

Statistical Inference on Hierarchical Simultaneous Autoregressive Models with Missing Data

Anjana Wijayawardhana*, Thomas Suesse, and David Gunawan

School of Mathematics and Applied Statistics, University of Wollongong,
Wollongong, Australia

Abstract

Efficient estimation methods for simultaneous autoregressive (SAR) models with missing data in the response variable have been well-developed in the literature. It is common practice to introduce a measurement error into SAR models. The measurement error serves to distinguish the noise component from the spatial process. However, the previous literature has not considered adding a measurement error to the SAR models with missing data. The maximum likelihood estimation for such models with large datasets is challenging and computationally expensive. This paper proposes two efficient likelihood-based estimation methods: the marginal maximum likelihood (ML) and expectation-maximisation (EM) algorithms for estimating SAR models with both measurement errors and missing data in the response variable. The spatial error model (SEM) and the spatial autoregressive model (SAM), two popular SAR model types, are considered. The missing data mechanism is assumed to follow missing at random (MAR). While naive calculation approaches lead to computational complexities of $O(n^3)$, where n is the total number of observations, our computational approaches for both the marginal ML and EM algorithms are designed to reduce the computational complexity. The performance of the proposed methods is investigated empirically using simulated and real datasets.

Keywords: spatial error model; spatial autoregressive model; measurement errors; marginal likelihood; expectation-maximisation algorithm; computational complexity

*Corresponding author: anjanaw@uow.edu.au

1 Introduction

Spatial regression models, often called simultaneous autoregressive (SAR) models, have been extensively used in the field of spatial statistics and econometrics (Anselin, 1988; LeSage and Pace, 2009; Gómez-Rubio et al., 2021). They extend linear regression models by accounting for spatial dependencies between different observations, making them useful for analysing spatial data. The dependence between observations in SAR models is captured by a spatial weight matrix (also known as a contiguity matrix); see Section 2 for further details on the spatial weight matrix. SAR models have been used in a wide range of applications, including ecology (Tognelli and Kelt, 2004; Ver Hoef et al., 2018), social sciences (Angrist and Lang, 2004; Ammermueller and Pischke, 2009), criminology (Glaeser et al., 1996), and financial market analysis (Longstaff, 2010).

There are two commonly used SAR models. The first is the spatial error model (SEM), where the spatial dependence is incorporated in the error term. The second is the spatial autoregressive model (SAM), where the spatial dependence is directly modelled in the equation for the response variable. In addition, more complex variants of SAR models exist. For example, the spatial Durbin model (SDM) (Anselin, 1988), which considers spatially correlated covariates, the SAM with heteroskedastic errors (Su, 2012), and SAR models with measurement errors (Burden et al., 2015; Suesse, 2018a). Furthermore, Dong and Harris (2015) expanded the applicability of SAR models accounting for geographically hierarchical data structures by incorporating regional-level random effects.

The estimation methods for SAR models with no missing values in the response variable are well-developed in the literature. For example, Ord (1975) introduced an efficient maximum likelihood (ML) estimation method by leveraging the eigenvalues of the spatial weight matrix. When dealing with sparse spatial weight matrices, efficient sparse Cholesky factorisation algorithms are employed (Pace and Barry, 1997; Pace, 1997; Pace and LeSage, 2004) within the ML estimation framework. Other popular estimation methods, such as the method of moments (MOM) (Kelejian and Prucha, 1999, 2001; Lee, 2007), Bayesian methods (Hepple, 1979; Anselin, 1988; LeSage, 1997), and instrumental variable (IV) estimation methods (Lee, 2003), have also been developed in the literature.

Having missing values in the response variable is common in practice. In this case, SAR models are often incorrectly estimated using the spatial weight matrix constructed from observed data only. The resulting estimates are biased and inconsistent (Wang and Lee, 2013; Benedetti et al., 2020). Estimation methods for SAR models under the missing at random (MAR) mechanism (Little and Rubin, 2019) have been explored extensively. For example, LeSage and Pace (2004) employed an iterative algorithm, which is similar to the expectation-

maximisation (EM) algorithm of Dempster et al. (1977), for estimating SAR models under MAR. Although the distribution of unobserved data given observed data is multivariate normal for SAR models, LeSage and Pace (2004) expressed it as a product of univariate normals to avoid computationally demanding matrix inversions. They approximated the mean of the SAM using the Neumann approximation for computational efficiency. These alterations made their algorithm invalid as a proper EM algorithm. Suesse and Zammit-Mangion (2017) addressed this issue by implementing minor adjustments to the algorithm proposed by LeSage and Pace (2004), thus transforming it into a valid EM algorithm. These adjustments included the consideration of exact terms rather than the approximations utilised by LeSage and Pace (2004). Suesse (2018b) introduced an alternative approach that directly maximises the marginal log-likelihood of the observed data. This is feasible because the marginal densities for observed data in both the SEM and SAM are available in closed form. This is achieved by integrating out the unobserved data from the joint density of observed and unobserved data, resulting in multivariate normal densities. Consequently, the marginal log-likelihood of the observed data for both SEM and SAM is available in closed form. They demonstrated that the marginal likelihood method is generally faster than the EM algorithm, and it does not suffer from the non-convergence issues associated with the EM algorithm. Kelejian and Prucha (2010) proposed an instrumental-variable (IV) estimator for SAM with missing responses, and Luo et al. (2021) developed an inverse probability weighting (IPW) based robust estimator for the SAM.

In spatial statistics, the observations are noisy measurements of the underlying spatial unobserved latent process. As a result, a measurement error is usually added to SAR models. For example, Bivand et al. (2015) and Gómez-Rubio et al. (2021) used the Integrated Nested Laplace Approximation (INLA) method of Rue et al. (2009) for estimating SAR models with measurement errors. Burden et al. (2015) and Suesse (2018a) used the ML estimation method for the SEM with measurement errors. Burden et al. (2015) utilised a spatial random effects model proposed by Cressie and Johannesson (2008) to efficiently approximate the covariance matrix of the SEM, while Suesse (2018a) employed efficient sparse matrix operations to estimate SEM and SAM with measurement errors.

Our article makes a number of contributions. First, we introduce SAR models that account for measurement errors, denoted as hierarchical SAR (H-SAR) models, with missing data in the response variable. Second, two ML estimation methods for estimating the H-SAR models with missing data are proposed. The first is the marginal ML estimation algorithm. The second is the EM algorithm. The EM algorithm involves two steps. The first step is known as the E-step, which computes the expectation of the complete log-likelihood with respect to the distribution of unobserved data given observed data and a fixed parameter

vector. The second step is known as the M-step, which maximises the expectation with respect to the model parameters; see Section 3.3 for further details.

Naive implementations of these two algorithms are expensive for large datasets, mainly due to the calculations of matrix inversions and determinants. The third contribution is to reduce computational costs. We propose efficient computational strategies that significantly lower computational complexity while preserving the accuracy of parameter estimates. For the marginal ML method, our computational approach significantly decreases the complexity from $O(n^3)$ to a maximum of either $O(n^{3/2}n_o)$ or $O(nn_o^2)$, where n is the total number of observations and n_o is the number of observed data. In the case of the EM method, the naive calculation of the terms leads to a complexity of $O(n^3)$. Our approach typically requires less than $O(n^3)$. Fourth, we establish a guideline for selecting the most appropriate method for different real-world scenarios.

The remainder of this paper is organised as follows. Section 2 discusses the hierarchical SAR models. Section 3 presents the two estimation methods: the marginal ML estimation method and the EM algorithm. Efficient computational strategies to reduce the complexity of the algorithm are also discussed. In Section 4, we evaluate the performance of the estimation methods using simulated datasets. Section 5 discusses the real data application. Section 6 concludes. The paper has an online supplement containing some further technical and empirical results.

2 SAR models

In this section, we first present standard SAR models. Then, hierarchical SAR models are considered.

2.1 Standard SAR Models

We consider the spatial autoregressive model (SAM),

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{W}\mathbf{y} + \mathbf{e}, \quad (2.1)$$

and the spatial error model (SEM),

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \\ \mathbf{u} &= \rho\mathbf{W}\mathbf{u} + \mathbf{e}, \end{aligned} \quad (2.2)$$

where $\mathbf{y} = (y_1, y_2, \dots, y_n)^\top$ is the vector of response variables observed at n spatial locations $s_1, \dots, s_n$, $\mathbf{X}$ is the $n \times r$ matrix of covariates, and $\mathbf{W}$ is the $n \times n$ spatial weight matrix. The error term $\mathbf{e}$ is assumed to follow a multivariate normal distribution with the mean vector $\mathbf{0}$ and covariance matrix $\sigma_y^2 \mathbf{I}_n$, where σ_y^2 is a variance parameter, and $\mathbf{I}_n$ is the $n \times n$ identity matrix, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_r)^\top$ is the vector of fixed effects parameters, and ρ is the spatial autocorrelation parameter (Anselin, 1988; Allison, 2001; LeSage and Pace, 2009).

Let W_{ij} be the i^{th} row and j^{th} column entry of the spatial weight matrix $\mathbf{W}$. The entry W_{ij} is non-zero if the unit i is the neighbour of the unit j . By definition, the diagonal of the spatial weight matrix $\mathbf{W}$ is zero, i.e. $W_{ii} = 0$. Several strategies for constructing $\mathbf{W}$ have been proposed in the literature (see Ord (1975); Anselin (1988); Kelley Pace and Barry (1997) for further details). The spatial weight matrix $\mathbf{W}$ has the dimension of $n \times n$, where n is the number of observations. We assume that $\mathbf{W}$ is sparse and symmetric, as commonly observed in numerous real-world scenarios. However, this may not necessarily be the case. Section 4.1 provides a detailed explanation of the methodology used to construct the spatial weight matrices that are employed in the simulation studies in this paper.

For the standard SEM and SAM, when the error vector $\mathbf{e}$ is normally distributed, the response variable $\mathbf{y}$ is also normally distributed with the mean vector $\boldsymbol{\mu}_y$ and the covariance matrix $\boldsymbol{\Sigma}_y$ given in Table 2.1.

Table 2.1: Expressions for $\boldsymbol{\mu}_y$, and $\boldsymbol{\Sigma}_y$ for SAM and SEM with $\mathbf{A} = \mathbf{I}_n - \rho \mathbf{W}$.

Term	SAM	SEM
$\boldsymbol{\mu}_y$	$\mathbf{A}^{-1} \mathbf{X} \boldsymbol{\beta}$	$\mathbf{X} \boldsymbol{\beta}$
$\boldsymbol{\Sigma}_y$	$\sigma_y^2 (\mathbf{A}^\top \mathbf{A})^{-1}$	$\sigma_y^2 (\mathbf{A}^\top \mathbf{A})^{-1}$

To ensure a valid covariance matrix in SAM and SEM, it is crucial that the correlation parameter ρ does not take on any of the values $\frac{1}{\lambda_{(1)}}, \frac{1}{\lambda_{(2)}}, \dots, \frac{1}{\lambda_{(n)}}$, where $\lambda_{(1)}, \lambda_{(2)}, \dots, \lambda_{(n)}$ are the eigenvalues of $\mathbf{W}$ sorted in ascending order (Li et al., 2012). When $\mathbf{W}$ is normalised either by row or column, with each row or column summing to 1, the range of ρ is constrained to $\frac{1}{\lambda_{(1)}} < \rho < 1$ (LeSage and Pace, 2009).

2.2 Hierarchical Simultaneous Autoregressive models

In many practical applications, the observations are a noisy manifestation of the underlying scientific process. A measurement error is usually added to the spatial statistical model. Given the observed spatial data $\mathbf{z} = (z_1, z_2, \dots, z_n)^\top$ at spatial locations $s_1, \dots, s_n$, the model is

$$z_i = y_i + \epsilon_i \quad i = 1, 2, \dots, n, \quad (2.3)$$

where y_i represents the spatial latent process, and ϵ_i is the additive measurement error at the spatial location s_i . We assume that ϵ_i follows a normal distribution with a mean of 0 and a variance σ_ϵ^2 .

The SAM with measurement errors is derived by substituting the latent process $\mathbf{y}$ in Equation (2.3) with the model described in Equation (2.1). Similarly, the SEM with measurement errors is obtained by replacing the latent process $\mathbf{y}$ in Equation (2.3) with the model specified in Equation (2.2).

