\[
\nabla_\gamma \varphi_i = (\nabla_u \varphi_i) \frac{\partial u}{\partial \gamma} + (\nabla_\alpha \varphi_i) \frac{\partial \alpha}{\partial \gamma} = (\nabla_u \varphi_i) \frac{\partial u}{\partial \gamma}.
\]

Fix \( \epsilon > 0 \). By Lemma 3 (and the fact that \( u(\gamma) \) is continuous), \( |\nabla_\gamma \varphi_i(\gamma)| \leq 1 + \epsilon \) for \( \gamma \) sufficiently close to \( \gamma_0 \). Consequently, for every \( \gamma \) close enough to \( \gamma_0 \), and for each \( k = 1, \ldots, q \),

\[
|\nabla_\gamma \varphi_i| = \left| (\nabla_u \varphi_i) \frac{\partial u}{\partial \gamma_k} \right| \leq (1 + \epsilon) \left| \frac{\partial u}{\partial \gamma_k} \right|
\]

and

\[
|\nabla_\gamma \cos^{-1} r_i| \leq (1 + \epsilon) \left| \frac{\partial u}{\partial \gamma_k} \right| = (1 + \epsilon) \left| L(\gamma_0)^{-1} \frac{\partial L}{\partial \gamma_k} P_i \right| \leq (1 + \epsilon) \left\| L(\gamma_0)^{-1} \right\| \| d_\gamma L \|.
\]

By assumption, \( V(\gamma) \) is continuous and its norm is bounded and bounded away from zero. Therefore, the same conditions hold for \( L(\gamma) \), and \( \left\| L(\gamma_0)^{-1} \right\| \| d_\gamma L \| \) is bounded. This establishes a bound for \( |\nabla_\gamma \cos^{-1} r_i| \) that is uniform in \( n \) if the bounds for \( \| V(\gamma) \| \) and \( \| d_\gamma L \| \) are.

\( \square \)

**Lemma 5.** Let

\[
A_i(\beta, \gamma) = P (z_i(\beta, \gamma_0) \leq x < z_i(\beta, \gamma))
\]

and

\[
B_i(\beta, \gamma) = P (z_i(\beta, \gamma) \leq x < z_i(\beta, \gamma_0)),
\]

and suppose the conditions of Lemma 4 are met.

Then there is a neighborhood \( U \) of \( \theta_0 \) and an \( M_n > 0 \) such that for all \( \theta \in U \),

\[
A_i(\beta, \gamma) \leq M_n |\theta - \theta_0|
\]

and

\[
B_i(\beta, \gamma) \leq M_n |\theta - \theta_0|.
\]

51
Moreover, if the uniformity conditions in Lemma 4 are met, and if the gradient of \( m_i(\beta, \gamma) \) is uniformly bounded (in \( n \)), then there is an \( M > 0 \) such that \( M_n \leq M \) for all \( n \).

**Proof.** Let \( t_i(\gamma) = \cos^{-1} r_i(\gamma) \geq 0 \). A straightforward probability calculation, based on well-known facts about the bivariate normal distribution, produces \( A_i(\beta, \gamma) = A(m, s, t) \), where

\[
A(m, s, t) = \int_x^\infty \frac{\Phi \left( \frac{x - m - rsw}{s \sqrt{1 - r^2}} \right)}{\Phi \left( \frac{x - m - \cos(t)sw}{s \sin(t)} \right)} \phi(w)dw = \int_x^\infty \frac{\Phi \left( \frac{x - m - rsw}{s \sqrt{1 - r^2}} \right)}{\Phi \left( \frac{x - m - \cos(t)sw}{s \sin(t)} \right)} \phi(w)dw,
\]

(20)

\( m = m_i(\beta, \gamma), s = s_i(\gamma), r = r_i(\gamma), \) and \( t = t_i(\gamma) \). Note that \( A \rightarrow 0 \) as \( m \rightarrow 0, s \rightarrow 1, \) and \( t \downarrow 0 \). For each value of \( t > 0 \), there is a neighborhood containing \( t \) over which the integrand in (20) is bounded. Therefore, the derivatives of \( A \) with respect to \( m, s, \) and \( t \) can be obtained by differentiating the integrand and integrating the result. \( A \) does not possess a derivative at the point \( (m, s, t) = (0, 1, 0) \), but it is differentiable for \( t > 0 \) and the corresponding derivatives remain bounded as \( (m, s, t) \rightarrow (0, 1, 0) \). Consequently, the derivatives are bounded in a convex neighborhood \( U' \) of \( (0, 1, 0) \) by a constant \( M' > 0 \).

