NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

University of Wollongong

Working Paper

04-17

Computational Aspects of the EM Algorithm for Spatial Econometric Models with Missing Data

Thomas Suesse and Andrew Zammit-Mangion

National Institute for Applied Statistics Research Australia, University of Wollongong,
Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845.
Email: anica@uow.edu.au

Computational Aspects of the EM Algorithm for Spatial Econometric Models with Missing Data

Thomas Suesse and Andrew Zammit-Mangion

National Institute for Applied Statistics Research Australia
University of Wollongong, New South Wales 2522, Australia

Abstract

Maximum-likelihood (ML) estimation with spatial econometric models is a long-standing problem that finds application in several areas of economic importance. The problem is particularly challenging in the presence of missing data, since there is an implied dependence between all units, irrespective of whether they are observed or not. Out of the several approaches adopted for ML estimation in this context, that of LeSage and Pace (2004) stands out as one of the most commonly used with spatial econometric models due its ability to scale with the number of units. Here, we review their algorithm, and consider several similar alternatives that are also suitable for large datasets. We compare the methods through an extensive empirical study and conclude that, while the approximate approaches are suitable for large sampling ratios, for small sampling ratios the only reliable algorithms are those that yield exact ML or restricted ML estimates.

Keywords: EM algorithm; missing data; spatial autoregressive models; spatial-errors models

1 Introduction

Spatial autoregressive models (SAMs) and spatial-errors models (SEMs) are popular spatial econometric linear models that take the dependence of the response variable of neighbouring units into account. These n units are typically spatial locations, although in principle any distance metric may be used to establish proximity upon which neighbourhood definition is based. The unit dependence is usually represented by a contiguity matrix $\mathbf{W}$, which has non-zero entries W_{ij} if i is treated as a neighbour of j, and where $\mathbf{W}$ is not necessarily

symmetric. Common methods to form $\mathbf{W}$ are first order contiguity relations and nearest neighbours [1, 2]. Often $\mathbf{W}$ is sparse in the sense that many entries W_{ij} are zero. Maximum likelihood (ML) estimation for such models where all units are observed was pioneered by Ord [1].In recent years, the standard SAMs and SEMs have been extended to model added complexity. For example [3] considers a model that combines a SAM with a SEM, [4] considers a SAM with heteroskedastic disturbances and [5] considers a non-parametric model for the mean.

The $n \times n$ matrix $\mathbf{W}$ is constructed based on the n units of interest. If data pertaining to each unit (hereby termed 'complete' data) are observed, then ML estimation can be accomplished in a straightforward fashion [1]. However, if missing data are present, that is, Y is only observed for n_s units, where $n_s < n$, estimation becomes more challenging as the contiguity matrix is still defined for all the n units.

Estimation with such spatial autoregressive models under the presence of missing data has been extensively studied using, for example, the generalised method of moments and least squares [6] and approximate Bayesian methods [7]. There are also a few extensions of these estimation methods to more complex models. For example, [8] considers a spectral EM algorithm for a spatial model of a similar form to the SAM and the SEM but with an added measurement error term.

An approach that is given special importance in this article is the method of LeSage and Pace [9] (denoted by LP04 in the remainder of this paper), who mainly focused on ML estimation using an iterative algorithm. LP04 also outlined how one could proceed using Markov Chain Monte Carlo methods in a Bayesian approach, although no implementation details were given. The ML approach of LP04 is most widely used due to its ability to scale well with complete-data size n. As we show later, although LP04's approach resembles an expectation maximisation (EM) algorithm (a natural option in a missing-data context, see [10]), it is not an EM algorithm, due to the various assumptions adopted. Specifically, when viewed as an EM algorithm, it becomes clear that the algorithm in LP04 alters the E-steps and M-steps such that some computational bottlenecks are avoided. In this paper we review their approach, and show how it can be made into a (valid) EM algorithm with a few modifications.This EM algorithm is, as expected, less favourable computationally than its simplified counterpart. We therefore additionally propose other approximations that may be used in order to achieve tractability. However, altering the

M-steps and E-steps is not advisable in general; in this article, we find that at low sampling ratios ($n_s \ll n$), it is crucial that the exact EM algorithm, or its restricted counterpart that we also derive and implement, are employed in order to obtain reliable estimates.

In Section 2, we introduce the class of spatial econometric models of interest, review the general EM algorithm and propose a variant that allows for restricted ML estimates. In Section 3 we show how LP04 is in fact an approximate EM algorithm and offer various approximations that also could be used to speed up the EM algorithm for larger datasets. In Section 4 a simulation study is conducted to compare the various approaches for the SAM and SEM models; we see that LP04's approach and other approximate methods work well for large sampling ratios ($n_s > n/2$), but do not perform well for smaller ratios ($n_s \ll n$). In Section 5 we confirm the simulation results on a housing dataset describing house prices in Lucas County (Ohio, USA) consisting of $n = 25,357$ units. Section 6 concludes the work.

2 Modelling and estimation

2.1 The SAM and the SEM

The two lattice models of focus in this paper are the spatial autoregressive model (SAM) and the spatial-errors model (SEM). Let $\mathbf{y} = (y_1, \ldots, y_n)^T$ be the n-vector of the response variable, $\mathbf{X}$ be the $n \times p$ design matrix containing the explanatory variables and $\mathbf{W}$ be the $n \times n$ contiguity matrix described in Section 1 with $W_{ii} = 0$ for all i. The SAM is given by

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{1}$$

while the SEM by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{v}, \tag{2}$$

$$\mathbf{v} = \rho\mathbf{W}\mathbf{v} + \mathbf{e}, \tag{3}$$

where $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$, $\mathbf{I}_n$ is the identity matrix of size $n \times n$ and σ^2 is a variance parameter. The parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ contains the fixed effects parameters and ρ is the spatial dependence or autocorrelation parameter. The two models imply that $\mathbf{y}$ is multivariate normal with mean $E(\mathbf{y}) \equiv \boldsymbol{\mu}$ and

covariance $\text{Cov}(\mathbf{y}) \equiv \boldsymbol{\Sigma}$. The terms of interest associated with these two models are specified in Table 9 in Appendix A. To have a valid covariance matrix $\boldsymbol{\Sigma}$ for the SAM and the SEM, it is required that $\rho \notin \{\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_n}\}$, where $\lambda_1, \ldots, \lambda_n$ are the n eigenvalues of the matrix $\mathbf{W}$ [11]. When $\mathbf{W}$ is row- or column-normalised (rows or columns sum to 1), as is common in practice, then $\lambda_i \leq 1$ and ρ is typically restricted to $-1 \leq \rho \leq 1$ for which $\boldsymbol{\Sigma}$ always exists.

2.2 Maximum Likelihood Estimation using the EM algorithm

Let u denote the subset of units for which the response variable is unobserved and s denote the units forming the sample for which the response variable is observed. Let $\mathbf{y}_u$ denote the vector containing missing responses and $\mathbf{y}_s$ the vector with observed responses. The complete-data vector is then denoted by $\mathbf{y} \equiv (\mathbf{y}_s^T, \mathbf{y}_u^T)^T$.

The likelihood of $\mathbf{y}_s$ can be obtained by integrating over the unobserved data,

$$f(\mathbf{y}_s|\boldsymbol{\theta}) = \int f(\mathbf{y}|\boldsymbol{\theta}) \mathrm{d}\mathbf{y}_u, \tag{4}$$

where $f(\mathbf{y}|\boldsymbol{\theta})$ is the density of the complete data and $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \rho, \sigma^2)^T$ contains the unknown parameters. It is difficult to maximise the likelihood (4) directly, and iterative methods such as the EM algorithm are convenient alternatives for this purpose.

The EM algorithm is an iterative algorithm that consists of two key steps, the E(expectation)-step and the M(maximisation)-step [12]. The E-step calculates the expected value of the complete-data log-likelihood, $\log f(\mathbf{y}; \boldsymbol{\theta})$, with respect to the conditional density of $\mathbf{y}_u$ given $\mathbf{y}_s$ and a fixed parameter vector $\boldsymbol{\theta}'$. This expectation is denoted by $E_{u|s;\boldsymbol{\theta}'}(\cdot)$ or $E_{u|s}(\cdot)$ in short. This expectation is then maximised with respect to $\boldsymbol{\theta}$ in the subsequent M-step. In summary the EM algorithm proceeds as follows:

- (E-Step) calculate $Q(\boldsymbol{\theta}|\boldsymbol{\theta}') \equiv E_{u|s;\boldsymbol{\theta}'}(\log f(\mathbf{y}; \boldsymbol{\theta}))$.
- (M-Step) maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}')$ with respect to $\boldsymbol{\theta}$.
- Set the new $\boldsymbol{\theta}'$ equal to the result of the M-step.
- Repeat the above three steps until convergence.

To perform the E-step, an expression for the complete-data log-likelihood is required. Let $\boldsymbol{\Sigma} \equiv \sigma^2\mathbf{V}$. The complete-data log-likelihood expressed in terms of $\boldsymbol{\mu}$, $\mathbf{M} \equiv \mathbf{V}^{-1}$, $\omega \equiv \sigma^2$ and $\mathbf{r} \equiv \mathbf{y} - \boldsymbol{\mu}$ has the following form:

$$\log f(\mathbf{y};\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\omega + \frac{1}{2}\log|\mathbf{M}| - \frac{1}{2\omega}\mathbf{r}^T\mathbf{M}\mathbf{r}. \tag{5}$$

Table 9 lists the expressions needed for the evaluation of the complete-data log-likelihood $L_c \equiv \log f(\mathbf{y};\boldsymbol{\theta})$. Note that no matrix inversions are needed for $\mathbf{r}^T\mathbf{M}\mathbf{r}$ but only matrix multiplications. An expected information matrix, $\mathcal{J}(\boldsymbol{\beta}, \rho, \sigma^2 \mid \mathbf{y})$, for the complete-data case was provided by Ord [1].

The numerically difficult term in Table 9 is $\log|\mathbf{M}|$, which is proportional to $\log|\mathbf{A}|$ with $\mathbf{A} = \mathbf{I} - \rho\mathbf{W}$ generally sparse since $\mathbf{W}$ is usually sparse. The `R` package `Matrix` [13] can calculate determinants efficiently for symmetric sparse matrices. Recall, however, that $\mathbf{W}$ is not constrained to be symmetric for the models we consider. For some types of non-symmetric $\mathbf{W}$, for example row-normalised $\mathbf{W}$, instead of calculating $|\mathbf{A}|$ directly, the determinant of $\mathbf{I}-\rho\mathbf{W}_{sim}$ can be found, where $|\mathbf{A}| = |\mathbf{I} - \rho\mathbf{W}_{sim}|$ and $\mathbf{W}_{sim}$ is a *similar* symmetric matrix in the sense that it has the same determinant as the non-symmetric $\mathbf{W}$. Let $\widetilde{\mathbf{W}}$ be the unnormalised symmetric contiguity matrix from which the row-normalised $\mathbf{W}$ is constructed. Then $\mathbf{W} = Diag(\mathbf{z})^{-1}\widetilde{\mathbf{W}}$, where $Diag(\mathbf{z})$ returns a diagonal matrix with the vector $\mathbf{z}$ containing the values of the row-sums of $\widetilde{\mathbf{W}}$ on its diagonal. In terms of the determinant $|\mathbf{W}| = |\,Diag(\mathbf{z})^{-1}\widetilde{\mathbf{W}}| = |\,Diag(\sqrt{\mathbf{z}})^{-1}\widetilde{\mathbf{W}}\,Diag(\sqrt{\mathbf{z}})^{-1}| = |\mathbf{W}_{sim}|$, where in this case the square-root is taken element-wise and $\mathbf{W}_{sim} \equiv Diag(\sqrt{\mathbf{z}})^{-1}\widetilde{\mathbf{W}}\,Diag(\sqrt{\mathbf{z}})$, which is symmetric. Note that $|\mathbf{I} - \rho\mathbf{W}|$ is required, and hence

$$|\mathbf{I} - \rho\mathbf{W}_{sim}| = \begin{cases} (\rho)^n \left|\frac{1}{\rho}\mathbf{I} + (-\mathbf{W}_{sim})\right|; & \rho > 0, \\ (-\rho)^n \left|\left(-\frac{1}{\rho}\right)\mathbf{I} + \mathbf{W}_{sim}\right|; & \rho < 0, \end{cases}$$

needs to be computed. The Cholesky decomposition of $\mathbf{W}_{sim}$ needed for the determinant can be easily updated when $c\mathbf{I}$, $c > 0$, is added to a matrix (here $-\mathbf{W}_{sim}$ for $\rho > 0$ and $\mathbf{W}_{sim}$ for $\rho < 0$), where c is a constant. This updating is implemented effectively in the package `spdep` [14] using the package `Matrix` [13, 15].