In spatial statistics, the model given in Equation (2.3) is often called the data model (Arab et al., 2008). The data model characterises the distribution of the observed data, z_i , given the underlying hidden process, y_i , for $i = 1, \dots, n$. Conversely, the model describing the underlying latent process (the process of interest), such as the SAM in Equation (2.1) and the SEM in Equation (2.2), is usually called the process model (Arab et al., 2008). A joint spatial statistical model defined through a data model and a process model is often called a hierarchical spatial model (Arab et al., 2008). In this paper, we call SAR models with measurement errors as the hierarchical simultaneous autoregressive model (H-SAR), the SAM with measurement errors as the hierarchical SAM (H-SAM), and the SEM with measurement errors as the hierarchical SEM (H-SEM).

We now derive the marginal distribution of $\mathbf{z}$. Let $\boldsymbol{\phi} = (\boldsymbol{\beta}^\top, \rho, \omega, \theta)^\top$ be the vector of model parameters of the H-SAR model, where $\omega = \sigma_\epsilon^2$ and $\theta = \sigma_y^2/\sigma_\epsilon^2$. The joint distribution of $\mathbf{z}$ and $\mathbf{y}$ is

$$f(\mathbf{z}, \mathbf{y} \mid \boldsymbol{\phi}) = f(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\phi})f(\mathbf{y} \mid \boldsymbol{\phi}), \quad (2.4)$$

where $f(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\phi}) = \prod_{i=1}^n f(z_i \mid y_i, \boldsymbol{\phi})$, with $f(z_i \mid y_i, \boldsymbol{\phi})$ a normal distribution having mean y_i and variance σ_ϵ^2 . In addition, $f(\mathbf{y} \mid \boldsymbol{\phi})$ represents the density of standard SAR models outlined in Equations (2.1) and (2.2) for SEM and SAM, respectively. Then, the marginal distribution of $\mathbf{z}$ is obtained by integrating out $\mathbf{y}$ from the joint distribution of $\mathbf{z}$ and $\mathbf{y}$ in Equation (2.4), and is given by:

$$f(\mathbf{z} \mid \boldsymbol{\phi}) = \int f(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\phi})f(\mathbf{y} \mid \boldsymbol{\phi})d\mathbf{y}. \quad (2.5)$$

Since the measurement error distribution follows a multivariate normal distribution, the marginal distribution of $\mathbf{z}$ in H-SEM and H-SAM also follows a multivariate normal distribution with the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ given in Table 2.2. The constraints on the correlation parameter ρ in H-SAM and H-SEM, ensuring a valid covariance matrix $\boldsymbol{\Sigma}$, align with those for standard SEM and SAM, as elucidated by Suesse (2018a). The log-likelihood function of $\mathbf{z}$ in terms of the model parameters $\boldsymbol{\phi} = (\boldsymbol{\beta}^\top, \rho, \omega, \theta)^\top$ is

Table 2.2: Expressions for $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\mathbf{V}$ for H-SAM and H-SEM with $\mathbf{A} = \mathbf{I}_n - \rho\mathbf{W}$.

Term	H-SAM	H-SEM
$\boldsymbol{\mu}$	$\mathbf{A}^{-1}\mathbf{X}\boldsymbol{\beta}$	$\mathbf{X}\boldsymbol{\beta}$
$\boldsymbol{\Sigma} = \omega\mathbf{V}$	$\omega(\mathbf{I}_n + \theta(\mathbf{A}^\top\mathbf{A})^{-1})$	$\omega(\mathbf{I}_n + \theta(\mathbf{A}^\top\mathbf{A})^{-1})$
$\mathbf{V}$	$(\mathbf{I}_n + \theta(\mathbf{A}^\top\mathbf{A})^{-1})$	$(\mathbf{I}_n + \theta(\mathbf{A}^\top\mathbf{A})^{-1})$
$\mathbf{M} = \mathbf{V}^{-1}$	$(\mathbf{I}_n + \theta(\mathbf{A}^\top\mathbf{A})^{-1})^{-1}$	$(\mathbf{I}_n + \theta(\mathbf{A}^\top\mathbf{A})^{-1})^{-1}$

$$\log f(\mathbf{z}; \omega, \theta, \rho, \boldsymbol{\beta}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\omega) + \frac{1}{2}\log|\mathbf{V}^{-1}| - \frac{1}{2\omega}\mathbf{r}^\top\mathbf{V}^{-1}\mathbf{r}, \quad (2.6)$$

where $\mathbf{r} = \mathbf{z} - \boldsymbol{\mu}$ is the vector of residuals. Expressions for the $\mathbf{V}$ and $\boldsymbol{\mu}$ are given in Table 2.2.

Suesse (2018a) developed an efficient ML estimation method for estimating H-SAR models without missing data. The next section discusses the proposed ML estimation methods for estimating the H-SAR models with missing responses under the missing at random (MAR) mechanism.

3 Estimation Methods

Section 3.1 discusses the H-SAR under the MAR mechanism, Section 3.2 discusses the marginal maximum likelihood (marginal ML) estimation method, and Section 3.3 discusses the EM algorithm.

3.1 Hierarchical simultaneous autoregressive models under MAR

Let $\mathbf{z}_o$ be the subset of $\mathbf{z}$ with n_o units, and $\mathbf{z}_u$ be the subset of $\mathbf{z}$ with n_u units, where n_o and n_u are the numbers of observed and missing response variables, respectively. The complete-data vector is denoted by $\mathbf{z} = (\mathbf{z}_o^\top, \mathbf{z}_u^\top)^\top$. The matrices $\mathbf{X}$ and $\mathbf{W}$ are divided into distinct parts as follows:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_o \\ \mathbf{X}_u \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} \mathbf{W}_{oo} & \mathbf{W}_{ou} \\ \mathbf{W}_{uo} & \mathbf{W}_{uu} \end{pmatrix}, \quad (3.1)$$

where $\mathbf{X}_o$ and $\mathbf{X}_u$ are the corresponding matrices of covariates for the observed and unobserved data, respectively, and $\mathbf{W}_{oo}$, $\mathbf{W}_{ou}$, $\mathbf{W}_{uo}$, and $\mathbf{W}_{uu}$ represent the sub-matrices of $\mathbf{W}$.

When dealing with missing data in the response variable, the conventional approach of specifying the H-SAR model based solely on observed locations becomes inadequate. We

refer to this model specification as *observed data models* (ODMs). The ODM specification constructs the spatial weight matrix ($\mathbf{W}$) exclusively using the location information from observed locations or units only. They use $\mathbf{z}_o$, $\mathbf{X}_o$, and $\mathbf{W}_{oo}$ to specify the H-SAR model. As a result, the ODM uses inaccurate specifications of the correlation structure of the marginal density of observed data $\mathbf{z}_o$ (LeSage and Pace, 2004), leading to inconsistent and biased parameter estimates (Wang and Lee, 2013; Benedetti et al., 2020).

We now briefly explain the MAR mechanism. Consider the vector $\mathbf{m}$ of length n containing 1's and 0's. If an element in $\mathbf{z}$ is missing, then the corresponding element in $\mathbf{m}$ is 1 and 0, otherwise. The vector $\mathbf{m}$ is often called the missing data indicator vector (Little and Rubin, 2019). The probability distribution of $\mathbf{m}$ given $\mathbf{z}$ is denoted by $p(\mathbf{m} \mid \mathbf{z}, \boldsymbol{\psi})$, where $\boldsymbol{\psi}$ is the vector of parameters that governs the missing data generating process. Under the MAR mechanism, this probability distribution is independent of the unobserved data but it depends on the observed data $\mathbf{z}_o$, i.e., $p(\mathbf{m} \mid \mathbf{z}, \boldsymbol{\psi}) = p(\mathbf{m} \mid \mathbf{z}_o, \boldsymbol{\psi})$ (Rubin, 1976). Under the MAR mechanism, assuming that $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ are distinct, the ML estimation method focuses on maximising the marginal likelihood of the observed data $\mathbf{z}_o$, ignoring the missing data mechanism $\mathbf{m}$; see Little and Rubin (2019) for the detailed proof.

To compute the marginal likelihood of $\mathbf{z}_o$, the unobserved data $\mathbf{z}_u$ must be integrated out from the complete data density of $\mathbf{z}$,

$$f(\mathbf{z}_o; \boldsymbol{\phi}) = \int f(\mathbf{z}; \boldsymbol{\phi}) d\mathbf{z}_u, \quad (3.2)$$

where $f(\mathbf{z}; \boldsymbol{\phi})$ is the complete data density of $\mathbf{z}$.

Unlike the ODM, the accurate specification of the marginal density of $\mathbf{z}_o$ depends on the weight matrix $\mathbf{W}$, constructed based on the complete set of n spatial locations. We call this model the *complete locations model* (CLM). We employ two approaches to maximise the marginal likelihood in Equation (3.2). The first approach involves directly maximising the marginal log-likelihood of the observed data. In the second approach, we utilise the EM algorithm. In Section 4.2, we compare the two approaches in terms of accuracy and computational cost.

3.2 Marginal maximum likelihood estimation method

This section discusses the proposed marginal ML estimation method. Since $\mathbf{z}$ is a multivariate normal random variable, $\mathbf{z}_o$ also follows a multivariate normal distribution with the mean vector $\boldsymbol{\mu}_o$ and the covariance matrix $\boldsymbol{\Sigma}_{oo} = \omega \mathbf{V}_{oo}$ (Petersen et al., 2008); see Equation (3.4). To compute the log-likelihood of the marginal distribution of $\mathbf{z}_o$, replace $\mathbf{V}$ with $\mathbf{V}_{oo}$, $\boldsymbol{\mu}$ with $\boldsymbol{\mu}_o$, $\mathbf{z}$ with $\mathbf{z}_o$, and n with n_o in Equation (2.6). This yields the following expression for the

marginal log-likelihood of $\mathbf{z}_o$:

$$\log f(\mathbf{z}_o; \omega, \theta, \rho, \boldsymbol{\beta}) = -\frac{n_o}{2} \log(2\pi) - \frac{n_o}{2} \log(\omega) + \frac{1}{2} \log |\mathbf{V}_{oo}^{-1}| - \frac{1}{2\omega} \mathbf{r}_o^\top \mathbf{V}_{oo}^{-1} \mathbf{r}_o, \quad (3.3)$$

where $\mathbf{r}_o = \mathbf{z}_o - \boldsymbol{\mu}_o$, and the $\boldsymbol{\mu}$ vector and the matrix $\mathbf{V}$ are partitioned as

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_o \\ \boldsymbol{\mu}_u \end{pmatrix}, \quad \mathbf{V} = \begin{pmatrix} \mathbf{V}_{oo} & \mathbf{V}_{ou} \\ \mathbf{V}_{uo} & \mathbf{V}_{uu} \end{pmatrix}. \quad (3.4)$$

Direct maximisation of Equation (3.3) is challenging, and to reduce the dimension of the maximisation problem, we obtain the concentrated marginal log-likelihood by substituting the ML estimates of $\hat{\omega}$ and $\hat{\boldsymbol{\beta}}$ into Equation (3.3).

We now define a new matrix $\tilde{\mathbf{X}}$ to make the notation below simpler. For H-SEM, the matrix $\tilde{\mathbf{X}} = \mathbf{X}$, and for H-SAM the matrix $\tilde{\mathbf{X}} = \mathbf{A}^{-1} \mathbf{X}$. By maximising Equation (3.3) with respect to $\boldsymbol{\beta}$ and ω (taking partial derivatives and setting to zero) while holding θ and ρ fixed, we obtain the closed form ML estimates for $\boldsymbol{\beta}$ and ω as follows:

$$\hat{\boldsymbol{\beta}}(\rho, \theta) = \left(\tilde{\mathbf{X}}_o^\top \mathbf{V}_{oo}^{-1} \tilde{\mathbf{X}}_o \right)^{-1} \tilde{\mathbf{X}}_o^\top \mathbf{V}_{oo}^{-1} \mathbf{z}_o, \quad (3.5)$$

and

$$\hat{\omega}(\rho, \theta) = \frac{\mathbf{r}_o^\top \mathbf{V}_{oo}^{-1} \mathbf{r}_o}{n_o}. \quad (3.6)$$

The proof is given in Section S4 of the online supplement.

By substituting $\hat{\boldsymbol{\beta}}(\rho, \theta)$, and $\hat{\omega}(\rho, \theta)$ in Equations (S4.3) and (S5.5) in the log-likelihood in Equation (3.3), the concentrated marginal log-likelihood L_c has the form:

$$L_c(\theta, \rho) = c - \frac{n_o}{2} \log(\hat{\omega}) + \frac{1}{2} \log |\mathbf{V}_{oo}^{-1}|, \quad (3.7)$$

where $c = -\frac{n_o}{2} \log(2\pi) - \frac{n_o}{2}$ is a constant.