Thus, for any fixed \( t' > 0 \) with \( (0, 1, t') \in U' \), we have

\[
A(m, s, t) = A(0, 1, t') + \frac{\partial A}{\partial m} \bigg|_{m=m^*} m + \frac{\partial A}{\partial s} \bigg|_{s=s^*} (s - 1) + \frac{\partial A}{\partial t} \bigg|_{t=t^*} (t - t'),
\]

where \( m^* \) is between 0 and \( m, s^* \) is between 1 and \( s, \) and \( t^* \) is between \( t' \) and \( t \). [See Serfling (1980), Section 1.12, Theorem B.] From the properties of matrix norms, it follows that

\[
|A(m, s, t) - A(0, 1, t')| \leq \left( \left| \frac{\partial A}{\partial m} \right|_{m=m^*}, \left| \frac{\partial A}{\partial s} \right|_{s=s^*}, \left| \frac{\partial A}{\partial t} \right|_{t=t^*} \right) \cdot |(m - 1, t - t')^T|.
\]
Letting \( t' \downarrow 0 \), and using the bound for the derivatives and the convexity of \( U' \), we have

\[
|A(m, s, t)| \leq M'(m, s - 1, t)^T.
\]

Along the hyperplane \( t = 0 \), a similar probability calculation produces a bound that, without loss of generality, can be assumed to be \( \leq M' \). (Otherwise, take \( M' \) to be the minimum of the two bounds.)

By Lemma 4, \( t_i(\gamma) \) has bounded gradient near \( \gamma_0 \). Since

\[
\frac{\partial s_i}{\partial \nu_k} = \frac{1}{2} s_i(\gamma)^{-1} P_i(d_{\nu_k} L^T) V(\gamma_0) L(\gamma) P_i^T + P_i L(\gamma)^T V(\gamma_0) (d_{\nu_k} L) P_i^T,
\]

the gradient of \( s_i(\gamma) \) is bounded by \( |s_i(\gamma)|^{-1} \), \( \|d_{\nu_k} L^T\| \), \( \|V(\gamma_0)\| \), and \( \|d_{\nu_k} L\| \). The latter two quantities are bounded by assumption; being upper-triangular, \( d_{\nu_k} L^T \) has the same norm as its transpose; and \( s_i(\gamma) \) can be made arbitrarily close to 1. Finally,

\[
\frac{\partial m_i}{\partial \nu_k} = P_i(d_{\nu_k} L^T) X(\beta - \beta_0)
\]

and

\[
\frac{\partial m_i}{\partial \beta} = P_i L(\gamma)^T X,
\]

and \( X \) has bounded elements, so the gradient of \( m_i(\beta, \gamma) \) is bounded (although not necessarily uniformly in \( n \)). It follows that there is a neighborhood \( U \) of \( \theta_0 \) and an \( M'' > 0 \) such that \( |m_i(\beta, \gamma)| < M'' |\theta - \theta_0|, |s_i(\gamma) - 1| < M'' |\theta - \theta_0|, \) and \( |t_i(\gamma)| < M'' |\theta - \theta_0| \).