We now derive the EM algorithm for the SEM and the SAM. Recall that the E-step of the EM algorithm requires the calculation of $E_{u|s,\boldsymbol{\theta}'}(\log f(\mathbf{y};\boldsymbol{\theta}))$, where $\log f(\mathbf{y};\boldsymbol{\theta})$ for this problem is given by (5). The first three terms of (5)

do not depend on $\mathbf{y}_u$, and therefore in the E-step we are mostly concerned with calculating the term $E_{u|s;\boldsymbol{\theta}'}(\mathbf{r}^T\mathbf{M}\mathbf{r})$. For notational convenience, from here on we omit the dependence of the expectation operator on $\boldsymbol{\theta}'$.

Partition the vector $\boldsymbol{\mu}$ and the matrices $\mathbf{V}$ and $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\mu} \equiv \begin{pmatrix} \boldsymbol{\mu}_s \\ \boldsymbol{\mu}_u \end{pmatrix}, \mathbf{V} \equiv \begin{pmatrix} \mathbf{V}_{ss} & \mathbf{V}_{su} \\ \mathbf{V}_{us} & \mathbf{V}_{uu}. \end{pmatrix}, \boldsymbol{\Sigma} \equiv \begin{pmatrix} \boldsymbol{\Sigma}_{ss} & \boldsymbol{\Sigma}_{su} \\ \boldsymbol{\Sigma}_{us} & \boldsymbol{\Sigma}_{uu}. \end{pmatrix}. \tag{6}$$

It is widely known [16] that the conditional distribution of $\mathbf{y}_u|\mathbf{y}_s$ is multivariate normal with mean

$$\begin{aligned} E_{u|s}(\mathbf{y}_u) \equiv \boldsymbol{\mu}_{u|s} &= \boldsymbol{\mu}_u + \boldsymbol{\Sigma}_{us}\boldsymbol{\Sigma}_{ss}^{-1}(\mathbf{y}_s - \boldsymbol{\mu}_s) \\ &= \boldsymbol{\mu}_u + \mathbf{V}_{us}\mathbf{V}_{ss}^{-1}(\mathbf{y}_s - \boldsymbol{\mu}_s), \end{aligned} \tag{7}$$

and covariance matrix

$$\begin{aligned} \mathrm{Cov}_{u|s}(\mathbf{y}_u) \equiv \boldsymbol{\Sigma}_{u|s} &= \boldsymbol{\Sigma}_{uu} - \boldsymbol{\Sigma}_{us}\boldsymbol{\Sigma}_{ss}^{-1}\boldsymbol{\Sigma}_{su} \\ &= \omega\left\{\mathbf{V}_{uu} - \mathbf{V}_{us}\mathbf{V}_{ss}^{-1}\mathbf{V}_{su}\right\}. \end{aligned} \tag{8}$$

Now re-write $E_{u|s}(\mathbf{r}^T\mathbf{M}\mathbf{r})$ as

$$\begin{aligned} E_{u|s}(\mathbf{r}^T\mathbf{M}\mathbf{r}) &= E_{u|s}\left((\mathbf{y} - \boldsymbol{\mu})^T\mathbf{M}(\mathbf{y} - \boldsymbol{\mu})\right) \\ &= E_{u|s}(\mathbf{y}^T\mathbf{M}\mathbf{y}) - 2E_{u|s}(\mathbf{y})^T\mathbf{M}\boldsymbol{\mu} + \boldsymbol{\mu}^T\mathbf{M}\boldsymbol{\mu}, \end{aligned} \tag{9}$$

where $E_{u|s}(\mathbf{y}) = (\mathbf{y}_s^T, \boldsymbol{\mu}_{u|s}^T)^T$ and $\boldsymbol{\mu}_{u|s} \equiv E_{u|s}(\mathbf{y}_u)$. The second term of the right-hand side of (9) is straightforward to find, while the first term is

$$E_{u|s}(\mathbf{y}^T\mathbf{M}\mathbf{y}) = \mathbf{y}_s^T\mathbf{M}_{ss}\mathbf{y}_s + 2\mathbf{y}_s^T\mathbf{M}_{su}E_{u|s}(\mathbf{y}_u) + E_{u|s}(\mathbf{y}_u^T\mathbf{M}_{uu}\mathbf{y}_u). \tag{10}$$

Computing second-order expectations in $E_{u|s}(\mathbf{y}^T\mathbf{M}\mathbf{y})$ hence reduces to computing $E_{u|s}(\mathbf{y}_u^T\mathbf{M}_{uu}\mathbf{y}_u)$, which can be re-written as

$$\begin{aligned} E_{u|s}(\mathbf{y}_u^T\mathbf{M}_{uu}\mathbf{y}_u) &= \mathrm{tr}\left\{E_{u|s}\left(\mathbf{y}_u\mathbf{y}_u^T\right)\mathbf{M}_{uu}\right\} \\ &= \mathrm{tr}\left\{\left(\boldsymbol{\mu}_{u|s}\boldsymbol{\mu}_{u|s}^T + \boldsymbol{\Sigma}_{u|s}\right)\mathbf{M}_{uu}\right\} \\ &= \boldsymbol{\mu}_{u|s}^T\mathbf{M}_{uu}\boldsymbol{\mu}_{u|s} + \mathrm{tr}\left\{\boldsymbol{\Sigma}_{u|s}\mathbf{M}_{uu}\right\}, \end{aligned}$$

where $\mathrm{tr}\{\cdot\}$ denotes the trace operator. The second equality follows from $\mathrm{Var}(\mathbf{y}) =$

$E(\mathbf{y}\mathbf{y}^T) - E(\mathbf{y})E(\mathbf{y})^T$ which holds for any random vector. Now we apply the formula for the inverse of a partitioned matrix [17, Corollary 8.5.12] and obtain

$$\mathbf{M}_{uu} \equiv \left(\mathbf{V}^{-1}\right)_{uu} = \left(\mathbf{V}_{uu} - \mathbf{V}_{us}\mathbf{V}_{ss}^{-1}\mathbf{V}_{su}\right)^{-1}.$$

From (8), we see that $\mathbf{M}_{uu}^{-1}$ is proportional to $\mathbf{\Sigma}_{u|s}$ and hence

$$\begin{aligned}\mathbf{\Sigma}_{u|s} &= \omega\left(\mathbf{V}_{uu} - \mathbf{V}_{us}\mathbf{V}_{ss}^{-1}\mathbf{V}_{su}\right)\\ &= \omega\mathbf{M}_{uu}^{-1}.\end{aligned}$$

It is important to note that, while the conditional expectation is with respect to (and hence $\mathbf{\Sigma}_{u|s}$ is a function of) the current parameter estimates, say ρ' and ω', $\mathbf{M}$ is a function of the parameters to be optimised over, namely ρ. Writing out the dependencies explicitly, we obtain

$$\mathrm{tr}\left\{\mathbf{\Sigma}_{u|s}(\rho',\omega')\mathbf{M}_{uu}(\rho)\right\} = \omega'\mathrm{tr}\left\{\mathbf{M}_{uu}(\rho')^{-1}\mathbf{M}_{uu}(\rho)\right\}.$$

In summary we can re-write the term $E_{u|s}(\mathbf{r}^T\mathbf{M}\mathbf{r})$ in (9) as

$$E_{u|s}(\mathbf{r}^T\mathbf{M}\mathbf{r}) = \mathbf{r}_{u|s}^T\mathbf{M}\mathbf{r}_{u|s} + \omega'\mathrm{tr}\{\mathbf{M}_{uu}(\rho')^{-1}\mathbf{M}_{uu}(\rho)\}, \tag{11}$$

with $\mathbf{r}_{u|s} \equiv E_{u|s}(\mathbf{y}) - \boldsymbol{\mu}$. Finally, the expected complete-data log-likelihood $E_{u|s}\log f(\mathbf{y};\boldsymbol{\theta})$ can be written as

$$-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\omega + \frac{1}{2}\log|\mathbf{M}| - \frac{\mathbf{r}_{u|s}^T\mathbf{M}\mathbf{r}_{u|s} + \omega'\mathrm{tr}\left\{\mathbf{M}_{uu}(\rho')^{-1}\mathbf{M}_{uu}(\rho)\right\}}{2\omega}. \tag{12}$$

In each M-step, $Q(\boldsymbol{\theta}|\boldsymbol{\theta}') \equiv E_{u|s}\log f(\mathbf{y};\boldsymbol{\theta})$ has to be maximised with respect to $\boldsymbol{\theta}$, i.e. ρ, $\boldsymbol{\beta}$ and ω. To ease calculations, at each M-step we use the concentrated log-likelihood by replacing $\boldsymbol{\beta}$ and ω by the maximisers, $\hat{\omega}$ and $\hat{\boldsymbol{\beta}}$, so that the remaining expression only depends on ρ. These maximisers are found by differentiating the expected log-likelihood and setting equal to zero. Solving gives

$$\begin{aligned}\hat{\omega} &= \frac{\mathbf{r}_{u|s}^T\mathbf{M}\mathbf{r}_{u|s} + \omega'\mathrm{tr}\left\{\mathbf{M}_{uu}(\rho')^{-1}\mathbf{M}_{uu}(\rho)\right\}}{n},\\ \hat{\boldsymbol{\beta}} &= \left(\widetilde{\mathbf{X}}^T\mathbf{M}\widetilde{\mathbf{X}}\right)^{-1}\widetilde{\mathbf{X}}^T\mathbf{M}E_{u|s}(\mathbf{y}),\end{aligned} \tag{13}$$

where, for the SAM $\widetilde{\mathbf{X}} = \mathbf{A}^{-1}\mathbf{X}$, while for the SEM $\widetilde{\mathbf{X}} = \mathbf{X}$.

2.3 Pseudo Restricted Maximum Likelihood

The ML method does not take into account the degrees of freedom that are involved in estimating the fixed effects, leading (generally) to biased variance components. Restricted maximum likelihood (REML) accounts for these degrees of freedom by applying the ML method to a linear transformation $\mathbf{Ky}$, where $\mathbf{K}$ is chosen such that $\mathbf{KX} = \mathbf{0}$ and $\text{rank}(\mathbf{K}) = n - p$, with p the number of columns of the design matrix $\mathbf{X}$. A requirement for REML is that $\mathbf{K}$ does not depend on the parameters governing the mean, and one such choice for $\mathbf{K}$ is the projection matrix $\mathbf{P} \equiv \mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ [18, 19].

For the SEM, $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ and $\mathbf{K} = \mathbf{P}$. For the SAM, since $\boldsymbol{\mu} = (\mathbf{A}(\rho)^{-1}\mathbf{X})\boldsymbol{\beta}$, a natural choice would be to use $\mathbf{P}$ with $\mathbf{X}$ replaced by $\widetilde{\mathbf{X}} = \mathbf{A}^{-1}\mathbf{X}$. However the distribution of the transformation depends on the 'fixed effect' parameter ρ governing the mean and the (co)variance. Since this violates the above requirement, REML cannot be applied to the SAM without introducing errors. We term a variant of REML that ignores this violation as 'Pseudo-REML' (P-REML), and we use this for the SAM. Such a 'Pseudo-REML' procedure has been applied by Samart and Chambers [20] for fitting linear models with nested errors for probability linked data. There, the authors demonstrated that the 'Pseudo-REML' method still outperforms standard ML, that is, it leads to improved estimation of variance parameters.