To obtain the ML estimates for θ and ρ , we maximise the concentrated marginal log-likelihood in Equation (3.7) using `optim()` function in R. Upon examining Equations (S4.3), (S5.5), and (3.7), it becomes clear that three computationally expensive terms are involved in the optimisation process. They are $\log |\mathbf{V}_{oo}^{-1}|$, $\mathbf{r}_o^\top \mathbf{V}_{oo}^{-1} \mathbf{r}_o$, and $\tilde{\mathbf{X}}_o^\top \mathbf{V}_{oo}^{-1} \tilde{\mathbf{X}}_o$. Therefore, efficient calculations of matrix $\mathbf{V}_{oo}$ are crucial for computing these terms. Below, we propose two computational approaches to calculate the sub-matrix $\mathbf{V}_{oo}$.

We call the first approach the *direct method* since it involves directly extracting the

sub-matrix $\mathbf{V}_{oo}$ from the larger matrix $\mathbf{V}$ as

$$\mathbf{V}_{oo} = [(\mathbf{I}_n + \theta(\mathbf{A}^\top \mathbf{A})^{-1})]_{oo}. \quad (3.8)$$

The analysis of computational complexity of calculating $\mathbf{V}_{oo}$ using the direct method is discussed in Section 4.1.1. The direct method gives an overall computational complexity of $O(n^2)$, where n is the total number of observations; see Section 4.1 for further details. For the second method, the $\mathbf{V}_{oo}$ sub-matrix is obtained by re-expressing it with an additional sparse matrix $\mathbf{B}_o$. This approach leverages the sparsity of $\mathbf{W}$ (through $\mathbf{A}$), and $\mathbf{B}_o$ to simplify the computation of $\mathbf{V}_{oo}$. First, the matrix $\mathbf{V}$ in Table 2.2 is written as $\mathbf{V} = \mathbf{A}^{-1} (\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n) (\mathbf{A}^\top)^{-1}$. The proof is given in Section S5.2 in the online supplement. Then the sub-matrix $\mathbf{V}_{oo}$ is obtained by $\mathbf{V}_{oo} = ((\mathbf{A}^\top)^{-1}\mathbf{B}_o^\top)^\top (\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n) ((\mathbf{A}^\top)^{-1}\mathbf{B}_o^\top)$, where $\mathbf{B}_o$ is a sparse matrix such that $\mathbf{B}_o = [\mathbf{I}_o | \mathbf{0}]$, where $\mathbf{I}_o$ is the $n_o \times n_o$ identity matrix, and $\mathbf{0}$ is the $n_o \times n_u$ zero matrix. For convenience of notation, let $(\mathbf{A}^\top)^{-1}\mathbf{B}_o^\top = \mathbf{A}_{\mathbf{B}_o}^{-1}$. Then, the matrix $\mathbf{V}_{oo}$ is

$$\mathbf{V}_{oo} = (\mathbf{A}_{\mathbf{B}_o}^{-1})^\top (\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}) \mathbf{A}_{\mathbf{B}_o}^{-1}. \quad (3.9)$$

The calculation of $\mathbf{A}_{\mathbf{B}_o}^{-1}$ is computationally more efficient and stable compared to the direct calculation of $\mathbf{V}_{oo}$ using the inverse of $\mathbf{A}^\top \mathbf{A}$ as in Equation (3.8). The advantages are twofold. First, $\mathbf{A}$ is sparser than $\mathbf{A}^\top \mathbf{A}$, which reduces the computational complexity as there are fewer nonzero elements to consider. This can significantly speed up the calculations, especially for large sparse matrices. Second, to calculate $\mathbf{A}_{\mathbf{B}_o}^{-1}$, we do not need to explicitly compute $\mathbf{A}^{-1}$. Instead, we can use a system of linear equations to solve for it, which is generally faster and more stable numerically, particularly for sparse matrices. Therefore, we solve the system $\mathbf{A}^\top \mathbf{A}_{\mathbf{B}_o}^{-1} = \mathbf{B}_o^\top$, for $\mathbf{A}_{\mathbf{B}_o}^{-1}$ and, then, calculate $\mathbf{V}_{oo}$. We call the approach the *parameterisation method*.

In Section 4.1, we thoroughly examine the computational complexities associated with both methods. Table 4.1 outlines the computational complexities linked to the fundamental operations needed for calculating $\mathbf{V}_{oo}$ using the expression in Equation (3.8). Table 4.2 summarizes the complexities associated with the essential operations required for computing $\mathbf{V}_{oo}$ based on the expression in Equation (3.9).

3.3 EM algorithm

The EM algorithm, introduced by Dempster et al. (1977), is a widely used iterative algorithm for parameter estimation in the presence of missing data (Lauritzen, 1995; Enders, 2003) or latent variables (Bishop, 1998). The EM algorithm maximises Equation 3.2 by iteratively

estimating the missing data or the latent variables and updating the model parameters.

The EM algorithm consists of two main steps: the E (expectation) step and the M (maximisation) step. In the E-step, the algorithm computes the expected value of the complete-data log-likelihood in Equation (2.6), with respect to the conditional density of $\mathbf{z}_u$ given $\mathbf{z}_o$ and a fixed parameter vector $\boldsymbol{\phi}'$. This expectation is denoted by $E_{u|o,\boldsymbol{\phi}'}(\cdot)$. In the subsequent M-step, this expectation is maximised with respect to the vector of parameters $\boldsymbol{\phi}$. In summary, the EM algorithm for H-SAR models proceeds as follows:

- Starts with assigning initial values for $\boldsymbol{\phi}' = \boldsymbol{\phi}'_0$.
- E-step: Calculate

$$\begin{aligned} E_{u|o,\boldsymbol{\phi}'} [\log f(\mathbf{z}; \boldsymbol{\phi})] &= Q(\boldsymbol{\phi} | \boldsymbol{\phi}') \\ &= \int \log f(\mathbf{z}; \boldsymbol{\phi}) f(\mathbf{z}_u | \mathbf{z}_o, \boldsymbol{\phi} = \boldsymbol{\phi}') d\mathbf{z}_u, \end{aligned} \quad (3.10)$$

where $f(\mathbf{z}_u | \mathbf{z}_o, \boldsymbol{\phi} = \boldsymbol{\phi}')$ denotes the conditional density of $\mathbf{z}_u$ given $\mathbf{z}_o$ and a fixed parameter vector $\boldsymbol{\phi}'$.

It is well known that as $\mathbf{z}$ follows a multivariate normal distribution, the conditional distribution of $\mathbf{z}_u | \mathbf{z}_o$ also follows a multivariate normal distribution (Petersen et al., 2008), characterized by the mean vector $\boldsymbol{\mu}_{u|o}$ and the covariance matrix $\boldsymbol{\Sigma}_{u|o}$ given by

$$\boldsymbol{\mu}_{u|o} = \boldsymbol{\mu}_u + \mathbf{V}_{uo} \mathbf{V}_{oo}^{-1} (\mathbf{z}_o - \boldsymbol{\mu}_o), \quad (3.11)$$

and

$$\boldsymbol{\Sigma}_{u|o} = \omega \{ \mathbf{V}_{uu} - \mathbf{V}_{uo} \mathbf{V}_{oo}^{-1} \mathbf{V}_{ou} \}. \quad (3.12)$$

- M-step: Maximise $Q(\boldsymbol{\phi} | \boldsymbol{\phi}')$ with respect to the parameter vector $\boldsymbol{\phi}$.
- Iterate the E-step and the M-step until convergence.

By substituting $\mathbf{M}$ for $\mathbf{V}^{-1}$ in the complete-data log-likelihood of $\mathbf{z}$ in Equation (2.6), we obtain the following equation:

$$\log f(\mathbf{z}; \omega, \theta, \rho, \boldsymbol{\beta}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\omega) + \frac{1}{2} \log |\mathbf{M}| - \frac{1}{2\omega} \mathbf{r}^\top \mathbf{M} \mathbf{r}. \quad (3.13)$$

By taking the conditional expectation of Equation (3.13) with respect to $f(\mathbf{z}_u | \mathbf{z}_o, \boldsymbol{\phi} = \boldsymbol{\phi}')$,

we obtain

$$Q(\boldsymbol{\phi} \mid \boldsymbol{\phi}') = E_{u|o,\boldsymbol{\phi}'}(\log f(\mathbf{z}; \omega, \theta, \rho, \boldsymbol{\beta})) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\omega) + \frac{1}{2}\log|\mathbf{M}| - \frac{1}{2}E_{u|o,\boldsymbol{\phi}'}(\mathbf{r}^\top \mathbf{M} \mathbf{r}). \quad (3.14)$$

For simplicity, we no longer explicitly denote the dependence of the expectation operator on the fixed parameter vector $\boldsymbol{\phi}'$. i.e. $E_{u|o}(\cdot) = E_{u|o,\boldsymbol{\phi}'}(\cdot)$. Suesse and Zammit-Mangion (2017) simplified the conditional expectation of the final term in Equation (3.14) for the standard SAR models (for SAM and SEM). It is important to note that in the case of H-SAR, the covariance structure of the distribution of $\mathbf{z}$ deviates from that of standard SAR models. Consequently, the matrix $\mathbf{M}$ and the vector $\mathbf{r}$ exhibit distinct forms in H-SAR models compared to standard SAR models. Utilising these derivations, we can express the term $Q(\boldsymbol{\phi} \mid \boldsymbol{\phi}')$ for H-SAR as:

$$Q(\boldsymbol{\phi} \mid \boldsymbol{\phi}') = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\omega) + \frac{1}{2}\log|\mathbf{M}| - \frac{\mathbf{r}_{u|o}^\top \mathbf{M} \mathbf{r}_{u|o} + \omega' \text{tr}\{\mathbf{M}_{uu}(\theta', \rho')^{-1} \mathbf{M}_{uu}(\theta, \rho)\}}{2\omega}, \quad (3.15)$$

where $\text{tr}\{\cdot\}$ is the matrix trace¹, $\mathbf{r}_{u|o} = E_{u|o}(\mathbf{z}) - \boldsymbol{\mu}$, $E_{u|o}(\mathbf{z}) = (\mathbf{z}_o^\top, \boldsymbol{\mu}_{u|o}^\top)^\top$, and the matrix $\mathbf{M}$ is partitioned as

$$\mathbf{M} = \begin{pmatrix} \mathbf{M}_{oo} & \mathbf{M}_{ou} \\ \mathbf{M}_{uo} & \mathbf{M}_{uu} \end{pmatrix}. \quad (3.16)$$

In the M-step, we maximise the term $Q(\boldsymbol{\phi} \mid \boldsymbol{\phi}')$ in Equation (3.15) with respect to the vector of parameters $\boldsymbol{\phi} = (\rho, \boldsymbol{\beta}^\top, \theta, \omega)^\top$. By differentiating Equation (3.15) with respect to $\boldsymbol{\beta}$ and ω , and setting the derivatives to zero, we obtain closed-form expressions for the ML estimators of these parameters in terms of θ and ρ :

$$\hat{\boldsymbol{\beta}}(\rho, \theta) = \left(\tilde{\mathbf{X}}^\top \mathbf{M} \tilde{\mathbf{X}} \right)^{-1} \tilde{\mathbf{X}}^\top \mathbf{M} E_{u|o}(\mathbf{z}), \quad (3.17)$$

and

$$\hat{\omega}(\rho, \theta) = \frac{\mathbf{r}_{u|o}^\top \mathbf{M} \mathbf{r}_{u|o}}{n}. \quad (3.18)$$

The proof is given in Section S5.1 of the online supplement.

By substituting $\hat{\boldsymbol{\beta}}(\rho, \theta)$ from Equation (3.17) and $\hat{\omega}(\rho, \theta)$ from Equation (3.18) into Equa-

¹The trace of a square matrix $\mathbf{M}$, denoted as $\text{tr}\{\mathbf{M}\}$, is the sum of its diagonal elements: $\text{tr}\{\mathbf{M}\} = \sum_{i=1}^n M_{ii}$.

tion (3.15), we simplify $Q(\phi \mid \phi')$ to

$$Q(\theta, \rho \mid \phi') = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\hat{\omega}(\rho, \theta)) + \frac{1}{2}\log|\mathbf{M}| - \frac{n}{2}, \quad (3.19)$$

which is a function of θ and ρ . We obtain the ML estimates for θ and ρ by maximising (3.19) using the `optim()` function in R.