With \( M_n = 3M'M'' \), we have

\[
|A_i(\beta, \gamma)| \leq M'|m_i(\beta, \gamma), s_i(\gamma) - 1, t_i(\gamma)^T| \leq 3M'M''|\theta - \theta_0| \leq M_n|\theta - \theta_0|.
\]

If \( \|V(\gamma)\| \) is uniformly bounded and bounded away from zero, and if \( \|d_{\nu_k} L\| \) is uniformly bounded, then the gradients of \( s_i(\gamma) \), and \( t_i(\gamma) \) are uniformly bounded. For \( t_i(\gamma) \), this follows from Lemma 4, and for \( s_i(\gamma)^2 \) it follows from the previous paragraph. If, in
addition, \( m_i(\beta, \gamma) \) is uniformly bounded, then \( 3MM'' \) is a bound for all \( n \). A similar proof establishes the same result for \( B_i(\beta, \gamma) \). □

**Lemma 6.** Let

\[
AA_{ij}(\beta, \gamma) = P \left( z_i(\beta_0, \gamma_0) \leq x < z_i(\beta, \gamma), z_j(\beta_0, \gamma_0) \leq x < z_j(\beta, \gamma) \right),
\]

\[
AB_{ij}(\beta, \gamma) = P \left( z_i(\beta_0, \gamma_0) \leq x < z_i(\beta, \gamma), z_j(\beta, \gamma) \leq x < z_j(\beta_0, \gamma_0) \right),
\]

\[
BA_{ij}(\beta, \gamma) = P \left( z_i(\beta, \gamma) \leq x < z_i(\beta_0, \gamma_0), z_j(\beta_0, \gamma_0) \leq x < z_j(\beta_0, \gamma_0) \right),
\]

and

\[
BB_{ij}(\beta, \gamma) = P \left( z_i(\beta, \gamma) \leq x < z_i(\beta_0, \gamma_0), z_j(\beta, \gamma) \leq x < z_j(\beta_0, \gamma_0) \right).
\]

Also, let \( A_i(\beta, \gamma) \) and \( B_i(\beta, \gamma) \) be defined as in Lemma 5, and let

\[
\sigma_{ij}(\beta, \gamma) = \begin{bmatrix}
\sigma_{i0j} \\
\sigma_{j1i} \\
\sigma_{ij2}
\end{bmatrix} = \begin{bmatrix}
\text{Cov}[z_i(\beta, \gamma), z_j(\beta, \gamma)] \\
\text{Cov}[z_i(\beta_0, \gamma), z_j(\beta, \gamma)] \\
\text{Cov}[z_i(\beta_0, \gamma_0), z_j(\beta, \gamma)]
\end{bmatrix}.
\]

If the conditions of Lemma 4 are met, then

\[
AA_{ij}(\beta, \gamma) - A_i(\beta, \gamma)A_j(\beta, \gamma) = O(\|\sigma_{ij}\|),
\]

and similar relationships hold for \( AB_{ij}(\beta, \gamma) \), \( BA_{ij}(\beta, \gamma) \), and \( BB_{ij}(\beta, \gamma) \). The bound for these quantities is uniform if the bounds for \( m_i(\beta, \gamma) \) and the bounds in Lemma 4 are uniform in \( n \).

**Proof.** As in the proof of Lemma 5, we use a limiting Taylor’s Theorem argument. We first express \( AA_{ij} \) as the integral of the joint density of \( z_i(\beta, \gamma), z_i(\beta_0, \gamma_0), z_j(\beta, \gamma), \) and \( z_j(\beta_0, \gamma_0) \). Defining the covariance and mean of the corresponding multivariate normal,
let

$$W_{ij} = \begin{bmatrix}
    s_i^2 & s_i^2 \cos(t_i) & \sigma_{ij0} & \sigma_{ij1} \\
    s_i^2 \cos(t_i) & 1 & \sigma_{ij2} & 0 \\
    \sigma_{ij0} & \sigma_{ij2} & s_j^2 & s_j^2 \cos(t_j) \\
    \sigma_{ij1} & 0 & s_j^2 \cos(t_j) & 1
\end{bmatrix}, \quad (21)$$

$$W_{ij}^* = \begin{bmatrix}
    s_i^2 & s_i^2 \cos(t_i) & 0 & 0 \\
    s_i^2 \cos(t_i) & 1 & 0 & 0 \\
    0 & 0 & s_j^2 & s_j^2 \cos(t_j) \\
    0 & 0 & s_j^2 \cos(t_j) & 1
\end{bmatrix},$$