A computationally simple method to obtain REML estimates for complete data is by adding $\log c(\boldsymbol{\theta}) = -\frac{1}{2}\log|\widetilde{\mathbf{X}}^T\boldsymbol{\Sigma}^{-1}\widetilde{\mathbf{X}}|$ to the complete-data log-likelihood $\log f(\mathbf{y};\boldsymbol{\theta})$, where

$$\log c(\boldsymbol{\theta}) = -\frac{1}{2}\log|\widetilde{\mathbf{X}}^T\boldsymbol{\Sigma}^{-1}\widetilde{\mathbf{X}}| = -\frac{1}{2}\log|\widetilde{\mathbf{X}}^T\mathbf{M}\widetilde{\mathbf{X}}| + \frac{p}{2}\log\omega,$$

and maximising the resulting expression $\log g(\mathbf{y};\boldsymbol{\theta}) \equiv \log f(\mathbf{y};\boldsymbol{\theta}) + \log c(\boldsymbol{\theta})$ [21]:

$$\log g(\mathbf{y};\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{n-p}{2}\log\omega - \frac{1}{2}\log|\widetilde{\mathbf{X}}^T\mathbf{M}\widetilde{\mathbf{X}}| + \frac{1}{2}\log|\mathbf{M}| - \frac{1}{2\omega}\mathbf{r}^T\mathbf{Mr}. \tag{14}$$

Maximising this expression yields REML estimates in the complete data case. As we are interested in the incomplete data case, we aim at maximising the marginal function

$$g(\mathbf{y}_s;\boldsymbol{\theta}) = \int g(\mathbf{y};\boldsymbol{\theta})\mathrm{d}\mathbf{y}_u,$$

which is obtained by integrating out the unobserved $\mathbf{y}_u$. To achieve this objective, we add the constant $\log c(\boldsymbol{\theta})$ to $Q(\boldsymbol{\theta}|\boldsymbol{\theta}')$, see (12), and maximise the resulting expression in the M-step. Notice that for a given $\boldsymbol{\theta}$, the E-step is the same as in the standard ML-EM case since $\log c(\boldsymbol{\theta})$ is independent of $\mathbf{y}_u$. In the Appendix B.2 we show that this method yields estimates that maximise $\log g(\mathbf{y}_s;\boldsymbol{\theta})$, or equivalently, $g(\mathbf{y}_s;\boldsymbol{\theta})$. This proposed computationally simple method is different from the approach by Diffey et al. [22] for linear mixed models, where the random effects (unobserved) and response variable (observed) are independent.

2.4 Observed Information

To obtain standard errors, the calculation of the observed information $-\frac{\partial^2 f(\mathbf{y}_s)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}$ is required. This is often complicated and instead the observed information is calculated indirectly by applying the missing information principle [23]:

$$\texttt{observed information} = \texttt{complete information} - \texttt{missing information},$$

or, in formulae,

$$-\frac{\partial f(\mathbf{y}_s;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} = E_{u|s}\left(-\frac{\partial f(\mathbf{y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right) - E_{u|s}\left(-\frac{\partial f(\mathbf{y}_u|\mathbf{y}_s;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right). \tag{15}$$

Appendix C gives explicit expressions for the quantities required to compute the observed information for the two models we analyse.

3 Computation and approximations

The approach in LP04 considers an iterative algorithm very similar to the EM algorithm described above but with some important differences that render their E-step and M-step only approximate. In this section we place their algorithm within the EM framework and also propose other approximations that could also be used in principle to achieve a similar computational benefit.

3.1 Approximating the Conditional Mean (E-Step)

The calculation of the conditional mean $\boldsymbol{\mu}_{u|s} \equiv E_{u|s}(\mathbf{y}_u)$ can be costly because of the matrix inverses $\mathbf{V} = \mathbf{M}^{-1}$ and $\mathbf{V}_{ss}^{-1}$, see (7). To avoid these costly matrix manipulations for $\boldsymbol{\mu}_{u|s}$, LP04 suggested expressing the distribution of $\mathbf{y}_u|\mathbf{y}_s$ as

a product of univariate normals, often called the pseudo or composite likelihood [24], through

$$\mathbf{C} = \mathbf{I}_n - [Diag(\mathbf{M})]^{-1}\mathbf{M}, \tag{16}$$

$$E(\mathbf{y}_u|\mathbf{y}_s) = \boldsymbol{\mu}_u + \mathbf{C}_{us}(\mathbf{y}_s - \boldsymbol{\mu}_s), \tag{17}$$

where $Diag(\mathbf{M})$ is the diagonal of $\mathbf{M}$. This formula is based on formula (A.2) provided in Gelman et al. [25, p. 579],

$$(y_i|y_j \text{ all } j \neq i) \sim N\left(\mu_i + \sum_{j\neq i} C_{ij}(y_j - \mu_j), 1/[(\boldsymbol{\Sigma})^{-1}]_{ii}\right), \tag{18}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of all $\mathbf{y} \equiv (y_1, y_2, \ldots, y_n)^T$ and C_{ij} are the elements of $\mathbf{C}$.

Equation (17) with $\mathbf{C}$ specified in (16) only holds when we are interested in the conditional distribution of y_i given all the other $n-1$ variables $y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n$ for all i that are contained in the subset u. However above we require $\mathbf{y}_u|\mathbf{y}_s$ or equivalently $\mathbf{y}_i|\mathbf{y}_s$ for all $i \in u$, where we condition on $n_s \leq n-1$ variables. Unless $n_s = n-1$, the formula (16) does not hold in general. Despite this, we see that (17) may still be useful as it serves potentially as an computationally fast approximation of the conditional mean. We denote this approximation as *LP*.

For computing the conditional mean (19) exactly, [26] suggest the computationally simpler expression (see also [27, Section A.2])

$$\boldsymbol{\mu}_{u|s} = \boldsymbol{\mu}_u - (\boldsymbol{\Sigma}^{-1})^{-1}_{uu}(\boldsymbol{\Sigma}^{-1})_{us}(\mathbf{y}_s - \boldsymbol{\mu}_s) \tag{19}$$

$$= \boldsymbol{\mu}_u - \mathbf{M}^{-1}_{uu}\mathbf{M}_{us}(\mathbf{y}_s - \boldsymbol{\mu}_s). \tag{20}$$

Note that although this expression only requires blocks of $\mathbf{M}$, which are known and sparse for the SAM and SEM, it still requires the calculation of the inverse of $\mathbf{M}_{uu}$. This is unproblematic for small n_u (small number of missing units), but still computational costly for large n_u, say $n_u > 2,000$. In practice, the inverse is solved using backward-forward solves. We denote this method as *exact*.

Until now, we have only considered the term in (20) that contains the matrix $\mathbf{M}_{uu}$. However, for the SAM, the calculation of $\boldsymbol{\mu}$ can also be costly for large

n, since $\boldsymbol{\mu} = \mathbf{A}^{-1}\mathbf{X}\boldsymbol{\beta}$. LP04 suggested using the expansion

$$\mathbf{A}^{-1}\mathbf{X}\boldsymbol{\beta} \approx \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{W}\mathbf{X}\boldsymbol{\beta} + \rho^2\mathbf{W}^2\mathbf{X}\boldsymbol{\beta} + \dots \ , \tag{21}$$

and to compute the $n \times p$ matrices $\{\mathbf{W}^j\mathbf{X}; j = 1, 2, \dots\}$ prior to commencing the EM algorithm. However, our experience is that solving the left-hand-side of (21) poses no difficulty since $\mathbf{X}$ is not an $n \times n$ matrix but one of size $n \times p$, usually with $p \ll n$. We therefore compute $\boldsymbol{\mu} = \mathbf{A}^{-1}\mathbf{X}\boldsymbol{\beta}$ exactly by solving $\mathbf{A}\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ for $\boldsymbol{\mu}$ for the SAM without calculating $\mathbf{A}^{-1}$.

3.2 Approximating the trace term in $\hat{\omega}$ (M-step)

The calculation of the term $\operatorname{tr}\left\{\mathbf{M}_{uu}(\rho')^{-1}\mathbf{M}_{uu}(\rho)\right\}$ affecting ω can be costly because of the number of solves to be carried out when n_u is large. As a result, it is natural to consider approximations to simplify this computation; after all it is well known that as long as the expected log-likelihood $Q(\boldsymbol{\theta}|\boldsymbol{\theta}')$ increases in the M-step, the (now generalised) EM algorithm still guarantees an improvement in the marginal log-likelihood $L_s \equiv \log f(\mathbf{y}_s|\boldsymbol{\theta})$ in each step, that is, convergence to the ML estimates [12, Chapter 3]. Unfortunately, computing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}')$ (exactly) results in the same computational bottleneck. When running an EM algorithm to convergence with an approximate M-step the hope is that the unknown $Q(\boldsymbol{\theta}|\boldsymbol{\theta}')$ increases within each M-step, although this, obviously, will not always be the case.

When the E-step is computed exactly, the iterative algorithm of LP04, run until convergence, is identical to the EM algorithm as described here when:

- (*LP-1*) replacing $\mathbf{y}_u$ with $E_{u|s}(\mathbf{y}_u)$ (LP04, p. 236) and using complete-data SAM/SEM algorithms for the M-step. This is equivalent to maximising (12) without the trace term, and
- (*LP-2*) letting $\hat{\omega} = (\mathbf{A}\mathbf{r})_s^T(\mathbf{A}\mathbf{r})_s/n_s$ (LP04 p. 240, see Appendix A for a comparison with their notation) for the SAM when using the concentrated log-likelihood.

Although LP04 implements both *LP-1* and *LP-2* simultaneously, we separated them to study the effects of these approximations individually. We denote the simultaneous application of these approximations as *LP-3*. Evidently, LP04's algorithm is an EM algorithm with the M-step being approximated at each iteration. As we shall see in the simulation section, this can result in unpredictable

estimates on convergence when the approximation is poor, namely when n_u is large relative to n_s.

As such there are several other approximations that can be considered, some of which might be more suitable for the case when n_u is large. The approximations we consider are:

- (*equal* ρ) using $\hat{\omega} = \frac{\mathbf{r}_{u|s}^T \mathbf{M} \mathbf{r}_{u|s} + \omega'(n - n_s)}{n}$ based on $\rho = \rho'$,
- (*equal* ρ *&* ω) letting $\hat{\omega} = \mathbf{r}_{u|s}^T \mathbf{M} \mathbf{r}_{u|s} / n_s$ based on $\rho = \rho'$ and $\omega = \omega'$,
- (*diag*) using $Diag(\mathbf{M}_{uu}(\rho))/\, Diag(\mathbf{M}_{uu}(\rho'))$ instead of $\mathbf{M}_{uu}^{-1}(\rho')\mathbf{M}_{uu}(\rho)$, where $Diag(\mathbf{M})$ returns the diagonal of matrix $\mathbf{M}$, and
- (*Taylor 1st/6th*) using a 1st and 6th order Taylor series approximation of $\mathbf{M}_{uu}^{-1}$.

For the Taylor series approximations we proceed as follows. We first write out $\mathbf{M}$ explicitly as

$$\begin{aligned}\mathbf{M} &= \mathbf{A}^T\mathbf{A} \\ &= \mathbf{I} - \rho(\mathbf{W} + \mathbf{W}^T) + \rho^2\mathbf{W}^T\mathbf{W} \\ &= \mathbf{I} - \rho\mathbf{C} + \rho^2\mathbf{D},\end{aligned}$$

where $\mathbf{C} = \mathbf{W} + \mathbf{W}^T$ and $\mathbf{D} = \mathbf{W}^T\mathbf{W}$. Hence

$$\mathbf{M}_{uu} = \mathbf{I}_{n_u} - \rho\mathbf{C}_{uu} + \rho^2\mathbf{D}_{uu}.$$

Now, if the inverse of a matrix $\mathbf{B}$ exists, then the inverse can be expressed as a Neumann series [28],

$$\mathbf{B}^{-1} = \sum_{j=0}^{\infty} (\mathbf{I}_n - \mathbf{B})^j. \tag{22}$$

We employ (22) but limit the series up to the 6th order, which should be a suitable compromise between accuracy and feasibility in several applications. The resulting approximation is

$$\mathbf{M}_{uu}^{-1} \approx \mathbf{I}_{n_u} + \sum_{k=1}^{12} \rho^k \widetilde{\mathbf{M}}_{uu,k}.$$

The matrices $\widetilde{\mathbf{M}}_{uu,k}$ can be calculated prior to the EM algorithm making the calculation of the approximate $\mathbf{M}_{uu}^{-1}$ very fast since matrix multiplications and

additions require less operations than matrix linear solves or inverses.