While maximising the expected log-likelihood in Equation (3.19), we encounter four computationally challenging terms: $\log|\mathbf{M}|$, $\tilde{\mathbf{X}}^\top \mathbf{M} \tilde{\mathbf{X}}$, $\mathbf{r}_{u|o}^\top \mathbf{M} \mathbf{r}_{u|o}$, and $\boldsymbol{\mu}_{u|o}$. Importantly, it is worth noting that it is unnecessary to compute the entire matrix $\mathbf{M} = (\mathbf{I}_n + \theta(\mathbf{A}^\top \mathbf{A})^{-1})^{-1}$ for any of these terms, as this would entail two matrix inversions. Instead, we only need specific terms related to $\mathbf{M}$ and its sub-matrices, as detailed in the following subsections.

3.3.1 Calculate $\log|\mathbf{M}|$

We have previously rewritten $\mathbf{V} = \mathbf{A}^{-1}(\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n)(\mathbf{A}^\top)^{-1}$. Since $\mathbf{M}$ is the inverse of $\mathbf{V}$, we express it as follows:

$$\mathbf{M} = \mathbf{V}^{-1} = \mathbf{A}^\top (\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n)^{-1} \mathbf{A}, \quad (3.20)$$

and the logarithm of the matrix $\mathbf{M}$ is

$$\log|\mathbf{M}| = \log|\mathbf{A}^\top| - \log|\mathbf{A}\mathbf{A}^\top + \mathbf{I}_n\theta| + \log|\mathbf{A}| = -\log|\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n| + \log|\mathbf{A}\mathbf{A}^\top|; \quad (3.21)$$

see Section S5.2 of the online supplement for the complete proof.

Calculating the determinant of $\mathbf{A}\mathbf{A}^\top$ using its Cholesky factorisation (denoted as $\mathbf{L}_{\mathbf{A}\mathbf{A}^\top}$) is computationally efficient since the matrix $\mathbf{A}\mathbf{A}^\top$ is sparse. Next, we can compute the determinant of $\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n$ by using `update()` function in R. The update function takes the Cholesky factor $\mathbf{L}_{\mathbf{A}\mathbf{A}^\top}$, the matrix $\mathbf{A}\mathbf{A}^\top$, and θ as the inputs, and it outputs $\mathbf{L}_{\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n}$, the Cholesky factor of $\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n$. This approach lowers the computational requirements of calculating the logarithm of the determinant of $\mathbf{M}$ in Equation (S5.7).

3.3.2 Computing the terms $\tilde{\mathbf{X}}^\top \mathbf{M} \tilde{\mathbf{X}}$ and $\mathbf{r}_{u|o}^\top \mathbf{M} \mathbf{r}_{u|o}$

The terms $\tilde{\mathbf{X}}^\top \mathbf{M} \tilde{\mathbf{X}}$ and $\mathbf{r}_{u|o}^\top \mathbf{M} \mathbf{r}_{u|o}$ play a crucial role in computing $\hat{\beta}(\rho, \theta)$ and $\hat{\omega}(\rho, \theta)$ in Equations (3.18) and (3.19). To facilitate the more efficient calculation, we introduce two additional terms: $\mathbf{C} = (\mathbf{L}_{\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n})^{-1} \tilde{\mathbf{X}}_{\mathbf{A}}$, and $\mathbf{c} = (\mathbf{L}_{\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n})^{-1} \mathbf{z}_{\mathbf{A}_{u|o}}$, where $\tilde{\mathbf{X}}_{\mathbf{A}} = \mathbf{A} \tilde{\mathbf{X}}$, and

$\mathbf{z}_{\mathbf{A}_{u|o}} = \mathbf{A}E_{u|o}(\mathbf{z})$. Subsequently, the expressions for $\tilde{\mathbf{X}}^\top \mathbf{M}\tilde{\mathbf{X}}$ and $\mathbf{r}_{u|o}^\top \mathbf{M}\mathbf{r}_{u|o}$ are given by:

$$\tilde{\mathbf{X}}^\top \mathbf{M}\tilde{\mathbf{X}} = \left((\mathbf{L}_{\mathbf{A}\mathbf{A}^\top + \theta \mathbf{I}_n})^{-1} \tilde{\mathbf{X}}_{\mathbf{A}} \right)^\top (\mathbf{L}_{\mathbf{A}\mathbf{A}^\top + \theta \mathbf{I}_n})^{-1} \tilde{\mathbf{X}}_{\mathbf{A}} = \mathbf{C}^\top \mathbf{C}, \quad (3.22)$$

and

$$\mathbf{r}_{u|o}^\top \mathbf{M}\mathbf{r}_{u|o} = E_{u|o}(\mathbf{z})^\top \mathbf{M}E_{u|o}(\mathbf{z}) - 2E_{u|o}(\mathbf{z})^\top \mathbf{M}\tilde{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{M}\tilde{\mathbf{X}}\boldsymbol{\beta}. \quad (3.23)$$

We can show that $E_{u|o}(\mathbf{z})^\top \mathbf{M}E_{u|o}(\mathbf{z}) = \mathbf{c}^\top \mathbf{c}$ and $E_{u|o}(\mathbf{z})^\top \mathbf{M}\tilde{\mathbf{X}} = \mathbf{c}^\top \mathbf{C}$ (see Section S5.1 of the online supplement for the proof). By substituting these terms and the expression for $\tilde{\mathbf{X}}^\top \mathbf{M}\tilde{\mathbf{X}}$ from Equation (S5.10) into Equation (S5.11), we obtain a much simpler expression for $\mathbf{r}_{u|o}^\top \mathbf{M}\mathbf{r}_{u|o}$ as:

$$\mathbf{r}_{u|o}^\top \mathbf{M}\mathbf{r}_{u|o} = \mathbf{c}^\top \mathbf{c} - 2\mathbf{c}^\top \mathbf{C}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}^\top \mathbf{C}^\top \mathbf{C}\hat{\boldsymbol{\beta}}. \quad (3.24)$$

By substituting $\tilde{\mathbf{X}}^\top \mathbf{M}\tilde{\mathbf{X}}$ in Equation (S5.10), and the simplified expression for $E_{u|o}(\mathbf{z})^\top \mathbf{M}\tilde{\mathbf{X}}$ into Equation (3.17), we obtain a simpler expression for $\hat{\boldsymbol{\beta}}(\rho, \theta)$ as:

$$\hat{\boldsymbol{\beta}}(\rho, \theta) = (\mathbf{C}^\top \mathbf{C})^{-1} \mathbf{C}^\top \mathbf{c}, \quad (3.25)$$

and by substituting $\mathbf{r}_{u|o}^\top \mathbf{M}\mathbf{r}_{u|o}$ in Equation (3.24) into Equation (3.18), we obtain a simpler expression for $\hat{\omega}(\rho, \theta)$ as

$$\hat{\omega}(\rho, \theta) = \frac{\mathbf{c}^\top \mathbf{c} - 2\mathbf{c}^\top \mathbf{C}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}^\top \mathbf{C}^\top \mathbf{C}\hat{\boldsymbol{\beta}}}{n} = \frac{\mathbf{c}^\top \mathbf{c} - \mathbf{c}^\top \mathbf{C}\hat{\boldsymbol{\beta}}}{n}. \quad (3.26)$$

3.3.3 Computing the conditional mean $\boldsymbol{\mu}_{u|o}$

To calculate the term $\mathbf{c}$, the vector $E_{u|o}(\mathbf{z})$ is required, which involves computing $\boldsymbol{\mu}_{u|o}$. Calculating $\boldsymbol{\mu}_{u|o}$ as in Equation (3.11), requires performing a sparse matrix inversion for calculating $\mathbf{V}$, and a dense matrix inversion for calculating $\mathbf{V}_{oo}^{-1}$. LeSage and Pace (2004) provide an alternative and simpler expression for $\boldsymbol{\mu}_{u|o}$ as follows:

$$\boldsymbol{\mu}_{u|o} = \boldsymbol{\mu}_u - \mathbf{M}_{uu}^{-1} \mathbf{M}_{uo}(\mathbf{z}_o - \boldsymbol{\mu}_o). \quad (3.27)$$

Even though Equation (3.27) contains sub-matrices of $\mathbf{M}$, it is possible to obtain these sub-matrices without explicitly calculating the full matrix $\mathbf{M}$. Let us define a new, sparse matrix $\mathbf{B}_u$ such that $\mathbf{B}_u = [\mathbf{0} \mid \mathbf{I}_u]$, where $\mathbf{I}_u$ is the $n_u \times n_u$ identity matrix, and $\mathbf{0}$ is the $n_u \times n_o$ zero matrix. Using $\mathbf{B}_u$, the matrix $\mathbf{M}_{uu}$ is expressed as

$$\mathbf{M}_{uu} = \mathbf{B}_u \mathbf{M} \mathbf{B}_u^\top = \mathbf{B}_u \mathbf{A}^\top (\mathbf{A} \mathbf{A}^\top + \mathbf{I}_n \theta)^{-1} \mathbf{A} \mathbf{B}_u^\top, \quad (3.28)$$

and the matrix $\mathbf{M}_{ou}$ is

$$\mathbf{M}_{ou} = \mathbf{B}_o \mathbf{M} \mathbf{B}_u^\top = \mathbf{B}_o \mathbf{A}^\top (\mathbf{A} \mathbf{A}^\top + \mathbf{I}_n \theta)^{-1} \mathbf{A} \mathbf{B}_u^\top, \quad (3.29)$$

where $\mathbf{B}_o$ is a sparse matrix such that $\mathbf{B}_o = [\mathbf{I}_o | \mathbf{0}]$ with $\mathbf{I}_o$ is the $n_o \times n_o$ identity matrix, and $\mathbf{0}$ is the $n_o \times n_u$ zero matrix.

By examining Equations (3.28) and (3.29), we observe that the term $(\mathbf{A} \mathbf{A}^\top + \mathbf{I}_n \theta)^{-1} \mathbf{A} \mathbf{B}_u^\top$ is common to both equations. Moreover, we can efficiently compute $(\mathbf{A} \mathbf{A}^\top + \mathbf{I}_n \theta)^{-1} \mathbf{A} \mathbf{B}_u^\top$ using the `solve()` function in R since the Cholesky factors of $\mathbf{A} \mathbf{A}^\top + \mathbf{I}_n \theta$ are available from previous computations.

Upon a comprehensive examination of the terms crucial for implementing the EM algorithm, it becomes apparent that the algorithm's overall computational complexity is predominantly influenced by the computation of the log determinant term, $\log|\mathbf{M}|$, and the conditional mean in Equation (3.27), which depends on $\mathbf{M}_{uu}$ and $\mathbf{M}_{ou}$ (see Table 4.3 for details of the computational complexity of computing $\boldsymbol{\mu}_{u|o}$ using $\mathbf{M}_{ou}$ and $\mathbf{M}_{uu}$). It is clear that the naive calculation of these terms results in computational complexity of $O(n^3)$, given that the computation of $\mathbf{M}$ involves the inverse of the dense matrix $\mathbf{M} = (\mathbf{I}_n + \theta(\mathbf{A}^\top \mathbf{A})^{-1})^{-1}$. Our proposed approaches reduce the complexity, ensuring it remains below this bound in most real-world scenarios. Further analysis of the computational complexity of the two proposed algorithms is discussed in the subsequent section.

4 Simulation study

In Section 3.2, we show that the computation of the matrix $\mathbf{V}_{oo}$ poses a significant computational issue when implementing the marginal ML method. Similarly, the calculations involving $\mathbf{M}_{uu}$, $\mathbf{M}_{ou}$, and $\log|\mathbf{M}|$ are computationally demanding for the implementation of the EM algorithm (see Section 3.3).

Section 4.1 presents simulation results for evaluating the proposed computational approaches to reduce the complexities associated with computing the demanding terms in marginal ML and EM algorithms. Then, in Section 4.2, we discuss the accuracy and computational cost of the marginal ML and EM methods using simulated datasets.

4.1 Computational complexities

Parameter estimation of SAR models involves a series of matrix and vector manipulations, and many of these operations, such as matrix multiplication, inversion, and factorisation,

have a computational complexity of $O(n^3)$ for dense matrices. Consequently, the overall complexity of SAR estimation algorithms is $O(n^3)$ if naive computation methods are utilised.