and $\xi_{ij} = (m_i, 0, m_j, 0)^T$, where $t_i$ and $t_j$ are defined as in the proof of Lemma 5 and the functional dependence on $(\beta, \gamma)$ has been omitted from the notation. Then

$$AA_{ij} = \frac{|W_{ij}|^{-1/2}}{(2\pi)^2} \int_{-\infty}^{\infty} \int_{-\infty}^{x} \int_{x}^{\infty} \int_{-\infty}^{\infty} \exp \left( -\frac{1}{2} (w - \xi_{ij})^T W_{ij}^{-1} (w - \xi_{ij}) \right) dw$$

(22)

and

$$A_iA_j = \frac{|W_{ij}^*|^{-1/2}}{(2\pi)^2} \int_{-\infty}^{\infty} \int_{-\infty}^{x} \int_{x}^{\infty} \int_{-\infty}^{\infty} \exp \left( -\frac{1}{2} w^T W_{ij}^{*-1} w \right) dw.$$ \quad (23)

Note that $\sigma_{ij1} \to 0$, $\sigma_{ij2} \to 0$, $\sigma_{ij0} \to 0$, $s_i \to 1$, $s_j \to 1$, $t_i \to 0$, $t_j \to 0$, $m_i \to 0$, and $m_j \to 0$ as $\theta \to \theta_0$. We first show that the derivatives of $AA_{ij}$ with respect to $\sigma_{ij1}$, $\sigma_{ij2}$, and $\sigma_{ij0}$ are bounded near $\theta = \theta_0$. Note that $W_{ij}$ and $W_{ij}^*$ are singular at this limit. We compute the derivative using formal matrix differentiation; simplification produces

$$d_k AA_{ij} = \frac{|W_{ij}|^{-1/2}}{(2\pi)^2} \int_{-\infty}^{\infty} \int_{-\infty}^{x} \int_{x}^{\infty} \int_{-\infty}^{\infty} \exp \left( -\frac{1}{2} w^T W_{ij}^{-1} w \right) h_k(w) dw,$$ \quad (24)

where

$$h_k(w) = \frac{1}{2} \left[ w^T W_{ij}^{-1} d_k W_{ij} W_{ij}^{-1} w - \text{tr}(W_{ij}^{-1} d_k W_{ij}) \right]$$
and $d_k$ denotes the matrix derivative with respect to one of the $\sigma_{ij}$ parameters appearing in (21). It is easily shown that

$$\text{tr}(W_{ij}^{-1}d_kW_{ij}) = 0$$

when $\sigma_{ij1} = \sigma_{ij2} = \sigma_{ij0} = 0$. After a change of variables to $u = U_{ij}w$, where $U_{ij}^T U_{ij} = W_{ij}^{-1}$, (24) produces

$$d_k AA_{ij} = \frac{1}{2(2\pi)^2} \int_{\Omega} \exp\left(-\frac{1}{2}u^T u\right) u^T (U_{ij} d_k W_{ij} U_{ij}^T) u du,$$  

(25)

where $\Omega \subset \mathbb{R}^4$ represents the appropriately transformed region. It can be shown that when $\sigma_{ij1} = \sigma_{ij2} = \sigma_{ij0} = 0$, the integral in (25) can be expressed as the sum of integrals of the form

$$\frac{\zeta(t_i, t_j)}{2} \int_{x}^{\infty} \int_{-\infty}^{\alpha_i x - \beta_i u_2} u_1^{d_1} u_2^{1-d_1} \phi(u_1) \phi(u_2) du_1 du_2 \int_{x}^{\infty} \int_{-\infty}^{\alpha_j x - \beta_j u_4} u_3^{d_3} u_4^{1-d_3} \phi(u_3) \phi(u_4) du_3 du_4,$$

where $\alpha_i = \sin(t_i)^{-1}$, $\beta_i = \cos(t_i)\sin(t_i)^{-1}$, $\alpha_j = \sin(t_j)^{-1}$, $\beta_j = \cos(t_j)\sin(t_j)^{-1}$, $d_i \in \{0, 1\}$, $d_j \in \{0, 1\}$, and $\zeta(t_i, t_j) = O(\sin(t_i)^{-d_i}\sin(t_j)^{-d_j})$. Direct computation of the double integrals shows that the entire integral is bounded. Thus, (24) remains bounded as $\theta \to \theta_0$.