Finally, we also consider the exact M-step involving the direct calculation of $\mathbf{M}_{uu}(\rho')^{-1}\mathbf{M}_{uu}(\rho)$. We denote this as *exact*.

4 Simulation Study

The aim of this simulation study is to show and evaluate the performance of the various approximations to the EM algorithm of Section 2 that are discussed in Section 3. Throughout, we use the contiguity matrix $\mathbf{W}$, available from the Corrected Boston Housing Data (CBHD) provided with the `R` package `spdep` [14]. These data were collected by Harrison and Rubinfeld [29] and contain $n = 506$ units. For the simulation study, we discarded the responses and covariates of the CBHD and used solely the spatial locations; the data were simulated at these for the SAM and the SEM with intercept $\beta_0 = 1$, slope $\beta_1 = 2$, spatial dependence parameter $\rho = 0.5$ and variance $\omega = 1$. The covariate X_i for unit i was sampled from a standard normal distribution for the SAM and from $N(0, V_{ii})$ for the SEM in order for the models to imply identical explained variance at each unit $R^2 = 1 - \mathrm{Var}(Y_i|\mathbf{X})/\mathrm{Var}(Y_i)$. It can be shown that $R^2 = 0.8$ under these settings for both the SEM and the SAM.

Table 1 shows the empirical mean of $\hat{\rho}$ and $\hat{\omega}$ over $10,000$ simulated datasets for the SEM with $n_s = 400$. In this case, where the majority of the units are observed, all methods, except LP-1, perform well and provide similar estimates from a practical point of view. LP-1 produces (biased) under-estimates for the variance parameter, ω, likely due to an over-confidence resulting from treating $\mathbf{y}_u$ as an observed quantity with value $E_{u|s}(\mathbf{y}_u)$, see Section 3.2. LP-3 reduces the bias for ω somewhat, but still performs less favourably than the other methods. Of all the trace approximations, we find that the methods *equal* ρ, *diag*, *Taylor 1st* and *Taylor 6th* provide good estimates in this study. The approximation to the conditional mean, *LP*, does not alter the results substantially. As expected, *P-REML* provides estimates that are slightly less biased than those obtained using maximum likelihood.

Table 2 shows the same as Table 1 but for $n_s = 100$. This table reveals the problems with approximating the steps in the EM algorithm when the number of observed units is low compared to the total number of units. The *exact* method shows slightly biased estimates: an empirical mean of $\hat{\rho}$ of 0.458 versus $\rho = 0.500$ and an empirical mean of $\hat{\omega}$ of 0.970 versus $\omega = 1.000$ (*P-REML* again

Table 1: Empirical mean of ρ and ω estimates over 10,000 simulations for the SEM model with $n = 506$, $n_s = 400$ and the parameters $\rho = 0.5$, $\omega = 1$, $\beta_0 = 1$, $\beta_1 = 2$

		Empirical mean of $\hat{\rho}$		Empirical mean of $\hat{\omega}$	
		conditional mean		conditional mean	
		exact	LP	exact	LP
trace method	exact	0.493	0.481	0.995	1.002
	equal ρ & ω	0.569	0.556	0.963	0.972
	equal ρ	0.517	0.505	0.984	0.992
	diag	0.481	0.470	1.000	1.007
	Taylor 6th	0.492	0.481	0.996	1.003
	Taylor 1st	0.498	0.486	0.978	0.986
	LP-1	0.569	0.556	0.761	0.768
	LP-2	0.529	0.525	0.964	0.971
	LP-3	0.569	0.556	0.947	0.958
	P-REML	0.496	0.485	0.998	1.006

has a slightly lower bias). This bias results from the relatively small $n_s = 100$, which leads to negative estimates of $\hat{\rho}$ for some simulation runs. The impact of the approximations used for low n_s highlights the need for care when operating in this regime, particularly when employing 'plug-in' approximations such as *equal ρ & ω* or the *LP* methods. Also, in this regime, the approximate E-step further results in estimates that are considerably different from those without the approximation.

The use of exact M-steps guarantees that the marginal log-likelihood L_s increases in each iteration. However the use of the approximate M-steps can in fact lead to decreasing values of L_s. We illustrate such a consequence in Figure 1, where we show the trajectory of L_s when estimating the parameters from a random dataset generated from a SEM using both the *exact* method and the approximate methods. Note how monotonicity in L_s is present for the *exact* method when both $n_s = 100$ (left panel) and $n_s = 400$ (right panel). When $n_s = 400$, all approximate methods improve L_s, and all methods converge to similar values. For $n_s = 100$ though, all methods (except *Taylor 6th* in this case) cause unpredictable behaviour in L_s, and thus the unreliable estimates observed so far.

Table 3 shows the coverage of the 95% Wald-type confidence intervals (CI) for β_0 and β_1 ($n_s = 100$) while Table 4 shows those for ρ and ω ($n_s = 400$),

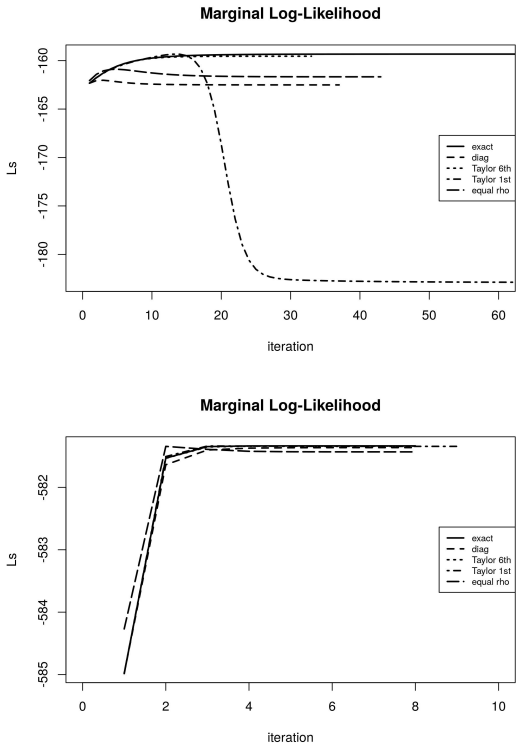

Figure 1: Values of the marginal log-likelihood L_s for various methods for the SEM and $n = 506$ with samples of sizes $n_s = 100$ (left panel) and $n_s = 400$ (right panel) with $X \sim N(0, 1)$.

Table 2: Empirical mean of ρ and ω estimates over 10,000 simulations for the SEM model with $n = 506$, $n_s = 100$ and the parameters $\rho = 0.5$, $\omega = 1$, $\beta_0 = 1$, $\beta_1 = 2$

		Empirical mean of $\hat{\rho}$		Empirical mean of $\hat{\omega}$	
		conditional mean		conditional mean	
		exact	LP	exact	LP
trace method	exact	0.447	0.258	0.977	1.156
	equal ρ & ω	0.882	0.699	0.473	0.905
	equal ρ	0.0995	0.0904	1.236	1.241
	diag	0.0351	0.0347	1.260	1.260
	Taylor 6th	0.370	0.252	1.092	1.161
	Taylor 1st	0.796	0.327	0.214	0.845
	LP-1	0.882	0.699	0.0934	0.179
	LP-2	0.762	0.666	0.508	0.930
	LP-3	0.882	0.699	0.391	0.822
	P-REML	0.478	0.281	0.983	1.172

that is, they show the proportion of simulations in which the true parameter θ lies inside the confidence interval $\hat{\theta} \pm 1.96\sqrt{\widehat{\text{Var}}(\hat{\theta})}$. While for β_0 and β_1 such Wald-type CI's tend to be appropriate, this is not necessarily the case for ρ and ω since for small n_s these parameter estimates cannot be assumed to be normally distributed. For example it is well known for simple linear regression that confidence intervals for $\hat{\omega} = \hat{\sigma}^2$ are based on a chi-square distribution of the sum of squares divided by σ^2 rather than on a normal distribution. The results show that the coverages based on the *exact* method (ML) and *P-REML* for the β parameters, usually of primary interest, are near the anticipated 95%, whereas most of the other approximations do not perform satisfactorily.

In Appendix D we show the tables for the same analysis but with the SAM instead of the SEM. We observe that all methods perform slightly better for the SAM than for the SEM. This is a consequence of the mean and not only the variance depending on ρ in the SAM, leading to more precise estimation of the parameter ρ. The conclusions are similar to those with the SEM: For large $n_s = 400$ (relative to $n = 506$) all methods perform relatively well, however for low sampling fractions, that is, for n_s/n small, ($n_s = 100$, $n = 506$) the detrimental effect of using various approximations within the E- and M-steps becomes apparent. Only for large sampling fractions (say $n_s/n \geq 40\%$), some approximations appear to be reliable, for example *equal* ρ and *Taylor 6th*.

Table 3: Coverage of Wald-type confidence intervals for β_0 and β_1 estimates over 10,000 simulations for the SEM model with $n = 506$, $n_s = 100$ and the parameters settings $\rho = 0.5$, $\omega = 1$, $\beta_0 = 1$, $\beta_1 = 2$

		Coverage for β_0		Coverage for β_1	
		conditional mean		conditional mean	
		exact	LP	exact	LP
trace method	exact	0.939	0.916	0.934	0.945
	equal ρ & ω	0.989	0.985	0.846	0.945
	equal ρ	0.903	0.902	0.945	0.945
	diag	0.899	0.899	0.944	0.944
	Taylor 6th	0.932	0.916	0.945	0.945
	Taylor 1st	0.958	0.873	0.570	0.895
	LP-1	0.975	0.737	0.467	0.597
	LP-2	0.981	0.972	0.861	0.932
	LP-3	0.988	0.983	0.803	0.932
	P-REML	0.942	0.922	0.936	0.946

To further investigate the relationship between performance and sampling ratio, we calculated the value of the log-likelihood ratio statistic $LLR = 2 \times \log f(\mathbf{y}_s|\hat{\boldsymbol{\theta}}) - 2 \times \log f(\mathbf{y}_s|\hat{\boldsymbol{\theta}}_{ML})$ of each method with estimate $\hat{\boldsymbol{\theta}}$ relative to the exact ML method with estimate $\hat{\boldsymbol{\theta}}_{ML}$. A large value (say, greater than 1) indicates that the approximation does not perform well, while a value of 0 indicates identical performance. Figure 2 shows the 90th percentiles over 10,000 simulations of the values of $\log_2(LLR + 1)$, where the dashed line indicates the value for which $LLR = 1$. We treat methods where the 90th percentile of $LLR > 1$ (and hence the 90th percentile of $\log_2(LLR + 1) > 1$) as unreliable: Such cases occur when less than 90% of the log-likelihood ratios obtained are less than one. For $n_s = 500$ all methods perform well; for $n_s = 400$ most methods work well, except *LP-1* (for both the SAM and the SEM), and *equal* ρ & ω and *LP-2* for the SAM. For $n_s \in \{200, 300\}$ the two methods *equal* ρ and *Taylor 6th* always perform well. For $n_s \in \{50, 100\}$ all approximate methods appear to perform unsatisfactorily.

5 Case Study

The Lucas County (Ohio, USA) housing data consist of $n = 25,357$ observations of single-family homes sold in the period 1993–1998; a full description of

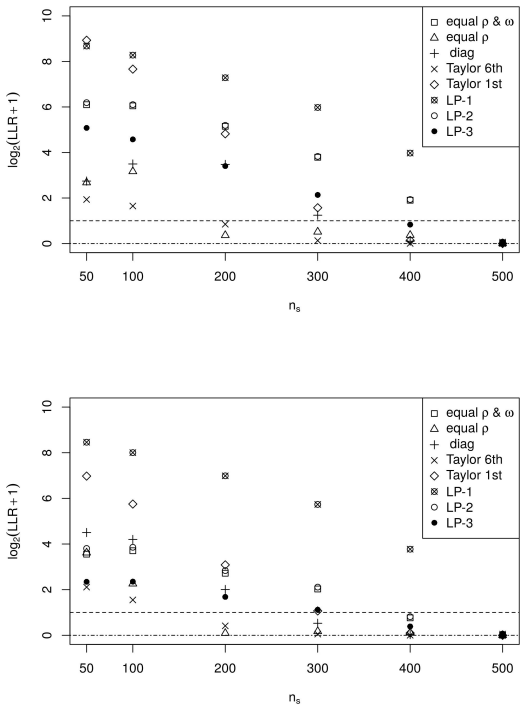

Figure 2: Value of the 90th percentile of $log_2(LLR + 1)$ versus n_s with the approximate ML methods: for the SEM with $X \sim N(0, V_{ii})$ (left panel) and SAM with $X \sim N(0, 1)$ (right panel). The dashed lines denotes $log_2(LLR+1) = 1$, while the dotted-dashed lines denotes $log_2(LLR + 1) = 0$. When the value is above the dashed line, the associated approximation method for the given n_s is considered unreliable.