In practical applications of SAR models, sparse weight matrices ($\mathbf{W}$) are commonly utilised. Matrix operations involving sparse matrices tend to be significantly faster than those involving dense matrices. The computational complexity of matrix multiplication and factorisations, such as the Cholesky factorisation for sparse matrices, depends on the number of nonzero values within the sparse matrix. For instance, while the Cholesky factorisation for dense matrices typically exhibits a complexity of $O(n^3)$, the complexity for sparse precision matrices is often reduced to $O(n)$ for temporal Gaussian Markov random fields (GMRFs), $O(n^{3/2})$ for spatial GMRFs, and $O(n^2)$ for spatio-temporal GMRFs, as discussed by Rue and Held (2005).

In our simulation study, we utilise sparse weight matrices ($\mathbf{W}$), allowing us to take advantage of the sparse matrix operations offered by the R package *Matrix* (Bates et al., 2022) for implementing our proposed methods. For example, the EM algorithm described in Section 3.3 necessitates computing the Cholesky factors of the matrix $\mathbf{A}\mathbf{A}^\top$. The *Matrix* package offers an efficient function for this task, which has a computational complexity of $O(n^{3/2})$ when $\mathbf{A}\mathbf{A}^\top$ is sparse. In contrast, the standard R `chol()` function entails a complexity of $O(n^3)$, even when $\mathbf{A}\mathbf{A}^\top$ is sparse. Likewise, for sparse matrix inversions and solving systems with sparse matrices, we rely on the efficient functions provided by the *Matrix* package. Further details regarding the functionality of these functions can be found in Section S1 of the online supplement.

We now discuss the theoretical and empirical computational complexities of the expensive matrix operations used in the marginal ML and EM methods. The theoretical computational complexity is established by analysing the structure of matrices involved in the specific operation. To derive the theoretical complexity of a specific matrix operation, we need to compute the number of scalar operations, known as flops (floating-point operations). For instance, consider the multiplication of two $n \times n$ dense matrices. This operation comprises n^3 scalar multiplications and $n^2(n - 1)$ scalar additions. In total, the computation involves $2n^3 - n^2$ flops. As a consequence, the overall computational complexity is $O(n^3)$.

On the other hand, the empirical complexity is estimated based on simulations. To illustrate, we now consider determining the empirical complexity of the matrix multiplication $\mathbf{A}\mathbf{A}^\top$. We first define a $\sqrt{n} \times \sqrt{n}$ grid (see Figure 4.1), and then followed by the construction of the spatial weight matrix $\mathbf{W}$ and the computation of $\mathbf{A} = \mathbf{I}_n - \rho\mathbf{W}$. We perform the matrix multiplication $\mathbf{A}\mathbf{A}^\top$ repeatedly for a fixed n . The average computing time is calculated based on 5000 replications. This process is repeated for various n values, and the resulting computing time is plotted against n (see Figures S1.1 and S1.2 in Section S1 of the online

		N		
	N	O	N	
		N		

Figure 4.1: The grid used for constructing $\mathbf{W}$ based on the Rook neighbourhood (also referred to as the W-neighbourhood).

supplement). Finally, the following regression line

$$t_i = b \times n_i^\alpha, \quad (4.1)$$

is fitted to estimate the computational complexity, where t_i denotes the average computing time of $\mathbf{A}\mathbf{A}^\top$ when the total number of observations $n = n_i$, b is a constant, and $\alpha \geq 0$ is the computational complexity. However, the estimation of the regression is done in the log scale; $\log(t_i) = \log(b) + \alpha \times \log(n_i)$, using the least square method.

We briefly explain the grid and neighbourhood structure used in constructing the spatial weight matrix $\mathbf{W}$. Consider a regular grid of size $\sqrt{n} \times \sqrt{n}$, with n observations. We define neighbours based on the Rook neighbourhood criterion (Moura and Fonseca, 2020) as shown in Figure 4.1, where O is the unit of interest and, N's are neighbours of O. It is clear that any given unit can have 2 to 4 neighbours. The rows of matrix $\mathbf{W}$ (before row normalization) that correspond to Rook neighbourhoods can accommodate a maximum of 4 non-zero elements. The neighbourhood structure of $\mathbf{W}$ is often called the W-neighbourhood in spatial econometrics literature (Mukherjee et al., 2014; Suesse, 2018a). Moreover, given that each row of $\mathbf{W}$ contains a constant maximum number of non-zero elements (4) regardless of the value of n , the neighbourhood structure of $\mathbf{W}$ is defined as a local neighbourhood.

The computational complexities, as detailed in the final columns of Tables 4.1, 4.2, and 4.3, are obtained through a combination of theoretical and empirical analysis. The analysis is based on the spatial weight matrix $\mathbf{W}$ following the Rook neighbourhood structure. The overall computational complexity of a term is determined by the maximum complexity among the individual matrix operations involved. In Tables 4.1 and 4.2, the final row reflects the overall complexity of computing the computationally expensive terms in the marginal ML method using the direct approach and the parameterisation approach respectively. See

Table 4.1: Complexity of calculating $\mathbf{V}_{oo}$ using the direct method. The cells in the complexity column marked with * indicate the empirical complexity derived from simulations, whereas the cells without * represent the theoretical complexity. See Section S1.2 of the online supplement for a comprehensive derivation of these complexities.

Term	Description	Complexity
$\mathbf{A}^\top \mathbf{A}$	Two $n \times n$ sparse matrix multiplication	$O(n)$
$(\mathbf{A}^\top \mathbf{A})^{-1}$	Inverse of $n \times n$ sparse matrix	$O(n^2)$ *
$\mathbf{V}_{oo} = (\mathbf{I}_n + \theta(\mathbf{A}^\top \mathbf{A})^{-1})_{oo}$	No matrix operations	-
Overall	—	$O(n^2)$

Table 4.2: Complexity of calculating $\mathbf{V}_{oo}$ using the parameterisation method. The cells in the complexity column marked with * indicate the empirical complexity derived from simulations, whereas the cells without * represent the theoretical complexity. See Section S1.2 of the online supplement for a comprehensive derivation of these complexities.

Term	Description	Complexity
$\mathbf{A}_{\mathbf{B}_o}^{-1} = (\mathbf{A}^\top)^{-1} \mathbf{B}_o^\top = \text{solve}(\mathbf{A}^\top, \mathbf{B}_o^\top)$	solving a system of two sparse matrices	$O(n^{1.5}n_o)$ *
$\mathbf{A}\mathbf{A}^\top$	Two $n \times n$ sparse matrix multiplication	$O(n)$
$\mathbf{R} = (\mathbf{A}_{\mathbf{B}_o}^{-1})^\top (\mathbf{A}\mathbf{A}^\top + \theta \mathbf{I})$	$n_o \times n$ times $n \times n$	$O(nn_o)$
$\mathbf{V}_{oo} = \mathbf{R}\mathbf{A}_{\mathbf{B}_o}^{-1}$	$n_o \times n$ times $n \times n_o$	$O(nn_o^2)$
Overall	—	$\max(O(n^{1.5}n_o), O(nn_o^2))$

Section S1 of the online supplement for further details on the analysis of the computational complexity of the marginal ML method and EM algorithm.

4.1.1 Computational complexity of marginal ML method

This section discusses the complexity of the marginal ML method, described in Section 3.2. The computation of the sub-matrix $\mathbf{V}_{oo}$ plays an important role in determining the overall complexity of the method. Tables 4.1 and 4.2 present the complexities associated with the individual terms of the two alternative approaches for computing the sub-matrix $\mathbf{V}_{oo}$. In the direct approach, we compute the complete matrix $\mathbf{V}$ and then extract the sub-matrix $\mathbf{V}_{oo}$. On the other hand, in the parameterisation method, we compute $\mathbf{V}_{oo}$ by reformulating it using a sparse matrix $\mathbf{B}_o = [\mathbf{I}_o | 0]$, represented as $\mathbf{V}_{oo} = ((\mathbf{A}^\top)^{-1} \mathbf{B}_o^\top)^\top (\mathbf{A}\mathbf{A}^\top + \theta \mathbf{I}_n) ((\mathbf{A}^\top)^{-1} \mathbf{B}_o^\top)$.

Tables 4.1 and 4.2 show that the overall complexity of the direct method for calculating $\mathbf{V}_{oo}$ depends only on the size of the dataset (n), while the complexity of the parameterisation method depends on both the size of the dataset (n) and the number of observed data points

(n_o). We now discuss the overall complexity of the parameterisation method under two different real-world scenarios.

In the context of conducting a survey based on a sample from a larger population, the population is considered as the complete dataset with n units. The sample comprises observed data with n_o units; the remaining n_u units are considered missing. Choosing a criterion to determine the value of n_o results in varying complexities.

Scenario 1: ($n_o \leq c$)

When the size of n_o is constrained by a constant c ($n_o \leq c$), the overall complexity is $O(n^{1.5})$.

Scenario 2: $n_o = \sqrt{n}$

When $n_o = \sqrt{n}$, the overall complexity becomes $O(n^2)$.

In the first scenario, the decision is made to keep n_o fixed regardless of any increase in n . This leads to an overall computational complexity of $O(n^{3/2})$ for calculating $\mathbf{V}_{oo}$. In the second scenario, n_o remains relatively small compared to n . For example, with $n = 1,000,000$, the number of observed units n_o is 1,000. The number of observed units n_o is chosen to be relatively small compared to n , but still varying with n such that the computational complexity is $O(n^2)$. Thus, in real-world scenarios, the parameterisation approach typically demonstrates computational complexities either smaller or equivalent to the direct method, which carries a complexity of $O(n^2)$ for any given scenario.

We strongly recommend utilising the parameterisation method to compute $\mathbf{V}_{oo}$. This recommendation arises from two key considerations. First, it avoids the direct inversion of a sparse $n \times n$ matrix with a computational complexity of $O(n^2)$. Second, in the direct method, even though $\mathbf{A}^\top \mathbf{A}$ is a sparse matrix, its inverse is a dense matrix. As n increases, the size of $(\mathbf{A}^\top \mathbf{A})^{-1}$ grows, and storing $(\mathbf{A}^\top \mathbf{A})^{-1}$ becomes impractical. In simulation studies, we observed that attempting to store $(\mathbf{A}^\top \mathbf{A})^{-1}$ on the National Institute for Applied Statistics Research Australia High Performance Computer cluster² failed when n exceeded approximately 32,500. On the other hand, utilising the parameterisation method allowed us to obtain $\mathbf{V}_{oo}$, even for large values of n up to 1,000,000.

As $\mathbf{V}_{oo}$ is a dense matrix, subsequent computations after its calculation include determining its determinant and solving systems that involve $\mathbf{V}_{oo}$; see the log-likelihood of $\mathbf{z}_o$ in Equation (3.3). Due to the computational complexity of these operations, particularly when n_o is large, the practical applicability and efficiency of the marginal ML method are limited. Thus, we recommend to use the marginal ML method when n_o is small.

²<https://hpc.niasra.uow.edu.au/>

Table 4.3: Complexity of calculating terms in matrices $\mathbf{M}_{ou}$ and $\mathbf{M}_{uu}$. The cells in the complexity column marked with * indicate the empirical complexity derived from simulations, whereas cells without * represent the theoretical complexity. Refer to Section S1.3 of the online supplement for a comprehensive derivation of these complexities.