Now, $AA_{ij} - A_{i} A_{j} = 0$ when $\sigma_{ij1} = \sigma_{ij2} = \sigma_{ij0} = 0$, whatever the values of $s_i$, $s_j$, $t_i$, $t_j$, $m_i$, and $m_j$ may be. With the restriction $t_i > 0$ and $t_j > 0$, $AA_{ij} - A_{i} A_{j}$ is a continuously differentiable function, so derivatives with respect to other parameters in (21) are zero at $\sigma_{ij1} = \sigma_{ij2} = \sigma_{ij0} = 0$. Consequently,

$$AA_{ij} - A_{i} A_{j} = \frac{\partial AA_{ij}}{\partial \sigma_{ij}} \bigg|_{\sigma_{ij0}} \sigma_{ij} = D(\theta^*)^T \sigma_{ij},$$

where $D(\theta)^T$, the derivative with respect to $\sigma_{ij}$, is a continuous map and is evaluated near $s_i = s_j = 1$, $m_i = m_j = 0$, and $t_i > 0$, $t_j > 0$. Since $D(\theta)$ remains bounded as
$t_i \to 0$ and $t_j \to 0$, an argument similar to that used in the proof of Lemma 5 shows that there is an $M_{ij} > 0$ such that

$$|AA_{ij} - A_iA_j| < M_{ij}|\sigma_{ij}|.$$ 

As in the proof of Lemma 5 the bound $M_{ij}$ is uniform if the bounds for $m_i(\beta, \gamma)$ and the bounds in Lemma 4 are uniform. The proofs for $AB_{ij}$, $BA_{ij}$ and $BB_{ij}$ are identical, except that the ranges of integration in (22) and (23) differ. 

**Theorem 4 (Clustered Designs).** Let $V(\gamma) = (V_1(\gamma), ..., V_N(\gamma))$ have block diagonal structure, with $N$ blocks having dimension at most $k \times k$, and for each block $h$ let $L_h(\gamma)L_h(\gamma)^T = V_h(\gamma)^{-1}$. Assume that, near $\gamma_0$, the $L^2$ norm $\|V_h(\gamma)\|$ is uniformly bounded and bounded away from zero, both in block number $h$ and in sample size $n$, and that for each $j = 1, ..., q$, $\|d_{\gamma_j} L_h\|$ is uniformly bounded in both $h$ and $n$. If $\hat{\theta}_n \xrightarrow{P} \theta_0$, then the conditions of Theorem 3 hold and the conclusion (6) follows.

**Proof.** For this proof, we employ double subscripting notation. In particular, let $y = (y_1^T, ..., y_N^T)^T$, where $\text{Var}[y_h] = V_h(\gamma)$ and $y_h$ and $y_{h'}$ are independent if $h \neq h'$, and let $d_h$ be the dimension of $V_h(\gamma)$. Let $P^*_h : \mathbb{R}^{d_h} \to \mathbb{R}$ be projection onto the $i$th component in block $h$. Define

$$z_h(\beta, \gamma) = P^*_h L_h(\gamma)^T(y_h - X_h\beta),$$

which is consistent with the corresponding definition above; define $a_{hi}(\beta, \gamma; \beta', \gamma')$, and all other quantities analogously (with double subscript notation).