Table 4: Coverage of Wald-type confidence intervals for ρ and ω over 10,000 simulations for the SEM model with $n = 506$, $n_s = 400$ and the parameters settings $\rho = 0.5$, $\omega = 1$, $\beta_0 = 1$, $\beta_1 = 2$

		Coverage for ρ		Coverage for ω	
		conditional mean		conditional mean	
		exact	LP	exact	LP
trace method	exact	0.915	0.918	0.938	0.947
	equal ρ & ω	0.578	0.685	0.877	0.901
	equal ρ	0.879	0.915	0.923	0.936
	diag	0.907	0.901	0.943	0.949
	Taylor 6th	0.918	0.921	0.942	0.948
	Taylor 1st	0.937	0.949	0.918	0.931
	LP-1	0.623	0.734	0.0311	0.0395
	LP-2	0.830	0.846	0.877	0.894
	LP-3	0.623	0.734	0.849	0.882
	P-REML	0.914	0.924	0.942	0.948

these data is available in the Spatial Econometrics toolbox (SET) for Matlab, see `http://www.spatial-econometrics.com/html/jplv7.zip`. The dataset is part of the `R` package `spdep` [14] and has been used by Bivand [30] to compare with one another several software packages designed to fit spatial regression models.

Bivand [30] used the logarithm of the house price ($\log(price)$) as the response variable. The chosen explanatory variables included powers of house age (age, age^2 and age^3), the logarithm of the lot size in square feet ($\log(lotsize)$), the number of rooms ($rooms$), the logarithm of the total living area in square feet (LTA), the number of $beds$ and an indicator for each of the years 1993–1998 ($syear$) when the house was sold. We use the same row-normalised sparse contiguity matrix $\mathbf{W}$ used by Bivand [30], while for fitting the various models we use a single core of type Intel Xeon E5-2620 v2 - 2.10GHz. Fitting the model for the full (observed) dataset takes around 5 seconds. In order to assess the algorithms with missing data, we set $s = \{1, 6, 11, \ldots, 25356\}$, which contains $n_s = 5,072$ elements. Systematic sampling was used for ease of reproducibility across different software.

We fitted the SAM to the data using the range of the methods considered in the simulation study. For this study, convergence was deemed to be reached when the sum of the absolute deviations from the vector of old and new esti-

mates was less than 10^{-6}. The number of iterations reduces by approximately 50% when the tolerance is set to 10^{-4}. Estimates of ρ and ω, the value of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}')$, either exact or approximate corresponding to the method, the observed (marginal) log-likelihood L_s, the number of EM-steps required, and the time in seconds needed for the EM algorithm to converge, are shown in Table 5. The first column of Table 5 lists the approximation method; a combination of the method used for the conditional mean (E-step), followed by the method applied. For comparison purposes, we also list the results of the Matlab code (denoted *Matlab LP*) kindly provided by James LeSage, that implements *LP-3* but uses the SET in Matlab to fit the SAM in the M-step.

As expected from the results in Section 4, for the chosen small ratio $n_s/n \approx 20\%$, the estimates for ρ and ω are inconsistent across the different approximations adopted. It is apparent from the marginal log-likelihood L_s that the only reliable methods in this case are those that do not approximate the E- and M-steps. The fact that the best-performing approximation method is the *LP/LP-2* method, one which approximates both the E- and the M-steps, highlights how unpredictable such approximations can be. The methods were more consistent relative to each other with larger sampling ratios.

Matlab LP yielded approximately $\hat{\rho} = 0.7830$ after only 22 EM-steps in approximately 27 seconds. This lies in contrast to the final estimate of $\hat{\rho} = 0.8841$ for *exact/LP-3* using `R` after 60 EM-steps in 1,550 seconds. The difference between the Matlab implementation and the `R` implementation arises from the way in which the SET carries out parameter estimation, which yields less accurate estimates but with considerable computational gain (The SET took approximately 1 second per M-step, compared to the approximately 26 seconds required for the `R` package `spdep`). This inaccuracy was also observed by Bivand [30, see Table 4] in the case of a fully observed data set. The resulting parameter estimates show that the approximate behaviour of the SET toolbox, deemed to be inconsequential in a full-data scenario, has a large impact on the final EM estimates, even if the differences in the estimates at each EM step are almost negligible.

Table 6 shows the β-coefficients and their standard errors for the *exact*, *P-REML* and *Taylor 6th* methods and also for *Matlab LP*, the code associated with the latter does not provide standard error estimates.

Similar conclusions were obtained from fitting a SEM, see Tables 7 and 8 (*Matlab LP* is not included since it only caters for SAMs): all approximate methods have not maximised L_s and the resulting estimates are unreliable. In

Table 5: Estimates of ρ and ω for a selected sample of size $n_s = 5,072$ for various methods with the SAM with starting values $\rho = 0.5$ and $\omega = 0.1$

Method cond. mean / trace	$\hat{\rho}$	$\hat{\omega}$	L_s	(approx.) $Q(\boldsymbol{\theta}\|\boldsymbol{\theta}')$	EM-steps	Time in seconds
exact/ exact	0.6197	0.0799	$-2,171.71$	$-6,282$	195	4,630
exact/ LP-1	0.8841	0.0081	$-9,278.81$	$-18,341$	60	1,550
exact/ LP-2	0.7521	0.0451	$-2,396.31$	-614	53	1,308
exact/ LP-3	0.8841	0.0325	$-3,286.83$	$-18,341$	60	1,550
Matlab LP	0.7830	0.1611	$-3,537.17$	$-21,336$	22	27
exact/ 6th	0.3992	0.1399	$-2,352.04$	$-11,899$	52	1,613
exact/ equal ρ	0.2866	0.1406	$-2,462.35$	$-6,282$	136	3,334
LP/ exact	0.4086	0.1233	$-2,319.28$	$-10,349$	300	4,154
LP/ LP-1	0.6567	0.0189	$-6,050.20$	$-11,564$	57	192
LP/ LP-2	0.6418	0.0838	$-2,195.42$	$-7,118$	43	683
LP/ LP-3	0.6567	0.0822	$-2,205.70$	$-11,564$	43	683
exact/ P-REML	0.6185	0.0803	$-2,171.72$	$-6,396$	195	4,711

this case, the methods closest to *exact* were *LP/LP-3, Taylor 6th* and *LP/LP-2*.

6 Conclusion

In this article we show that the algorithm in LP04 is an EM algorithm with approximate E- and M-steps. We compare their algorithm, and several other approximations, to exact methods and assess their utility. Through simulation studies on a well known dataset, we show that the proposed exact methods are the only 'trust-worthy' methods when the number of observed units is much less than the total number of units. It seems that in this regime, the exact and computationally demanding calculation of the term $\text{tr}(\mathbf{M}_{uu}(\rho')^{-1}\mathbf{M}_{uu}(\rho))$ cannot be avoided.

On the other hand the methods *exact/Taylor 6th* and *exact/equal* ρ are reliable when the sampling ratio is at least $n_s/n = 200/506 \approx 40\%$ and serve in this case as computationally faster alternatives. We deem these methods suitable for finding starting values for the exact methods. For example, in each M-step the exact $Q(\boldsymbol{\theta}'|\boldsymbol{\theta}')$ and $Q(\boldsymbol{\theta}|\boldsymbol{\theta}')$ can be calculated to check that the generalised EM criterion $Q(\boldsymbol{\theta}'|\boldsymbol{\theta}') < Q(\boldsymbol{\theta}|\boldsymbol{\theta}')$ is satisfied. Once this condition is violated one should then resort to exact M-steps.

A contribution of this work is the implementation of a *P-REML* EM approach in the context of spatial econometric models. We find that *P-REML*

Table 6: EM estimates for a selected sample of size $n_s = 5,072$ for various methods with the SAM

	Matlab LP	ML	P-REML	Taylor 6th
	est. (s.e.)	est. (s.e.)	est. (s.e.)	est. (s.e.)
Intercept	−0.3193 (−−)	0.0307 (0.1087)	0.0334 (0.1090)	0.6778 (0.1794)
age	0.6966 (−−)	1.1161 (0.0879)	1.1194 (0.0882)	1.7107 (0.1374)
age^2	−1.1609 (−−)	−1.9396 (0.1643)	−1.9461 (0.1648)	−3.1837 (0.2628)
age^3	0.2927 (−−)	0.5019 (0.0872)	0.5042 (0.0874)	0.9604 (0.1327)
$\log(lotsize)$	0.0212 (−−)	0.0425 (0.0048)	0.0427 (0.0048)	0.0771 (0.0082)
$rooms$	−0.0153 (−−)	−0.0098 (0.0060)	−0.0098 (0.0060)	−0.0053 (0.0081)
$\log(TLA)$	0.3551 (−−)	0.5191 (0.0210)	0.5203 (0.0210)	0.7221 (0.0323)
$beds$	0.0027 (−−)	−0.0084 (0.0088)	−0.0085 (0.0089)	−0.0206 (0.0120)
$syear1994$	0.0358 (−−)	0.0464 (0.0152)	0.0465 (0.0152)	0.0638 (0.0201)
$syear1995$	0.0662 (−−)	0.0830 (0.0148)	0.0831 (0.0148)	0.1047 (0.0196)
$syear1996$	0.0613 (−−)	0.0750 (0.0142)	0.0751 (0.0142)	0.0928 (0.0188)
$syear1997$	0.0841 (−−)	0.1130 (0.0140)	0.1132 (0.0140)	0.1437 (0.0186)
$syear1998$	0.1172 (−−)	0.1578 (0.0147)	0.1581 (0.0147)	0.1867 (0.0194)
ρ	0.7830 (−−)	0.6197 (0.0108)	0.6185 (0.0108)	0.3992 (0.0263)
ω	0.1611 (−−)	0.0798 (0.0018)	0.0803 (0.0018)	0.1399 (0.0059)

Table 7: Estimates of ρ and ω for a selected sample of size $n_s = 5,072$ for various methods with the SEM with starting values $\rho = 0.5$ and $\omega = 0.1$

Method cond. mean / trace	$\hat{\rho}$	$\hat{\omega}$	L_s	(approx.) $Q(\boldsymbol{\theta}\|\boldsymbol{\theta}')$	EM-steps	Time in seconds
exact/ exact	0.6888	0.0781	−2,564.30	−6,732.01	177	4,568
exact/ LP-1	0.9493	0.0054	−10,038.13	−20,816.29	154	5,570
exact/ LP-2	0.8760	0.0290	−3,049.66	−2,468.32	114	2,714
exact/ LP-3	0.9493	0.0218	−4,042.03	−20,816.29	154	5,570
exact/ 6th	0.3726	0.1561	−2,718.41	−13,173.44	41	1,131
exact/ equal ρ	0.0985	0.1779	−2,846.78	−14,143.62	38	956
LP/ exact	0.3226	0.1548	−2,741.60	−12,873.97	71	2,011
LP/ LP-1	0.7187	0.0221	−4,035.04	−8,895.99	65	402
LP/ LP-2	0.7551	0.0876	−2,752.13	−9,076.47	69	818
LP/ LP-3	0.7187	0.0913	−2,681.81	−8,895.99	65	402
exact/ P-REML	0.6869	0.0787	−2,564.33	−6,872.78	177	4,259

Table 8: EM estimates for a selected sample of size $n_s = 5,072$ for various methods with the SEM