Operation	Description	Complexity
$\mathbf{A}_{B_u} = \mathbf{A}\mathbf{B}_u^\top$	Two sparse matrix multiplication	$O(nn_u)$
$\mathbf{A}\mathbf{A}^\top$	Two $n \times n$ sparse matrix multiplication	$O(n)$
$\mathbf{L}_{\mathbf{A}\mathbf{A}^\top}$	Cholesky of local neighbourhood	$O(n^{1.5})$
$\mathbf{L}_{\mathbf{A}\mathbf{A}^\top + \theta \mathbf{I}_n}$	Cholesky update	$O(n^{1.5})$ *
$\mathbf{S} = \text{solve}(\mathbf{L}_{\mathbf{A}\mathbf{A}^\top + \theta \mathbf{I}_n}, \mathbf{A}_{B_u})$	solving a system of two sparse matrices using Cholesky factors	$O(n^{1.5}n_u)$ *
$\mathbf{A}_{B_o}^\top = \mathbf{B}_o\mathbf{A}^\top$	$n_o \times n$ times $n \times n$	$O(nn_o)$
$\mathbf{M}_{uu} = \mathbf{A}_{B_u}^\top \mathbf{S}$	$n_u \times n$ times $n \times n_u$	$O(n_u^2)$
$\mathbf{M}_{ou} = \mathbf{A}_{B_o}^\top \mathbf{S}$	$n_o \times n$ times $n \times n_u$	$O(n_un_o)$
$\mathbf{T} = \text{solve}(\mathbf{M}_{uu}, \mathbf{M}_{uo}(\mathbf{z}_o - \boldsymbol{\mu}_o))$	solving a system with $n_u \times n_u$ dense matrix	$O(n_u^3)$

4.1.2 Computational complexity of EM algorithm

This section discusses the complexity of the EM algorithm, described in Section 3.3. The most computationally demanding aspects of the EM algorithm involve calculating the log-determinant, $\log|\mathbf{M}|$ and the conditional mean, $\boldsymbol{\mu}_{u|o}$. Notably, the computation of submatrices $\mathbf{M}_{uu}$ and $\mathbf{M}_{ou}$ includes all the necessary calculations for determining $\log|\mathbf{M}|$. For a comprehensive understanding of the overall computational complexity of the EM algorithm, exploring the details of computing $\mathbf{M}_{uu} = \mathbf{B}_u\mathbf{A}^\top (\mathbf{A}\mathbf{A}^\top + \mathbf{I}_n\theta)^{-1} \mathbf{A}\mathbf{B}_u^\top$, $\mathbf{M}_{ou} = \mathbf{B}_o\mathbf{A}^\top (\mathbf{A}\mathbf{A}^\top + \mathbf{I}_n\theta)^{-1} \mathbf{A}\mathbf{B}_u^\top$, and then computing $\boldsymbol{\mu}_{u|o} = \boldsymbol{\mu}_u - \mathbf{M}_{uu}^{-1}\mathbf{M}_{uo}(\mathbf{z}_o - \boldsymbol{\mu}_o)$ are sufficient. Table 4.3 provides a detailed summary of the essential matrix operations and their complexity for computing $\boldsymbol{\mu}_{u|o}$.

By reviewing Table (4.3), it is clear that the overall complexity of calculating $\mathbf{T}$ in Table (4.3) using $\mathbf{M}_{uu}$ and $\mathbf{M}_{ou}$ depends on the values of n , n_o , and n_u . The overall complexity can be written as $\max(O(n^{1.5}n_u), O(nn_o), O(n_on_u), O(n_u^3))$. We undertake an analysis of the overall computational complexity for calculating $\mathbf{T}$ across two distinct real-world scenarios.

Scenario 1: $n_u \leq c, n_o = n - c$

When n_u is limited by a constant c ($n_u \leq c$), n_o equals $n - c$. Consequently, the overall

complexity is determined by the maximum of either $O(n^{1.5})$ or $O(n^2 - cn)$. Since the highest order polynomial is 2, the overall complexity can be simplified to $O(n^2)$.

Scenario 2: $n_u = \sqrt{n}$, $n_o = n - \sqrt{n}$.

This leads to the overall complexity of $O(n^2)$.

Similar to the parameterisation approach in the marginal ML method, we delve into explaining the complexities of the two real-world scenarios for the EM algorithm. In the first scenario, the sampling design involves fixing n_u regardless of any increase in n . In this context, the calculation of $\mathbf{T}$ exhibits a complexity of $O(n^2)$. In the second scenario as well, where n_u is relatively small compared to n , the overall complexity remains $O(n^2)$.

It is noteworthy that, for any positive value of n , these overall complexities are found to be less than the complexity derived from the naive calculation of terms, which is $O(n^3)$. This implies that the implementation of the EM algorithm utilising our proposed computational approaches entails lesser computational complexity compared to the implementation involving the naive computation of terms, particularly under certain real-world scenarios.

To find the solution $\mathbf{T}$ for the equation $\mathbf{T} = \mathbf{M}_{uu}^{-1}\mathbf{M}_{uo}(\mathbf{z}_o - \boldsymbol{\mu}_o)$, where $\mathbf{M}_{uu}$ is an $n_u \times n_u$ symmetric and dense matrix, it is necessary to perform the Cholesky decomposition of $\mathbf{M}_{uu}$ (See Section S1.3.7 of the online supplement for the detailed implementation). However, computing the Cholesky decomposition becomes computationally challenging for large values of n_u due to the dense nature of $\mathbf{M}_{uu}$. Therefore, we recommend employing the EM algorithm when n_u is small.

In Figure 4.2, we present a plot of the average computing time (across 1000 replications) required to calculate the marginal likelihood of $\mathbf{z}_o$ in the marginal ML method (using the parameterisation approach) and to compute the term $Q(\boldsymbol{\phi} \mid \boldsymbol{\phi}')$ in the EM algorithm for H-SAM and H-SEM. This analysis is conducted with a fixed total number of observations (n) while varying the proportion of missing values (n_u). We set $n = 5000$, and consider 10%, 20%, ..., 90% of 5000 as the number of observed units n_o . The plot provides support for our complexity analysis and its findings, indicating that the marginal ML approach is better when the value of n_o is small, whereas the EM algorithm becomes more useful when n_o is large for both H-SEM and H-SAM.

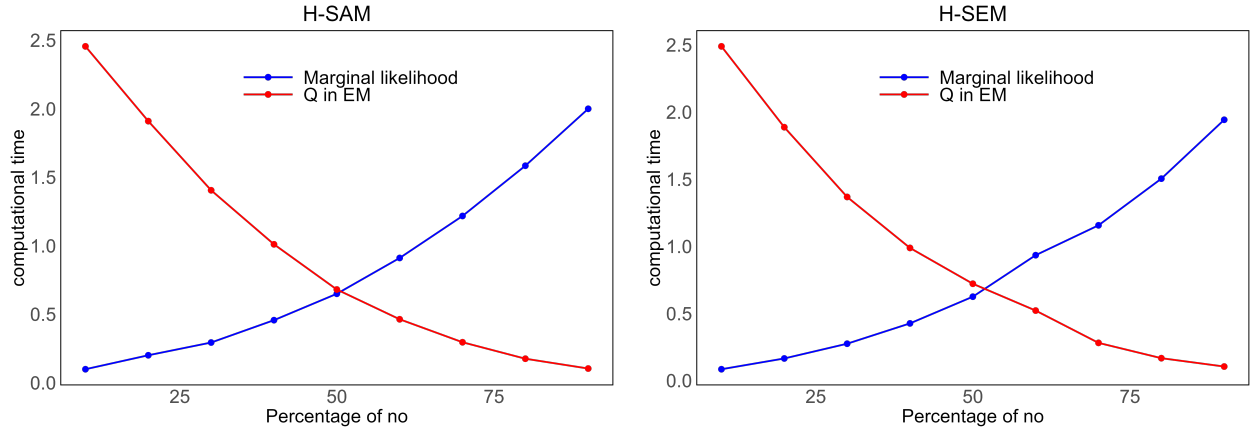


Figure 4.2: Average computing time of calculating the marginal log-likelihood of $\mathbf{z}_o$ for marginal ML, and $Q(\phi \mid \phi')$ for EM-algorithm for H-SAM (left) and H-SEM (right)

4.2 Comparing estimation methods

In the subsequent sections, we use the following terminology: the *full data model* (FDM) refers to the model fitted using the entire dataset. The *observed data model* (ODM), on the other hand, refers to the model fitted solely based on the observed locations. For this model, the weight matrix, design matrix, and the vector of the response variable are constructed exclusively from observed locations, denoted as $\mathbf{W}_{oo}$, $\mathbf{X}_o$, and $\mathbf{z}_o$ respectively. As discussed in Section 3.1, the ODM specification leads to inconsistencies and biases in parameter estimates. Furthermore, the *complete locations model* (CLM) refers to the model fitted based on the complete locations, which accurately captures the marginal distribution of $\mathbf{z}_o$. We employ two suggested estimation approaches from Sections 3.2 and 3.3 to fit the CLM.

Theoretically, the EM algorithm and the marginal ML method are expected to yield similar estimates. However, the frequent non-convergence of the EM algorithm often leads to divergent estimates. On the other hand, when the underlying optimisation problem presents multiple local maxima, employing optimisers, such as the Newton–Raphson method, for the marginal ML method to obtain the global maximum is challenging. Simulation studies are conducted with two main objectives: (1) to assess the accuracy and the computational efficiency of the marginal ML method and the EM algorithm, and (2) to evaluate the inconsistencies and the bias in the estimates when the ODM is used in the presence of missing data.

We now describe the approaches to assess convergence for marginal ML and EM algorithms. For the marginal ML method, we utilise the R `optim()` function to numerically maximise the marginal log-likelihood with respect to parameters θ and ρ . We set 10^{-8} as the convergence threshold for the `optim()` function. Regarding the EM algorithm, the standard

Table 4.4: Mean estimates of ρ , σ_ϵ^2 , and σ_y^2 for different percentages of missing data for the ODM and CLM estimated using marginal ML and EM algorithms for H-SEM. CLM-I: marginal ML method and CLM-II: EM algorithm. The true values are $\rho = 0.8$, $\sigma_\epsilon^2 = 2$, and $\sigma_y^2 = 1$. The empirical means are computed from 2500 simulated datasets, each with $n = 625$ units. Entries marked with 'NC' indicate that the EM algorithm failed to converge for a significant number of simulations, preventing the reliable calculation of mean estimates.

Model	10%			20%			30%		
	ρ	σ_ϵ^2	σ_y^2	ρ	σ_ϵ^2	σ_y^2	ρ	σ_ϵ^2	σ_y^2
ODM	0.7631	2.0128	1.0337	0.7400	2.1273	0.9719	0.7139	2.2285	0.9166
CLM-I	0.7779	1.8777	1.1342	0.7743	1.8771	1.1424	0.7817	1.8757	1.1203
CLM-II	0.7779	1.8777	1.1342	0.7743	1.8770	1.1424	0.7817	1.8757	1.1203

Model	70%			80%			90%		
	ρ	σ_ϵ^2	σ_y^2	ρ	σ_ϵ^2	σ_y^2	ρ	σ_ϵ^2	σ_y^2
ODM	0.5302	2.0334	1.6499	0.5012	2.1295	1.3452	0.2781	1.1193	2.7998
CLM-I	0.7726	1.7241	1.2912	0.8162	1.9895	0.8016	0.5547	0.9749	1.9075
CLM-II	0.7726	1.7248	1.2901	0.8159	1.9879	0.8021	NC	NC	NC

approach involves comparing the Euclidean distance between estimated parameter vectors at the i^{th} and $(i - 1)^{th}$ iterations of the EM algorithm (Mader et al., 2014). If the distance falls below 10^{-9} , we determine that the algorithm has converged. We also set a maximum limit of 1,000 iterations for the EM algorithm due to practical computational constraints. It is noteworthy that during each EM iteration, the estimates for θ and ρ are obtained using the `optim()` function by maximising $Q(\theta, \rho \mid \phi')$ in Equation (3.19). For this numerical optimisation task, we also set the convergence threshold of the `optim()` function to 10^{-8} .

To generate synthetic data, we set $\rho = 0.8$, $\sigma_\epsilon^2 = 2$, and $\sigma_y^2 = 1$ for the H-SEM and H-SAM. The regression parameters are fixed at $\beta_0 = 1$ and $\beta_1 = 5$. The covariate is drawn from a normal distribution with a mean of 0 and a standard deviation of 1. The spatial weight matrix is generated based on a 25×25 grid that follows the Rook neighbourhood. In total, we simulate 2500 data sets, each with $n = 625$ units. The same initial values are used for both marginal ML and EM algorithms.

Tables 4.4 and 4.5 present the mean estimates of ρ , σ_ϵ^2 , and σ_y^2 for the H-SEM and H-SAM with different percentages of missing data using the ODM and CLM specifications. Tables 4.6 and 4.7 further compare marginal ML and EM algorithms for estimating H-SEM and H-SAM, respectively. The tables include the average computational time, convergence rates of the EM algorithm, the average number of EM iterations, and the average relative distances (ARD) between the estimates obtained from the marginal ML method and EM

Table 4.5: Mean estimates of ρ , σ_ϵ^2 , and σ_y^2 for different percentages of missing data for the ODM and CLM estimated using marginal ML and EM algorithm for H-SAM. CLM-I: marginal ML method and CLM-II: EM algorithm. The true values are $\rho = 0.8$, $\sigma_\epsilon^2 = 2$, and $\sigma_y^2 = 1$. The empirical means are computed from 2500 simulated datasets, each with $n = 625$ units. Entries marked with 'NC' indicate that the EM algorithm failed to converge for a significant number of simulations, preventing the reliable calculation of mean estimates.