The block-diagonal structure of $V(\gamma)$ insures that, locally, $V(\gamma)$ is uniformly bounded in $n$, for the norm of $V(\gamma)$ can be no more than $\max_h \|V_h(\gamma)\|$. Similarly, near $\gamma_0$, $V(\gamma)$ is uniformly bounded away from zero. It follows that $L(\gamma)$ is also uniformly bounded and
bounded away from zero. Similar considerations apply to $d_y L$. Since

$$\left| \frac{\partial m_{h\beta}}{\partial \beta} \right| \leq \| L_h(\gamma)^T \| \| X_h \|$$

and

$$\left| \frac{\partial m_{h\beta}}{\partial \gamma_k} \right| \leq \| d_{\gamma_k} L_h^T \| \| X_h \| |\beta_0 - \beta|,$$

a uniform bound is obtained for $m_{h\beta}(\gamma, \beta)$ using the bounds for $\| L_h(\gamma)^T \|$ and $\| d_{\gamma_k} L_h \|$ and the fact that $X_h$ has bounded elements and dimensions. Thus the uniformity conditions of Lemma 1 are met.

Writing $a_{hi} = a_{hi}(\beta, \gamma; \beta_0, \gamma_0)$, it follows from Lemma 1 that there is an $M > 0$ such $\text{Var}[a_{hi}] \leq M |\theta - \theta_0|$. Consequently, $\text{Cov}[a_{hi}, a_{hi'}] \leq (\text{Var}[a_{hi}][\text{Var}[a_{hi'}])^{1/2} \leq M |\theta - \theta_0|$

if $h = h'$ and $\text{Cov}[a_{hi}, a_{hi'}] = 0$ if $h \neq h'$. Thus,

$$n^{-1} \sum_{h, i, h', i'} |\text{Cov}[a_{hi}, a_{hi'}]| = n^{-1} \sum_{h = h', i = i'} \text{Var}[a_{hi}] + n^{-1} \sum_{h = h', i \neq i'} |\text{Cov}[a_{hi}, a_{hi'}]|$$

$$= n^{-1} \sum_{h = h', i = i'} \text{Var}[a_{hi}] + n^{-1} \sum_{h = h', i \neq i'} |\text{Cov}[a_{hi}, a_{hi'}]|$$

$$\leq n^{-1} \cdot n \cdot M |\theta - \theta_0| + n^{-1} \cdot N k^2 \cdot M |\theta - \theta_0|$$

$$\leq M |\theta - \theta_0| + k^2 M |\theta - \theta_0|$$

$$= (1 + k^2) M |\theta - \theta_0|.$$ 

This establishes condition (18) in Theorem 3.

Longitudinal mixed effects models (Laird and Ware, 1982) fall into the class of models described by Theorem 4, as long as the design matrix corresponding to the random effects is uniformly bounded and bounded away from zero, and the covariance matrix $\Delta$ of the random effects satisfies requirements similar to those required by $V(\gamma)$ in the statement of the theorem.
Corollary 1 (Longitudinal Mixed Effects Models). Let \( V(\gamma) = (V_1(\gamma), ..., V_N(\gamma)) \) have block diagonal structure, where for each \( h \), \( V_h(\gamma) \) has dimension \( d_h \leq k \),

\[
V_h(\gamma) = \gamma_1 I_{d_h} + Z_h \Delta(\gamma_2, ..., \gamma_q) Z_h^T,
\]

and \( Z_h \) is a \( d_h \times (q-1) \) matrix that is bounded and bounded away from zero. Furthermore, assume that \( \gamma_1 > 0 \), that \( \Delta(\gamma_2, ..., \gamma_q) \) has bounded norm near \( \gamma_0 \), and that \( d_h \Delta \) is bounded in a neighborhood of \( \gamma_0 \). If \( \hat{\theta}_n \overset{p}{\rightarrow} \theta_0 \), then the conditions of Theorem 3 hold and the conclusion (6) follows.

Condition (6) also applies in models that do not have block diagonal variance. The following theorem, which establishes (6) for linear model with ARMA errors, is an example of the utility of our method for unclustered designs.