	ML	P-REML	Taylor 6th	LP/LP-3
	est. (s.e.)	est. (s.e.)	est. (s.e.)	est. (s.e.)
Intercept	3.7244 (0.1811)	3.7178 (0.1815)	3.0715 (0.1979)	4.2229 (0.2090)
age	1.8950 (0.1719)	1.9008 (0.1721)	2.3692 (0.1748)	1.6909 (0.2038)
age^2	−4.2835 (0.2905)	−4.2929 (0.2909)	−5.0015 (0.3021)	−3.8849 (0.3416)
age^3	1.6249 (0.1479)	1.6277 (0.1482)	1.8124 (0.1580)	1.6025 (0.1719)
$\log(lotsize)$	0.1958 (0.0099)	0.1956 (0.0099)	0.1722 (0.0099)	0.2022 (0.0118)
$rooms$	0.0073 (0.0083)	0.0073 (0.0083)	0.0073 (0.0096)	0.0012 (0.0093)
$\log(TLA)$	0.7606 (0.0275)	0.7618 (0.0276)	0.8811 (0.0310)	0.6801 (0.0311)
$beds$	−0.0092 (0.0121)	−0.0094 (0.0122)	−0.0268 (0.0141)	0.0034 (0.0136)
$syear1994$	0.0700 (0.0194)	0.0700 (0.0195)	0.0760 (0.0228)	0.0635 (0.0217)
$syear1995$	0.1043 (0.0186)	0.1044 (0.0187)	0.1136 (0.0221)	0.0955 (0.0208)
$syear1996$	0.0975 (0.0180)	0.0975 (0.0181)	0.1010 (0.0212)	0.0987 (0.0202)
$syear1997$	0.1648 (0.0178)	0.1648 (0.0179)	0.1647 (0.0210)	0.1650 (0.0199)
$syear1998$	0.2007 (0.0184)	0.2006 (0.0184)	0.1977 (0.0217)	0.1826 (0.0206)
ρ	0.6888 (0.0095)	0.6869 (0.0096)	0.3726 (0.0221)	0.7187 (0.0072)
ω	0.0781 (0.0018)	0.0787 (0.0018)	0.1561 (0.0046)	0.0913 (0.0030)

improves estimation of the parameters ρ and ω determining the covariance matrix $\mathbf{\Sigma}$, in the sense that the *P-REML* estimates are never worse than the *exact* estimates. While this *P-REML* approach reduces the bias in ρ and ω, the improvement is more substantial for the SEM than for the SAM, presumably because strict REML is not possible with the SAM. Overall our results indicate that *P-REML* can be recommended over the classical ML approach in the presence of missing data with both the SEM and the SAM. This does not come with increased computational complexity: Note that the computational times required by *P-REML* are comparable to those of the standard EM algorithm. Future research will focus on developing faster variants of this REML approach for hidden data, and aim at extending it to a larger class of spatial econometric models.

`R` code replicating the studies in this article is available from the first author.

Acknowledgements

We are grateful to James LeSage for providing the Matlab code used in the studies, and to Noel Cressie and the reviewers who provided us with further

references and helpful comments.

References

[1] Ord K. Estimation methods for models of spatial interaction. Journal of the American Statistical Association. 1975;70(349):120–126.

[2] Pace R, Barry R. Fast CARs. Journal of Statistical Computation and Simulation. 1997;59(2):123–147.

[3] Lee Lf, Yu J. Estimation of spatial autoregressive panel data models with fixed effects. Journal of Econometrics. 2010;154(2):165–185.

[4] Lin X, Lee Lf. Gmm estimation of spatial autoregressive models with unknown heteroskedasticity. Journal of Econometrics. 2010;157(1):34–52.

[5] Su L. Semiparametric gmm estimation of spatial autoregressive models. Journal of Econometrics. 2012;167(2):543–560.

[6] Wang W, Lee L. Estimation of spatial autoregressive models with randomly missing data in the dependent variable. Econometrics Journal. 2013;16:73–102.

[7] Bivand RS, Gómez-Rubio V, Rue H. Approximate Bayesian inference for spatial econometrics models. Spatial Statistics. 2014;9:146–165.

[8] Fernández-Macho J. Missing observations in spatial models: a spectral em algorithm. Journal of Computational and Graphical Statistics. 2010; 19(3):684–701.

[9] Lesage J, Pace R. Models for spatially dependent missing data. Journal of Real Estate Finance and Economics. 2004;29(2):223–254.

[10] Goulard M, Laurent T, Thomas-Agnan C. About predictions in spatial autoregressive models: Optimal and almost optimal strategies. Toulouse School of Economics Working Paper. 2013;13:452.

[11] Li H, Calder CA, Cressie N. One-step estimation of spatial dependence parameters: Properties and extensions of the aple statistic. Journal of Multivariate Analysis. 2012;105(1):68–84.

[12] McLachlan G, Krishnan T. The EM algorithm and extensions. 2nd ed. New Jersey: Wiley; 2008.

[13] Bates D, Maechler M. Matrix: Sparse and Dense Matrix Classes and Methods. 2015; Available at http://CRAN.R-project.org/package=Matrix.

[14] Bivand R. spdep: Spatial Dependence: Weighting Schemes, Statistics and Models. 2015; Available at http://CRAN.R-project.org/package=spdep.

[15] Chen Y, Davis TA, Hager WW, Rajamanickam S. Algorithm 887: CHOLMOD, supernodal sparse Cholesky factorization and update/downdate. ACM Transactions on Mathematical Software. 2008; 35(3):22.

[16] Eaton ML. Multivariate statistics: A vector space approach. JOHN WILEY & SONS; 1983.

[17] Harville D. Matrix algebra from a statistician's perspective. New York: Springer; 2000.

[18] Verbyla AP. A conditional derivation of residual maximum likelihood. Australian Journal of Statistics. 1990;32(2):227–230.

[19] Hodges JS, Reich BJ. Adding spatially-correlated errors can mess up the fixed effect you love. The American Statistician. 2010;64(4):325–334.

[20] Samart K, Chambers R. Linear regression with nested errors using probability-linked data. Australian & New Zealand Journal of Statistics. 2014;56(1):27–46.

[21] Lindstrom MJ, Bates DM. Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. Journal of the American Statistical Association. 1988;83(404):1014–1022.

[22] Diffey S, Welsh A, Smith A, Cullis BR. A faster and computationally more efficient REML (PX) EM algorithm for linear mixed models. National Institute for Applied Statistics Research Australia, The University of Wollongong; 2013. Report 2-13.

[23] Orchard T, Woodbury M. A missing information principle: Theory and application. 1972; proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability.

[24] Besag J. Efficiency of pseudolikelihood estimation for simple gaussian fields. Biometrika. 1977;64(3):616–618.

[25] Gelman A, Carlin J, Stern H, Rubin D. Bayesian data analysis. Boca Raton, FL: Chapman & Hall/CRC; 2004.

[26] Pace R, Lesage J. Spatial econometric models, prediction. New York: Springer; 2008. p. 1095–1100; encyclopedia of Geographical Information Science, Shashi Shekar & Hui Xiong (Eds.).

[27] Rasmussen CE. Gaussian processes for machine learning. 2006;.

[28] Werner D. Functional analysis. Berlin: Springer; 2011.

[29] Harrison D, Rubinfeld DL. Hedonic housing prices and the demand for clean air. Journal of Environmental Economics and Management. 1978; 5:81–102.

[30] Bivand R. Comparing estimation methods for spatial econometrics techniques using r. Department of Economics, Norwegian School of Economics and Business Administration; 2010. Report; SAM 26 2010.

[31] Meng X, Rubin D. Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. Journal of the American Statistical Association. 1991;86(416):899–909.

A Summary of Main Terms

Table 9: Expressions for $\boldsymbol{\mu}$, $\mathbf{V}$, $\log|\mathbf{V}|$ and $\mathbf{r}^T\mathbf{V}^{-1}\mathbf{r}$ with $\mathbf{A} \equiv \mathbf{I} - \rho\mathbf{W}$, $\mathbf{r} \equiv \mathbf{y} - \boldsymbol{\mu}$, $\mathbf{r}_y \equiv \mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ and $\omega \equiv \sigma^2$ for the SAM and the SEM

	SAM	SEM
$\boldsymbol{\mu}$	$\boldsymbol{A}^{-1}\mathbf{X}\boldsymbol{\beta}$	$\mathbf{X}\boldsymbol{\beta}$
$\boldsymbol{\Sigma} = \omega\mathbf{V}$	$\omega(\mathbf{A}^T\mathbf{A})^{-1}$	$\omega(\mathbf{A}^T\mathbf{A})^{-1}$
$\mathbf{V}$	$(\mathbf{A}^T\mathbf{A})^{-1}$	$(\mathbf{A}^T\mathbf{A})^{-1}$
$\mathbf{M} = \mathbf{V}^{-1}$	$\mathbf{A}^T\mathbf{A}$	$\mathbf{A}^T\mathbf{A}$
$\log\lvert\mathbf{M}\rvert$	$2\log\lvert\mathbf{A}\rvert$	$2\log\lvert\mathbf{A}\rvert$
$\mathbf{r}^T\mathbf{V}^{-1}\mathbf{r}$	$\mathbf{r}_y^T\mathbf{r}_y$	$(\mathbf{A}\mathbf{r})^T\mathbf{A}\mathbf{r}$

B Showing Equivalence of Two Methods

Here we show that setting $\hat{\sigma}^2 = (\mathbf{Ar})_s^T(\mathbf{Ar})_s/n_s$ is identical to the approximation given on page 240 in LP04, where the formula for an estimate of $\omega = \sigma^2$ for the SAM for given ρ is

$$\hat{\sigma}^2 = \frac{1}{n_s}(e_0^s - \rho e_d^s)^T(e_0^s - \rho e_d^s). \tag{23}$$

In (23), $e_0 = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0$, $e_d = \mathbf{Wy} - \mathbf{X}\boldsymbol{\beta}_d$, $\boldsymbol{\beta}_0 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ and $\boldsymbol{\beta}_d = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Wy}$. The superscript s refers to the sample, that is, $e_0^s = (e_0)_s$.

Now using $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{Ay} = \boldsymbol{\beta}_0 - \rho\boldsymbol{\beta}_d$ for the SAM

$$\begin{aligned} e_0^s - \rho e_d^s &= (\mathbf{y} - \rho\mathbf{Wy} - \mathbf{X}(\boldsymbol{\beta}_0 - \rho\boldsymbol{\beta}_d))_s \\ &= (\mathbf{y} - \rho\mathbf{Wy} - \mathbf{X}\hat{\boldsymbol{\beta}})_s \\ &= (\mathbf{Ay} - \mathbf{X}\hat{\boldsymbol{\beta}})_s \\ &= (\mathbf{A}(\mathbf{y} - \mathbf{A}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}}))_s \\ &= (\mathbf{Ar})_s. \end{aligned}$$

Hence

$$\hat{\sigma}^2 = (\mathbf{Ar})_s^T(\mathbf{Ar})_s/n_s.$$

C Pseudo-REML EM Algorithm

C.1 Approach of Lindstrom and Bates

First we consider Verbyla's [18] REML algorithm and show its equivalence to the approach by Lindstrom and Bates [21]. Let $\mathbf{L}_1$ be a $p \times n$ matrix such that $\mathbf{L}_1\mathbf{X} = \mathbf{I}_p$, and let $\mathbf{L}_2$ be a $(n-p) \times n$ matrix such that $\mathbf{L}_2\mathbf{X} = \mathbf{0}$. Note that $\mathbf{K}$ in Section 2.3 is identical to $\mathbf{L}_2$. Decompose the data $\mathbf{y}$ into two parts, $\mathbf{y}_1 = \mathbf{L}_1\mathbf{y}$ and $\mathbf{y}_2 = \mathbf{L}_2\mathbf{y}$, where the marginal density of $\mathbf{y}_2$, $f(\mathbf{y}_2)$, only depends on the parameters affecting the variance, say $\boldsymbol{\psi}$, where $\boldsymbol{\psi} = (\omega, \rho)^T$. Verbyla [18] suggested to decompose the joint distribution

$$f(\mathbf{y}_1, \mathbf{y}_2) = f(\mathbf{y}_1|\mathbf{y}_2)f(\mathbf{y}_2),$$

and first to estimate the variance parameters $\boldsymbol{\psi}$ by maximising $f(\mathbf{y}_2)$ and then, given those, to obtain estimates of the fixed effects parameter $\boldsymbol{\beta}$ from maximis-

ing the conditional density $f(\mathbf{y}_1|\mathbf{y}_2)$, which leads to $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{M}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{M}\mathbf{y}$. The key step is the first step as this leads to REML estimates of the variance parameters, which then defines $\hat{\boldsymbol{\beta}}$. The log density of $\mathbf{y}_2$ has the following form [18]:

$$\log f(\mathbf{y}_2;\boldsymbol{\psi}) \propto -\frac{1}{2}\left((n-p)\log\omega + \log|\mathbf{M}| - \log|\mathbf{X}^T\mathbf{M}\mathbf{X}| - \frac{1}{\omega}\mathbf{y}^T\mathbf{P}\mathbf{y}\right).$$