Model	10%			20%			30%		
	ρ	σ_ϵ^2	σ_y^2	ρ	σ_ϵ^2	σ_y^2	ρ	σ_ϵ^2	σ_y^2
ODM	0.7519	1.8826	2.1226	0.6892	1.8075	3.4897	0.6136	1.5454	5.3717
CLM-I	0.8009	2.0182	0.9621	0.7994	2.0213	0.9658	0.7998	2.0069	0.9734
CLM-II	0.8009	2.0182	0.9621	0.7994	2.0213	0.9659	0.7998	2.0069	0.9734

Model	70%			80%			90%		
	ρ	σ_ϵ^2	σ_y^2	ρ	σ_ϵ^2	σ_y^2	ρ	σ_ϵ^2	σ_y^2
ODM	0.3444	0.0387	15.1225	0.2737	0.3197	16.2810	0.2849	3.7648	13.8161
CLM-I	0.7987	2.0980	0.9440	0.8017	2.2445	0.8367	0.7991	2.2564	0.7381
CLM-II	0.7987	2.0970	0.9444	0.8017	2.2406	0.8384	NC	NC	NC

algorithm.

Tables 4.4 and 4.5 show that the mean estimates from marginal ML and EM algorithms are nearly identical for H-SEM and H-SAM, except for the case of 80% missing values. When analysing datasets with 90% missing values, the EM algorithm demonstrates slow convergence. Consequently, we solely compare the estimates from the ODM with those from the marginal ML method. For the H-SAM, regardless of the percentage of missing values, the CLM specifications consistently yield estimates closer to the true values with lower Mean Square Errors (MSEs) compared to estimates from the ODM (see Section S2 of the online supplement for MSEs). In the case of H-SEM, with missing value percentages of 10%, 20%, and 30%, although both ODM and CLM produce accurate parameter estimates, the estimates from CLM are slightly closer to the true values with lower MSEs. For higher missing value percentages, such as 70%, 80%, and 90%, the mean estimates for H-SEM obtained from the ODMs significantly deviate from the true values compared to CLMs. Section S2 of the online supplement provides further results for this study. These results include mean estimates, mean squared errors, and parameter coverages for all model parameters, including fixed effects (β), across various levels of missing data percentages.

As shown in Tables 4.6 and 4.7, when dealing with low percentages of missing values (10%, 20%, 30%), the EM algorithm consistently achieves convergence well before it reaches the maximum iteration limit across almost all simulated data sets for both H-SEM and H-

Table 4.6: Comparisons of average computing time, convergence rate, average number of EM iterations, and the ARD measure for fitting ODM and CLM (using marginal ML method and EM algorithm) for H-SEM. The properties are computed from 2500 simulated datasets, each with $n = 625$ units.

Property		Missing percentages					
		10	20	30	70	80	90
Average computing time (s)	ODM	1.1880	1.1477	1.1309	0.9018	0.8848	0.8207
	Marginal ML	33.2491	28.8777	23.5002	6.6306	4.1570	2.3825
	EM	40.942	83.797	200.009	3009.488	6488.425	6754.785
Convergence rate EM algorithm	-	1	0.9975	0.9920	0.7885	0.4423	0.2867
Average No. of EM iterations	-	46	70	110	565	812	895
ARD	-	9.8703	1.4330	2.3029	9.4642	1.3102	3.4279
		$\times 10^{-5}$	$\times 10^{-4}$	$\times 10^{-4}$	$\times 10^{-4}$	$\times 10^{-3}$	$\times 10^{-3}$

SAM. However, when working with datasets with high missing data percentages of 70%, 80%, 90%, we observe that the percentages of the convergence of the EM algorithm are 0.7885, 0.4423, 0.2867 for H-SEM and 0.9870, 0.2826 and 0.1831 for H-SAM on the simulated data sets, respectively. As the percentage of missing values increases, so does the average number of EM iterations. The observed behaviors can be attributed to the slower convergence rate exhibited by the EM algorithm when operating with a smaller number of observed units. We compare the estimates obtained from the EM algorithm with the estimates obtained using the marginal ML method and the true parameter values. The final rows in Tables 4.6 and 4.7 present a summary of the average relative distance (ARD) between EM and marginal ML estimates, relative to the true values, where $ARD = \sqrt{\sum_{i=1}^{r+3} \left(\frac{\hat{\theta}_{i(\text{marginal})} - \hat{\theta}_{i(\text{EM})}}{\theta_{i(\text{true})}} \right)^2}$, with $r + 3$ representing the length of the parameter vector. The measure quantifies the variation between the two estimation methods.

In both H-SEM and H-SAM, when the missing percentage exceeds 80%, the ARD between EM estimates and marginal ML estimates becomes larger than when the missing data percentages are lower than 80%, as presented in the final rows of Tables 4.6 and 4.7. This indicates a notable difference between estimates obtained using the marginal ML and EM algorithms in scenarios with higher missing value percentages compared to those with lower percentages. The difference is reflected in the observed variations in the mean estimates beyond the second decimal point, as illustrated in the respective cells of Tables 4.4 and 4.5. Tables 4.6 and 4.7 show that estimating ODM is faster than estimating CLM using the

Table 4.7: Comparisons of average computing time, convergence rate, average number of EM iterations, and the ARD measure for fitting ODM and CLM (using marginal ML method and EM algorithm) for H-SAM. The properties are computed from 2500 simulated datasets, each with $n = 625$ units.

Property		Missing percentages					
		10	20	30	70	80	90
Average time consumption (s)	ODM	1.1380	1.0012	0.9482	0.9613	0.8565	0.8413
	Marginal ML	33.2988	25.4981	20.1207	7.9276	6.9077	5.5549
	EM	59.542	74.295	165.862	2216.192	6368.822	7256.036
Convergence rate EM algorithm	-	1	0.9996	0.9996	0.9870	0.2826	0.1831
Average No. of EM iterations	-	42	55	84	560	918	927
ARD	-	1.5954	2.4737	3.5323	1.8938	2.6936	3.8173
		$\times 10^{-4}$	$\times 10^{-4}$	$\times 10^{-4}$	$\times 10^{-3}$	$\times 10^{-2}$	$\times 10^{-2}$

two proposed methods for any percentage of missing values. The marginal ML method is faster for estimating H-SEM and H-SAM when the missing percentages are high. The EM algorithm is faster when the missing value percentages are small.

Based on our complexity analysis detailed in Section 4.1, the EM algorithm demonstrates superior computational performance over the marginal ML method when the percentage of missing values (n_u) is low. However, Tables 4.6 and 4.7 show that the marginal ML is faster than the EM algorithm for all missing value percentages. There are two reasons. First, the EM algorithm's extremely small convergence threshold of 10^{-9} requires a large number of iterations for convergence, thereby contributing to its higher computing time relative to the marginal ML method. Second, the computational time of both algorithms is also influenced by the total number of observations n . We conducted an additional simulation study where we fitted ODMs and CLMs for H-SAM with $n = 2,500$, having low (10%) and high (90%) percentages of missing data. We adjusted the convergence threshold of the EM algorithm to 10^{-6} and set the maximum number of EM iterations to 1,000. Similar to the simulation study with $n = 625$ and 90% missing values, we observed slow convergence of the EM algorithm when $n = 2,500$ and 90% missing values. Consequently, in the scenario where 90% of values are missing, we only compare the ODM with the CLM estimated using the marginal ML method. The results of this study are presented in Table 4.8.

When estimating H-SAM with datasets with 10% of missing values percentages, the EM algorithm is faster than the marginal ML method. However, when dealing with datasets with 90% of missing values percentages, the marginal ML method outperforms the EM algorithm.

Table 4.8: Mean estimates of ρ , σ_ϵ^2 , and σ_y^2 and average computation times in seconds for 10% and 90% of missing data for the ODM and CLM estimated using marginal ML and EM algorithms for H-SAM. The true values are $\rho = 0.8$, $\sigma_\epsilon^2 = 2$, and $\sigma_y^2 = 1$. Entries marked with 'NC' indicate that the EM algorithm failed to converge for a significant number of simulations, preventing the reliable calculation of mean estimates and ARD. The properties are computed from 100 simulated datasets, each with $n = 2,500$ units.

Model	10%				90%			
	ρ	σ_ϵ^2	σ_y^2	Time (s)	ρ	σ_ϵ^2	σ_y^2	Time (s)
ODM	0.7584	1.9769	2.0744	20.5496	0.2415	0.0265	18.9752	0.8812
Marginal ML	0.8092	2.1325	0.8840	3778.8418	0.7999	2.0578	0.9458	136.5278
EM	0.8092	2.1326	0.8839	691.8043	NC	NC	NC	429878.3
Convergence rate EM algorithm	1				NC			
Average No. of EM iterations	42				NC			
ARD	1.3792 $\times 10^{-4}$				NC			

These findings confirm the results of our complexity analysis in Section 4.1. Furthermore, estimates derived from CLM, utilising both ML and EM algorithms are closer to the true values compared to estimates obtained from ODM for both 10% and 90% missing value percentages.

5 Real Data Application

This section applies the estimation methods to a real dataset. The dataset on house prices in Lucas County, Ohio, USA, contains a total of 25,357 observations of single-family homes sold between 1993 and 1998. The Spatial Econometrics toolbox for Matlab provides a comprehensive description of the dataset³. The dataset is also included in the R package spData (Bivand et al., 2023).

We use the natural logarithm of housing prices ($\ln(\text{price})$) as the dependent variable and include several independent variables, including various powers of house age (age, age², and age³), the logarithm of the lot size in square feet ($\ln(\text{lotsize})$), the number of rooms (rooms), the logarithm of the total living area in square feet (LTA), the number of bedrooms, and a binary indicator for each year from 1993 to 1998 (syear) to represent the year of the house

³The dataset can be accessed at <http://www.spatial-econometrics.com/html/jplv7.zip>

sale. The same row-normalized sparse weight matrix $\mathbf{W}$, as introduced by Bivand (2010), is used.

Suesse (2018a) conducted an analysis on the same dataset, using standard SEM and SAM (i.e. SAR models without measurement errors) and SEM and SAM with measurement errors. They employed the same dependent variable and covariates as those used by Bivand (2010). The analysis revealed that when ignoring measurement errors, SAM gives a better fit than SEM for the dataset, whereas when incorporating measurement errors, the SEM with measurement errors outperforms the SAM with measurement errors. Therefore, ignoring measurement errors in the SAR models can affect the conclusion significantly.

We first estimate the H-SEM and H-SAM with the complete dataset (FDM) and use the estimates as the ground truth for comparing ODM and CLMs. We then generate two datasets with missing values in the response variables from the complete dataset. The two datasets have missing percentages of 10% and 90%. Subsequently, we estimate the ODM and the CLM using the marginal ML and EM algorithms. We start both the marginal ML and EM algorithms with identical initial values. We set the convergence threshold for the `optim()` function to 10^{-8} for the marginal ML method. For the EM algorithm, we set the convergence threshold for the `optim()` function to 10^{-8} for numerical maximisation within each iteration, and 10^{-6} to terminate the EM algorithm.

Table 5.1 shows estimates and standard errors for ρ , σ_ϵ^2 , and σ_y^2 obtained using the FDM, the ODM, the CLM with marginal ML method, and the CLM with EM algorithm for the H-SEM with 10% of missing data. The table includes marginal log-likelihood values (for the marginal ML and the EM algorithms), along with fitting time, the time to compute the standard error of the parameters, and the number of EM iterations.

According to Table 5.1, the EM algorithm is faster than the marginal ML method for estimating H-SEM when the missing data percentage is small (10% of missing data). The results are consistent with our findings provided in Section 4.1. The marginal ML and EM methods yield identical parameter estimates. Moreover, the parameter estimates obtained from CLM are closer to the estimates obtained from the FDM than those of the ODM, which relies solely on observed data locations; see Table S3.3 in Section S3 of the online supplement for more detailed results, including the estimates and standard errors of fixed effects (β).