Theorem 5 (ARMA Errors). Let \( \varepsilon(\beta) = y - X\beta \) be a time-series vector from an autoregressive moving-average process of order \( (\bar{p}, \bar{q}) \). Thus, for \( i > \max(\bar{p}, \bar{q}) \),

\[
B(\gamma) \tilde{\varepsilon}_i(\beta) = A(\gamma) \tilde{z}_i(\beta, \gamma),
\]

where \( \tilde{\varepsilon}_i(\beta) = (\varepsilon_{i-\bar{p}}(\beta), ..., \varepsilon_i(\beta))^T \), \( \tilde{z}_i(\beta, \gamma) = (z_{i-\bar{q}}(\beta, \gamma), ..., z_i(\beta, \gamma))^T \), and \( A(\gamma) \) and \( B(\gamma) \) are conformable matrices, depending on \( \gamma \), such that \( A_1(\gamma) \neq 0 \), \( A_{\bar{q}+1}(\gamma) = 1 \), \( B_1(\gamma) \neq 0 \), and \( B_{\bar{p}+1}(\gamma) = -1 \). If \( \hat{\theta}_n \overset{p}{\rightarrow} \theta_0 \), then the conditions of Theorem 3 hold and the conclusion (6) follows.

Proof. We first demonstrate that the conditions of Lemma 1 and Lemma 2 are met. Because the covariance terms of \( V(\gamma) \) die off exponentially for widely separated observations [see Davidson (1994), page 215], the \( L^2 \) norm of \( V(\gamma) \) is uniformly bounded in \( n \). In addition, the variance of \( \varepsilon_i(\beta) \) is a positive value (and constant after finitely many observations), so \( ||V(\gamma)|| \) is bounded away from zero and \( ||L(\gamma)|| \) is bounded. The matrix

59
$L(\gamma)$ must produce residuals that satisfy (26), so that $A(\gamma) E[z_i(\beta, \gamma)] = B(\gamma) \bar{X}_i(\beta_0 - \beta)$, where $\bar{X}_i$ is an appropriate submatrix of $X$. It follows that $m_i(\beta, \gamma)$ is uniformly bounded in $n$. Therefore the lemmas apply. Equation (26) implies that for $i, i' > \max(\bar{p}, \bar{q})$,

$$Cov[z_i(\beta, \gamma), z_{i'}(\beta', \gamma')] = B(\gamma)Cov[\bar{\epsilon}_i(\beta), \bar{\epsilon}_i(\beta')]|B(\gamma')^T.$$

Since $B(\gamma)$ and $B(\gamma')$ are bounded near $\gamma_0$ and $Cov[\bar{\epsilon}_i(\beta), \bar{\epsilon}_i(\beta')] = Cov[\bar{\epsilon}_i(\beta_0), \bar{\epsilon}_i(\beta_0)]$ approaches zero exponentially as $i - i' \to \infty$ (after finitely many terms), it follows that $Cov[z_i(\beta, \gamma), z_{i'}(\beta', \gamma')] = O(\rho^{i-i'})$ for some nonnegative $\rho < 1$ that depends only on $\gamma_0$. Lemmas 1 and 2 imply that $|Cov[a_i(\beta, \gamma, \beta_0, \gamma_0), a_{i'}(\beta, \gamma, \beta_0, \gamma_0)]| = O(\rho^{i-i'}|\theta - \theta_0|)$, thus

$$n^{-1} \sum_{i, i' = 1}^{n} |Cov[a_i(\beta, \gamma, \beta_0, \gamma_0), a_{i'}(\beta, \gamma, \beta_0, \gamma_0)]| = n^{-1} \sum_{i, i' = 1}^{n} O(\rho^{i-i'}|\theta - \theta_0|)$$

$$= O(|\theta - \theta_0|) \left(1 + n^{-1} \sum_{i \neq i'} O(\rho^{i-i'})\right)$$

$$= O(|\theta - \theta_0|).$$

The last equality follows from the fact that

$$\sum_{i \neq i'} \rho^{i-i'} = 2 \sum_{i < i'} \rho^{i-i'} \leq 4(n - 1) \sum_{i = 1}^{\infty} \rho^i = 4(n - 1)\rho(1 - \rho)^{-1}.$$ 

Therefore condition (18) of Theorem 3 is satisfied.

Models with more complicated correlation structures also fit within our theory. The following theorem justifies the method for models whose correlation is based on the Kronecker product of correlations from different components. Note that this theorem can be applied to a model involving a finite number of correlated time series. In this case, (27) is proven in a manner similar to the proof of Theorem 5, given in the appendix.