From [18], the term $\mathbf{y}^T\mathbf{P}\mathbf{y}$ can be written as

$$\begin{aligned}\mathbf{y}^T\mathbf{P}\mathbf{y} &= \mathbf{y}^T(\mathbf{M} - \mathbf{M}\mathbf{X}(\mathbf{X}^T\mathbf{M}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{M})\mathbf{y} \\ &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\rho))^T\mathbf{M}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\rho)) \\ &\equiv \hat{\mathbf{r}}(\rho)^T\mathbf{M}\hat{\mathbf{r}}(\rho),\end{aligned}$$

leading to

$$\log f(\mathbf{y}_2;\boldsymbol{\psi}) = c - \frac{1}{2}\left((n-p)\log\omega + \log|\mathbf{M}| - \log|\mathbf{X}^T\mathbf{M}\mathbf{X}| - \frac{1}{\omega}\hat{\mathbf{r}}(\rho)^T\mathbf{M}\hat{\mathbf{r}}(\rho)\right), \tag{24}$$

where c is a constant independent of $\boldsymbol{\psi}$. This means that maximising $f(\mathbf{y}_2;\boldsymbol{\psi})$ is equivalent to maximising $L_R \equiv \log g(\mathbf{y};\boldsymbol{\theta})$, see (14), but only if $\boldsymbol{\beta}$ is replaced by $\hat{\boldsymbol{\beta}}$. However maximising L_R leads to $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, which can be shown by differentiating and setting the expression to zero. Therefore maximising L_R is equivalent to maximising (24), that is, maximising the objective used by Lindstrom and Bates [21].

C.2 EM for finding $\arg\max_{\boldsymbol{\theta}} g(\mathbf{y}_s|\boldsymbol{\theta})$

Consider the function $\log g(\mathbf{y};\boldsymbol{\theta})$ and let us aim to maximise the marginal function $g(\mathbf{y}_s;\boldsymbol{\theta}) = \int g(\mathbf{y};\boldsymbol{\theta})\mathrm{d}\mathbf{y}_\mathbf{u}$ to obtain 'Pseudo-REML' estimates.

We note that $g(\mathbf{y};\boldsymbol{\theta}) = c(\boldsymbol{\theta})f(\mathbf{y};\boldsymbol{\theta})$, where $\log f(\mathbf{y};\boldsymbol{\theta})$ is log-likelihood defined in (5) and $\log c(\boldsymbol{\theta})$ is the constant that has been added in (14) to obtain REML estimates in the complete data case.

Since $f(\mathbf{y}_s,\mathbf{y}_u;\boldsymbol{\theta}) = f(\mathbf{y}_u|\mathbf{y}_s;\boldsymbol{\theta})f(\mathbf{y}_s;\boldsymbol{\theta})$ and

$$g(\mathbf{y}_s;\boldsymbol{\theta}) = \int g(\mathbf{y};\boldsymbol{\theta})d\mathbf{y}_\mathbf{u} = c(\boldsymbol{\theta})\int f(\mathbf{y};\boldsymbol{\theta})\mathrm{d}\mathbf{y}_\mathbf{u} = c(\boldsymbol{\theta})f(\mathbf{y}_s;\boldsymbol{\theta}),$$

we have that

$$\log g(\mathbf{y}_s;\boldsymbol{\theta}) = \log c(\boldsymbol{\theta}) + \log f(\mathbf{y}_s,\mathbf{y}_u;\boldsymbol{\theta}) - \log f(\mathbf{y}_u|\mathbf{y}_s;\boldsymbol{\theta}).$$

Therefore

$$\begin{aligned}\log g(\mathbf{y}_s;\boldsymbol{\theta}) &= \log c(\boldsymbol{\theta}) + \int \log f(\mathbf{y}_s,\mathbf{y}_u;\boldsymbol{\theta}) f(\mathbf{y}_u|\mathbf{y}_s;\boldsymbol{\theta}')\mathrm{d}\mathbf{y}_u - \int \log f(\mathbf{y}_u|\mathbf{y}_s;\boldsymbol{\theta}) f(\mathbf{y}_u|\mathbf{y}_s;\boldsymbol{\theta}')\mathrm{d}\mathbf{y}_u \\ &= \widetilde{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}') + H(\boldsymbol{\theta}|\boldsymbol{\theta}'),\end{aligned}$$

where

$$\widetilde{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}') = \log c(\boldsymbol{\theta}) + \int \log f(\mathbf{y}_s,\mathbf{y}_u;\boldsymbol{\theta}) f(\mathbf{y}_u|\mathbf{y}_s;\boldsymbol{\theta}')\mathrm{d}\mathbf{y}_u,$$

and

$$H(\boldsymbol{\theta}|\boldsymbol{\theta}') = -\int \log f(\mathbf{y}_u|\mathbf{y}_s;\boldsymbol{\theta}) f(\mathbf{y}_u|\mathbf{y}_s;\boldsymbol{\theta}')\mathrm{d}\mathbf{y}_u.$$

The standard EM algorithm [12] uses the same definition of H but the Q-function (here $\widetilde{Q}$) is usually defined without the constant $\log c(\boldsymbol{\theta})$, that is, $\widetilde{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}') = \log c(\boldsymbol{\theta}) + Q(\boldsymbol{\theta}|\boldsymbol{\theta}')$.

Following standard arguments for the EM algorithm, by Gibbs' inequality $H(\boldsymbol{\theta}|\boldsymbol{\theta}') - H(\boldsymbol{\theta}'|\boldsymbol{\theta}') \geq 0$, leading to

$$\log g(\mathbf{y}_s;\boldsymbol{\theta}) - \log g(\mathbf{y}_s;\boldsymbol{\theta}') \geq \widetilde{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}') - \widetilde{Q}(\boldsymbol{\theta}'|\boldsymbol{\theta}') \geq 0,$$

meaning that an improvement in the $\widetilde{Q}$-function (achieved in the M-step) leads to an improvement (larger value) in the $g(\mathbf{y}_s;\boldsymbol{\theta})$-function, which is to be maximised.

Since our EM algorithm uses $f(\mathbf{y}_u|\mathbf{y}_s;\boldsymbol{\theta}')$ in the E-step to calculate

$$E_{u|s}\log g(\mathbf{y};\boldsymbol{\theta}) = E_{u|s}\log\{c(\boldsymbol{\theta})f(\mathbf{y};\boldsymbol{\theta})\} = \log c(\boldsymbol{\theta}) + E_{u|s}\log f(\mathbf{y};\boldsymbol{\theta}) = \log c(\boldsymbol{\theta}) + Q(\boldsymbol{\theta}|\boldsymbol{\theta}'),$$

this implies that the proposed EM algorithm (adding $\log c(\boldsymbol{\theta})$ to $Q(\boldsymbol{\theta}|\boldsymbol{\theta}')$ in the M-step) maximises the marginal function $g(\mathbf{y}_s;\boldsymbol{\theta})$.

D Information Matrix

We first express the observed log-likelihood $L_s \equiv \log f(\mathbf{y}_s;\boldsymbol{\theta})$ in terms of the complete-data log-likelihood $L_c \equiv \log f(\mathbf{y};\boldsymbol{\theta})$ and the missing-data log-likelihood $L_m \equiv \log f(\mathbf{y}_u|\mathbf{y}_s;\boldsymbol{\theta})$ [31]:

$$\log f(\mathbf{y}_s|\boldsymbol{\theta}) = \log f(\mathbf{y}|\boldsymbol{\theta}) - f(\mathbf{y}_u|\mathbf{y}_s,\boldsymbol{\theta}),$$

or as

$$L_s = L_c - L_m.$$

Then L_c is given in (5) and L_{m} is given by

$$\log f(\mathbf{y}_u|\mathbf{y}_s;\boldsymbol{\theta}) = -\frac{n_u}{2}\log(2\pi) - \frac{n_u}{2}\log\omega + \frac{1}{2}\log|\mathbf{M}_{uu}| - \frac{1}{2\omega}\tilde{\mathbf{r}}_u^T\mathbf{M}_{uu}\tilde{\mathbf{r}}_u, \quad (25)$$

where $\tilde{\mathbf{r}}_u \equiv \mathbf{y}_u - \boldsymbol{\mu}_{u|s}$.

Define $\mathbf{B}_{ss} \equiv \mathbf{M}_{ss} - \mathbf{M}_{su}\mathbf{M}_{uu}^{-1}\mathbf{M}_{us}$ and $\mathbf{r}_s \equiv \mathbf{y}_s - \boldsymbol{\mu}_s$. After some lengthy algebra, it can be shown that the elements of the observed information matrix with $\tilde{L} = -L_s$ for the SAM are

$$\begin{aligned}
\frac{\partial^2\tilde{L}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} &= \frac{1}{\omega}\widetilde{\mathbf{X}}_s^T\mathbf{B}_{ss}\widetilde{\mathbf{X}}_s,\\
\frac{\partial^2\tilde{L}}{\partial\omega^2} &= -\frac{n_s}{2}\frac{1}{\omega^2} + \frac{1}{\omega^3}\mathbf{r}_s^T\mathbf{B}_{ss}\mathbf{r}_s\\
\frac{\partial^2\tilde{L}}{\partial\rho^2} &= \frac{1}{2}\mathrm{tr}\left(\mathbf{B}_{ss}^{-1}\frac{\partial\mathbf{B}_{ss}}{\partial\rho}\mathbf{B}_{ss}^{-1}\frac{\partial\mathbf{B}_{ss}}{\partial\rho} - \mathbf{B}_{ss}^{-1}\frac{\partial^2\mathbf{B}_{ss}}{\partial\rho^2}\right)\\
&\quad + \frac{1}{\omega}\left(\frac{\partial\mathbf{r}_s^T}{\partial\rho}\mathbf{B}_{ss}\frac{\partial\mathbf{r}_s}{\partial\rho} + \frac{\partial^2\mathbf{r}_s^T}{\partial\rho^2}\mathbf{B}_{ss}\mathbf{r}_s + 2\frac{\partial\mathbf{r}_s^T}{\partial\rho}\frac{\partial\mathbf{B}_{ss}}{\partial\rho}\mathbf{r}_s + \mathbf{r}_s^T\frac{\partial^2\mathbf{B}_{ss}}{\partial\rho^2}\mathbf{r}_s\right),\\
\frac{\partial^2\tilde{L}}{\partial\omega\partial\boldsymbol{\beta}} &= -\frac{1}{\omega^2}\widetilde{\mathbf{X}}_s^T\mathbf{B}_{ss}\mathbf{r}_s,\\
\frac{\partial^2\tilde{L}}{\partial\rho\partial\boldsymbol{\beta}} &= -\frac{1}{\omega}\left[\widetilde{\mathbf{X}}_s^T\frac{\partial\mathbf{B}_{ss}}{\partial\rho}\mathbf{r}_s + \widetilde{\mathbf{X}}_s^T\mathbf{B}_{ss}\left(\frac{\partial\mathbf{r}_s}{\partial\rho}\right) + \frac{\partial\widetilde{\mathbf{X}}_s^T}{\partial\rho}\mathbf{B}_{ss}\mathbf{r}_s\right],\\
\frac{\partial^2 L_s}{\partial\omega\partial\rho} &= -\frac{1}{2\omega^2}\left(2\mathbf{r}_s^T\mathbf{B}_{ss}\frac{\partial\mathbf{r}_s}{\partial\rho} + \mathbf{r}_s^T\frac{\partial\mathbf{B}_{ss}}{\partial\rho}\mathbf{r}_s\right),
\end{aligned}$$