**Theorem 6 (Products).** Let $V_1(\gamma)$ be a $k \times k$ covariance matrix (fixed with respect to $n$) and let $V_2(\gamma)$ be a covariance matrix whose norm is bounded and bounded away from
zero. Let $L_1(\gamma)$ and $L_2(\gamma)$ be the corresponding Cholesky matrices, and assume that the norm of $L_2(\gamma)$ has bounded gradient. Finally, assume that \( \|L_2(\gamma)^T X\| \) is bounded and that there is an $M > 0$ such that

\[
\sum_{i \neq j} \text{Cov}[z_i(\beta, \gamma), z_j(\beta', \gamma')] \leq M|\theta - \theta_0|, \tag{27}
\]

where $(\beta', \gamma') = (\beta, \gamma)$ or $(\beta', \gamma') = (\beta_0, \gamma_0)$. If $V(\gamma) = V_1(\gamma) \otimes V_2(\gamma)$, then the conditions of Theorem 3 apply to $V(\gamma)$ and the conclusion (6) follows.

Proof. Note that the Cholesky matrix of $V_1(\gamma) \otimes V_2(\gamma)$ is $L_1(\gamma) \otimes L_2(\gamma)$. Since $V_1(\gamma)$ has fixed dimension, it is easy to show that $V_1(\gamma) \otimes V_2(\gamma)$ has a norm that is bounded and bounded away from zero. A proof similar to that used for Theorem 4 can be used to establish condition (18) of Theorem 3.

Theorems 3, 4, 5, and 6 address condition (6) in Theorem 1. However, condition (7) is also difficult to verify in practice. Theorem 7, proven by Taylor expansion, gives plausible conditions that insure (7).

**Theorem 7.** In a neighborhood of $\theta_0$, let $\| V_n(\gamma) \|$ be uniformly bounded and uniformly bounded away from zero. Assume, in addition, that $L_n(\gamma)$ is continuously twice-differentiable, and that for each $j, j' = 1, \ldots, q$, $\| d_{\gamma} L_n(\gamma) \|$ and $\| d_{\gamma_j} d_{\gamma_j'} L_n(\gamma) \|$ are uniformly bounded near $\theta_0$. Finally, suppose that there is a constant vector $c_0$ such that

\[
n^{-1} \phi(x) \sum_{i=1}^{n} \left( \frac{x \partial^2 m_i}{2 \partial \gamma_1^2} \right) \left. + \frac{\partial m_i}{\partial \gamma} \right|_{\theta=\theta_0} \to c_0^T \tag{28}
\]

as $n \to \infty$. Then condition (7) of Theorem 1 is satisfied.

Proof. The proof is established by second-order Taylor expansion. The details appear in our technical report. \[\square\]
Condition (28) is difficult to verify for arbitrary sequences, but is plausible when there is some uniformity to the model sequence. For example, if there are exchangeable units of sampling, then (28) will be true. Note that \( \partial m_i / \partial \beta = P_i L(\gamma)^T X, \partial s^2_i / \partial \beta = 0, \)

\[
\frac{\partial s^2_i}{\partial \gamma} = P_i [(d_\gamma L^T) V(\gamma_0) L(\gamma) + L(\gamma)^T V(\gamma_0) (d_\gamma L)] P_i^T,
\]

and

\[
\frac{\partial m_i}{\partial \gamma} = P_i [d_\gamma L^T] X (\beta_0 - \beta).
\]

Evaluated at \( \theta_0, \partial m_i / \partial \beta = P_i L(\gamma_0)^T X \) and \( \partial m_i / \partial \gamma = 0. \) Thus, (28) reduces to plausible regularity in the covariate matrix \( X \) and in the correlation structure implied by \( V(\gamma). \)

Similar considerations apply to condition (12) of Theorem 2. We refrain from stating another theorem, but remark that conditions identical to those described for \( V(\gamma) \) and \( L(\gamma) \) in Theorem 7, together with a condition similar to (28), lead to the desired conclusion (12).