with

$$\frac{\partial \mathbf{r}_s}{\partial \rho} = -\frac{\partial \widetilde{\mathbf{X}}_s}{\partial \rho}\boldsymbol{\beta}, \qquad \frac{\partial \widetilde{\mathbf{X}}_s}{\partial \rho} = \left[\mathbf{A}^{-1}\mathbf{W}\mathbf{A}^{-1}\mathbf{X}\right]_s,$$

$$\frac{\partial^2 \mathbf{r}_s}{\partial \rho^2} = 2\left[\mathbf{A}^{-1}\mathbf{W}\mathbf{A}^{-1}\mathbf{W}\mathbf{A}^{-1}\mathbf{X}\right]_s\boldsymbol{\beta},$$

$$\frac{\partial \mathbf{B}_{ss}}{\partial \rho} = \frac{\partial \mathbf{M}_{ss}}{\partial \rho} + \mathbf{M}_{su}\mathbf{M}_{uu}^{-1}\frac{\partial \mathbf{M}_{uu}}{\partial \rho}\mathbf{M}_{uu}^{-1}\mathbf{M}_{us} - \frac{\partial \mathbf{M}_{su}}{\partial \rho}\mathbf{M}_{uu}^{-1}\mathbf{M}_{us} - \mathbf{M}_{su}\mathbf{M}_{uu}^{-1}\frac{\partial \mathbf{M}_{us}}{\partial \rho},$$

$$\begin{aligned}\frac{\partial^2 \mathbf{B}_{ss}}{\partial \rho^2} &= \frac{\partial^2 \mathbf{M}_{ss}}{\partial \rho^2} - \left(\frac{\partial^2 \mathbf{M}_{su}}{\partial \rho^2}\mathbf{M}_{uu}^{-1}\mathbf{M}_{us} + \left(\frac{\partial^2 \mathbf{M}_{us}}{\partial \rho^2}\mathbf{M}_{uu}^{-1}\mathbf{M}_{us}\right)^T\right)\\ &\quad + 2\left(\frac{\partial \mathbf{M}_{su}}{\partial \rho}\mathbf{M}_{uu}^{-1}\frac{\partial \mathbf{M}_{uu}}{\partial \rho}\mathbf{M}_{uu}^{-1}\mathbf{M}_{us} + \left(\frac{\partial \mathbf{M}_{su}}{\partial \rho}\mathbf{M}_{uu}^{-1}\mathbf{M}_{us}\right)^T\right)\\ &\quad -2\frac{\partial \mathbf{M}_{su}}{\partial \rho}\mathbf{M}_{uu}^{-1}\frac{\partial \mathbf{M}_{us}}{\partial \rho} + \mathbf{M}_{su}\mathbf{M}_{uu}^{-1}\frac{\partial^2 \mathbf{M}_{uu}}{\partial \rho^2}\mathbf{M}_{uu}^{-1}\mathbf{M}_{us}\\ &\quad - 2\mathbf{M}_{su}\mathbf{M}_{uu}^{-1}\frac{\partial \mathbf{M}_{uu}}{\partial \rho}\mathbf{M}_{uu}^{-1}\frac{\partial \mathbf{M}_{uu}}{\partial \rho}\mathbf{M}_{uu}^{-1}\mathbf{M}_{us},\end{aligned}$$

$$\frac{\partial \mathbf{M}}{\partial \rho} = -(\mathbf{W} + \mathbf{W}^T) + 2\rho\mathbf{W}^T\mathbf{W},$$

$$\frac{\partial^2 \mathbf{M}}{\partial \rho^2} = 2\mathbf{W}^T\mathbf{W}.$$

The elements of the observed information matrix for the SEM are

$$\frac{\partial^2 \tilde{L}}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} = \frac{1}{\omega}\mathbf{X}_s^T\mathbf{B}_{ss}\mathbf{X}_s,$$

$$\frac{\partial^2 \tilde{L}}{\partial \omega^2} = -\frac{n_s}{2}\frac{1}{\omega^2} + \frac{1}{\omega^3}\mathbf{r}_s^T\mathbf{B}_{ss}\mathbf{r}_s,$$

$$\frac{\partial^2 \tilde{L}}{\partial \rho^2} = \frac{1}{2}\mathrm{tr}\left(\mathbf{B}_{ss}^{-1}\frac{\partial \mathbf{B}_{ss}}{\partial \rho}\mathbf{B}_{ss}^{-1}\frac{\partial \mathbf{B}_{ss}}{\partial \rho} - \mathbf{B}_{ss}^{-1}\frac{\partial^2 \mathbf{B}_{ss}}{\partial \rho^2}\right) + \frac{1}{\omega}\left(\mathbf{r}_s^T\frac{\partial^2 \mathbf{B}_{ss}}{\partial \rho^2}\mathbf{r}_s\right),$$

$$\frac{\partial^2 \tilde{L}}{\partial \omega \partial \boldsymbol{\beta}} = -\frac{1}{\omega^2}\mathbf{X}_s^T\mathbf{B}_{ss}\mathbf{r}_s,$$

$$\frac{\partial^2 \tilde{L}}{\partial \rho \partial \boldsymbol{\beta}} = -\frac{1}{\omega}\mathbf{X}_s^T\frac{\partial \mathbf{B}_{ss}}{\partial \rho}\mathbf{r}_s,$$

$$\frac{\partial^2 L_s}{\partial \omega \partial \rho} = -\frac{1}{2\omega^2}\mathbf{r}_s^T\frac{\partial \mathbf{B}_{ss}}{\partial \rho}\mathbf{r}_s,$$

with

$$\frac{\partial \mathbf{B}_{ss}}{\partial \rho} = \frac{\partial \mathbf{M}_{ss}}{\partial \rho} + \mathbf{M}_{su}\mathbf{M}_{uu}^{-1}\frac{\partial \mathbf{M}_{uu}}{\partial \rho}\mathbf{M}_{uu}^{-1}\mathbf{M}_{us} - \frac{\partial \mathbf{M}_{su}}{\partial \rho}\mathbf{M}_{uu}^{-1}\mathbf{M}_{us} - \mathbf{M}_{su}\mathbf{M}_{uu}^{-1}\frac{\partial \mathbf{M}_{us}}{\partial \rho},$$

$$\begin{aligned}\frac{\partial^2 \mathbf{B}_{ss}}{\partial \rho^2} = \frac{\partial^2 \mathbf{M}_{ss}}{\partial \rho^2} &- \left(\frac{\partial^2 \mathbf{M}_{su}}{\partial \rho^2}\mathbf{M}_{uu}^{-1}\mathbf{M}_{us} + \left(\frac{\partial^2 \mathbf{M}_{us}}{\partial \rho^2}\mathbf{M}_{uu}^{-1}\mathbf{M}_{us}\right)^T\right) \\ &+ 2\left(\frac{\partial \mathbf{M}_{su}}{\partial \rho}\mathbf{M}_{uu}^{-1}\frac{\partial \mathbf{M}_{uu}}{\partial \rho}\mathbf{M}_{uu}^{-1}\mathbf{M}_{us} + \left(\frac{\partial \mathbf{M}_{su}}{\partial \rho}\mathbf{M}_{uu}^{-1}\mathbf{M}_{us}\right)^T\right) \\ &-2\frac{\partial \mathbf{M}_{su}}{\partial \rho}\mathbf{M}_{uu}^{-1}\frac{\partial \mathbf{M}_{us}}{\partial \rho} + \mathbf{M}_{su}\mathbf{M}_{uu}^{-1}\frac{\partial^2 \mathbf{M}_{uu}}{\partial \rho^2}\mathbf{M}_{uu}^{-1}\mathbf{M}_{us} \\ &- 2\mathbf{M}_{su}\mathbf{M}_{uu}^{-1}\frac{\partial \mathbf{M}_{uu}}{\partial \rho}\mathbf{M}_{uu}^{-1}\frac{\partial \mathbf{M}_{uu}}{\partial \rho}\mathbf{M}_{uu}^{-1}\mathbf{M}_{us}.\end{aligned}$$

E Further Simulation Results

Table 10: Empirical mean of ρ and ω estimates over 10,000 simulations for the SAM model with $n = 506$, $n_s = 400$ and the parameters settings $\rho = 0.5$, $\omega = 1$, $\beta_0 = 1$, $\beta_1 = 2$

		Empirical mean of $\hat{\rho}$		Empirical mean of $\hat{\omega}$	
		conditional mean		conditional mean	
		exact	LP	exact	LP
trace method	exact	0.497	0.494	0.993	0.997
	equal ρ & ω	0.518	0.515	0.984	0.988
	equal ρ	0.504	0.501	0.990	0.994
	diag	0.494	0.491	0.994	0.998
	Taylor 6th	0.497	0.494	0.994	0.998
	Taylor 1st	0.499	0.496	0.978	0.982
	LP-1	0.518	0.515	0.778	0.781
	LP-2	0.508	0.506	0.974	0.979
	LP-3	0.518	0.515	0.970	0.976
	P-REML	0.498	0.495	0.997	1.001

Table 11: Coverage of Wald-type confidence intervals for ρ and ω over 10,000 simulations for the SAM model with $n = 506$, $n_s = 400$ and the parameters settings $\rho = 0.5$, $\omega = 1$, $\beta_0 = 1$, $\beta_1 = 2$

		Coverage for ρ		Coverage for ω	
		conditional mean		conditional mean	
		exact	LP	exact	LP
trace method	exact	0.941	0.940	0.939	0.943
	equal ρ & ω	0.883	0.901	0.929	0.934
	equal ρ	0.938	0.944	0.935	0.941
	diag	0.936	0.933	0.940	0.943
	Taylor 6th	0.942	0.941	0.940	0.944
	Taylor 1st	0.948	0.948	0.923	0.927
	LP-1	0.846	0.870	0.044	0.049
	LP-3	0.876	0.899	0.910	0.919
	LP-2	0.925	0.930	0.912	0.919
	P-REML	0.943	0.942	0.943	0.946

Table 12: Empirical mean of ρ and ω estimates over 10,000 simulations for the SAM model with $n = 506$, $n_s = 100$ and the parameters $\rho = 0.5$, $\omega = 1$, $\beta_0 = 1$, $\beta_1 = 2$

		Empirical mean of $\hat{\rho}$		Empirical mean of $\hat{\omega}$	
		conditional mean		conditional mean	
		exact	LP	exact	LP
trace method	exact	0.494	0.431	0.972	1.069
	equal ρ & ω	0.639	0.549	0.866	0.991
	equal ρ	0.387	0.341	1.093	1.171
	diag	0.244	0.224	1.301	1.337
	Taylor 6th	0.462	0.418	1.072	1.111
	Taylor 1st	0.564	0.472	0.470	0.619
	LP-1	0.639	0.549	0.171	0.196
	LP-2	0.639	0.549	0.768	0.939
	LP-3	0.568	0.529	0.827	0.951
	P-REML	0.497	0.435	0.989	1.088

Table 13: Coverage of Wald-type confidence intervals for β_0 and β_1 estimates over 10,000 simulations for the SAM model with $n = 506$, $n_s = 100$ and the parameters $\rho = 0.5$, $\omega = 1$, $\beta_0 = 1$, $\beta_1 = 2$

		Coverage for β_0		Coverage for β_1	
		conditional mean		conditional mean	
		exact	LP	exact	LP
trace method	exact	0.972	0.974	0.907	0.973
	equal ρ & ω	0.829	0.931	0.729	0.949
	equal ρ	0.838	0.643	0.962	0.966
	diag	0.212	0.206	0.938	0.934
	Taylor 6th	0.994	0.971	0.978	0.979
	Taylor 1st	0.511	0.875	0.0185	0.118
	LP-1	0.089	0.467	0.000	0.000
	LP-2	0.232	0.810	0.490	0.899
	LP-3	0.865	0.946	0.568	0.894
	P-REML	0.972	0.978	0.925	0.975