Table 5.2 shows estimates and standard errors for ρ , σ_ϵ^2 , and σ_y^2 obtained using the FDM, the ODM, the CLM with marginal ML method, and the CLM with EM algorithm for the H-SAM with 90% of missing data. The table shows that the marginal ML method outperforms the EM algorithm in terms of computing time when the percentage of missing values is high (i.e. n_o is small). This confirms our results outlined in Section 4.1. Due

Table 5.1: Parameter estimates (est.) with their standard errors (S.E.), marginal log-likelihood values for the marginal ML and the EM algorithms, estimation time and the time to calculate the standard errors of the parameters (Time std.error) in seconds for fitting H-SEM using FDM-ML for the full data set, and ODM-ML, CLM-Marginal ML, and CLM-EM for a partially observed data set with 10% of missing data, and the number of EM iterations

Variables	FDM-ML		ODM-ML		CLM -Marginal ML		CLM-EM	
	est.	S.E.	est.	S.E.	est.	S.E.	est.	S.E.
ρ	0.9866	0.0002	0.8654	0.0049	0.9868	0.0002	0.9868	0.0002
σ_y^2	0.0004	0.0001	0.0226	0.0011	0.0004	0.0001	0.0004	0.0001
σ_ϵ^2	0.0685	0.0007	0.0603	0.0012	0.0691	0.0007	0.0691	0.0007
number of EM iterations	—		—		—		8	
estimation time (s)	14.249		4.963		1803.9157		1546.3783	
Time std.error (s)	5.849		5.52		59.2792		65.7387	
marginal log-likelihood	—		—		-5885.136		-5885.098	

to computational constraints, we terminated the EM algorithm after 100 iterations (after around 28 hours), and at this point, the EM algorithm had not yet converged. The lack of convergence highlights the typical slow convergence of the EM algorithm when handling high percentages of missing data, supporting our recommendation of employing the marginal ML method for large missing data percentages in Section 4.1; see Table S3.2 in Section S3 of the online supplement for further details. In this example, the computational cost per iteration of the marginal ML method and EM algorithm is expensive because the total number of observations (n) is large. Moreover, the online supplement provides a summary of the H-SAM with 10% missing data and the H-SEM with 90% missing data; see Tables S3.1 and S3.4 in Section S3 of the online supplement for further details.

To compare the accuracy of the parameter estimates obtained from the ODM specification and those observed from the CLM, we computed the MSE of the estimated parameters for both H-SEM and H-SAM as shown in Table 5.3. The estimates from the FDM are treated as true parameter values, and the MSE of the parameter estimates obtained using the EM algorithm is calculated using the formula $MSE_{EM} = \frac{1}{r+3} \sum_{i=1}^{r+3} (\hat{\theta}_i - \hat{\theta}_{i,EM})^2$, where $\hat{\theta}_i$ is the estimated value for the i -th parameter from the FDM, and $\hat{\theta}_{i,EM}$ is the estimated value for the i -th parameter obtained from EM algorithm, and $r+3$ is the total number of parameters. A similar measure is also defined for the marginal ML method.

As expected, parameter estimates obtained from the data with lower missing data per-

Table 5.2: Parameter estimates (est.) with their standard errors (S.E.), marginal log-likelihood values for the marginal ML and the EM algorithms, estimation time and the time to calculate the standard errors of the parameters (Time std.error) in seconds for fitting H-SAM using FDM-ML for the full data set, and ODM-ML, CLM-Marginal ML, and CLM-EM for a partially observed data set with 90% of missing data, and the number of EM iterations. Entries labeled as 'NC' indicate instances where the EM algorithm did not successfully converge.

Variables	FDM-ML		ODM-ML		CLM -Marginal ML		CLM-EM	
	est.	S.E.	est.	S.E.	est.	S.E.	est.	S.E.
ρ	0.6727	0.0001	0.0027	0.0311	0.6046	0.0201	NC	
σ_y^2	0.0399	0.0008	0.0882	0.0114	0.0682	0.0070	NC	
σ_ϵ^2	0.042	0.0009	0.0882	0.0234	0.0203	0.0080	NC	
number of EM iterations	—		—		—		100	
estimation time (s)	6.319		0.216		18.7909		101916.2943	
Time std.error (s)	5.926		0.836		46.3919		41.6773	
marginal log-likelihood	—		—		-1120.645		NC	

Table 5.3: Mean squared errors of estimates of ODM and CLMs (marginal ML and EM estimates) relative to the FDM estimates for different missing value percentages. Cells marked with an asterisk (*) indicate instances where the EM algorithm has not converged.

		H-SEM		H-SAM	
		10% missing	90% missing	10% missing	90% missing
ODM	ML	0.1663	0.4838	1.0392	1.1248
CLMs	Marginal- ML	0.0046	0.2556	4×10^{-5}	0.0172
	EM algorithm	0.0045	*	4×10^{-5}	*

centages (10%) exhibit lower MSE compared to those from scenarios with higher missing value percentages (90%), irrespective of whether the ODM or the CLM model specifications are used for H-SEM and H-SAM. The MSE estimates obtained from the ODM are consistently much larger than those of CLMs for the 10% and 90% missing data percentages for H-SEM and H-SAM.

6 Conclusion

The article proposes two approaches for estimating the parameters of SAR models with measurement errors in the presence of missing data. The first method involves directly maximising the marginal log-likelihood. The second method is based on the EM algorithm. The two methods are particularly effective when the spatial weight matrix $\mathbf{W}$ is sparse.

The computational complexity analysis and our examples suggest that when dealing with datasets with a small number of observed data (i.e., n_o is small), the marginal ML is more efficient than the EM algorithm. There are two main reasons. First, the marginal ML method becomes more advantageous in terms of computational efficiency for small n_o because calculating the inverse and determinant of matrix $\mathbf{V}_{oo}$ presents fewer computational challenges in scenarios where n_o is small. Second, for datasets with small n_o , the convergence of the EM algorithm tends to be slow, leading to a high number of iterations required to achieve convergence.

When dealing with datasets containing a small number of missing data (i.e., n_u is small), we recommend employing the EM algorithm due to its lower computational complexity compared to the marginal ML method. The EM algorithm's computational demands mainly depend on the terms $\mathbf{M}_{uu}$ and $\mathbf{M}_{ou}$. When n_u is small, the computational burden associated with calculating these terms is considerably small. In addition, the EM algorithm tends to converge rapidly for datasets with a small number of missing data n_u , resulting in a lower number of iterations required for convergence. On the contrary, the marginal likelihood method involves operations such as $\mathbf{V}_{oo}^{-1}$ and calculation of the determinant of $\mathbf{V}_{oo}$. These operations are computationally intensive and may become infeasible for datasets with very large values of n_o (small values of n_u) because $\mathbf{V}_{oo}$ is a dense matrix.

A notable limitation of our proposed methods arises when both n_o and n_u are large, leading to computational challenges due to dense matrix operations. For instance, consider a scenario where $n = 100,000$, and $n_o = 1,000$. In such a case, the marginal ML method is both preferable and computationally feasible, since n_o is small. However, as the value of n_o increases to 20,000, even though the marginal ML method remains preferable due to the small size of n_o compared to n , it becomes computationally challenging. Specifically,

computing the determinant of a $20,000 \times 20,000$ dense matrix ($|\mathbf{V}_{oo}|$; see the marginal log-likelihood in Equation (3.3)) becomes difficult. Conversely, when $n_o = 20,000$, n_u reaches 80,000 employing the EM algorithm also becomes problematic due to its computational demands.

In future research, it is important to address computational challenges arising from large n_o and n_u . This can be accomplished by employing methods to approximate the marginal covariance of $\mathbf{z}_o$, such as the reduced rank approach utilised by Burden et al. (2015), as demonstrated for the SEM with measurement errors, in the cases where the data is fully observed. Furthermore, while our study primarily investigates the estimation of H-SAR models under the missing at random (MAR) mechanism, there is a pressing need to extend investigations to encompass estimations with the missing not-at-random (MNAR) mechanism.

References

- Allison, P. D. (2001). *Missing data*. Sage publications.
- Ammermueller, A. and Pischke, J.-S. (2009). Peer effects in European primary schools: Evidence from the progress in international reading literacy study. *Journal of Labor Economics*, 27(3):315–348.
- Angrist, J. D. and Lang, K. (2004). Does school integration generate peer effects? evidence from Boston’s Metco program. *American Economic Review*, 94(5):1613–1634.
- Anselin, L. (1988). *Spatial econometrics: methods and models*, volume 4. Springer Science & Business Media.
- Arab, A., Hooten, M. B., and Wikle, C. K. (2008). Hierarchical spatial models. *Encyclopedia of GIS*, 14(1):425–431.
- Bates, D., Maechler, M., and Jagan, M. (2022). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.5-3.
- Benedetti, R., Suesse, T., and Piersimoni, F. (2020). Spatial auto-correlation and auto-regressive models estimation from sample survey data. *Biometrical Journal*, 62(6):1494–1507.
- Bishop, C. M. (1998). Latent variable models. In *Learning in graphical models*, pages 371–403. Springer.
- Bivand, R. (2010). Comparing estimation methods for spatial econometrics techniques using R. *NHH Dept. of Economics Discussion Paper*, (26).
- Bivand, R., Gómez-Rubio, V., and Rue, H. (2015). Spatial data analysis with R-INLA with some extensions. *Journal of statistical software*, 63:1–31.
- Bivand, R., Nowosad, J., and Lovelace, R. (2023). *spData: Datasets for Spatial Analysis*. R package version 2.2.2.
- Burden, S., Cressie, N., and Steel, D. G. (2015). The SAR model for very large datasets: a reduced rank approach. *Econometrics*, 3(2):317–338.

- Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(1):209–226.
- Davis, T. A. (2006). *Direct Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.
- Dong, G. and Harris, R. (2015). Spatial autoregressive models for geographically hierarchical data structures. *Geographical Analysis*, 47(2):173–191.
- Enders, C. K. (2003). Using the Expectation Maximization algorithm to estimate coefficient alpha for scales with item-level missing data. *Psychological Methods*, 8(3):322.
- Ford, W. (2015). Chapter 11 - Gaussian elimination and the LU decomposition. In Ford, W., editor, *Numerical Linear Algebra with Applications*, pages 205–239. Academic Press, Boston.
- Glaeser, E. L., Sacerdote, B., and Scheinkman, J. A. (1996). Crime and social interactions. *The Quarterly Journal of Economics*, 111(2):507–548.
- Gómez-Rubio, V., Bivand, R. S., and Rue, H. (2021). Estimating spatial econometrics models with Integrated Nested Laplace Approximation. *Mathematics*, 9(17):2044.
- Hepple, L. W. (1979). Bayesian analysis of the linear model with spatial dependence. In *Exploratory and Explanatory Statistical Analysis of Spatial Data*, pages 179–199. Springer.
- Kelejian, H. H. and Prucha, I. R. (1999). A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review*, 40(2):509–533.
- Kelejian, H. H. and Prucha, I. R. (2001). On the asymptotic distribution of the Moran I test statistic with applications. *Journal of Econometrics*, 104(2):219–257.
- Kelejian, H. H. and Prucha, I. R. (2010). Spatial models with spatially lagged dependent variables and incomplete data. *Journal of Geographical Systems*, 12(3):241–257.
- Kelley Pace, R. and Barry, R. P. (1997). Fast cars. *Journal of Statistical Computation and Simulation*, 59(2):123–145.
- Lauritzen, S. L. (1995). The EM algorithm for graphical association models with missing data. *Computational Statistics & Data Analysis*, 19(2):191–201.
- Lee, L.-f. (2003). Best spatial two-stage least squares estimators for a spatial autoregressive model with autoregressive disturbances. *Econometric Reviews*, 22(4):307–335.
- Lee, L.-f. (2007). GMM and 2SLS estimation of mixed regressive, spatial autoregressive models. *Journal of Econometrics*, 137(2):489–514.
- LeSage, J. and Pace, R. K. (2009). *Introduction to spatial econometrics*. Chapman and Hall/CRC.
- LeSage, J. P. (1997). Bayesian estimation of spatial autoregressive models. *International Regional Science Review*, 20(1-2):113–129.
- LeSage, J. P. and Pace, R. K. (2004). Models for spatially dependent missing data. *The Journal of Real Estate Finance and Economics*, 29(2):233–254.
- Li, H., Calder, C. A., and Cressie, N. (2012). One-step estimation of spatial dependence parameters: Properties and extensions of the APLE statistic. *Journal of Multivariate Analysis*, 105(1):68–84.

- Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.
- Longstaff, F. A. (2010). The subprime credit crisis and contagion in financial markets. *Journal of Financial Economics*, 97(3):436–450.
- Luo, G., Wu, M., and Xu, L. (2021). IPW-based robust estimation of the SAR model with missing data. *Statistics & Probability Letters*, 172:109065.
- Mader, W., Linke, Y., Mader, M., Sommerlade, L., Timmer, J., and Schelter, B. (2014). A numerically efficient implementation of the expectation maximization algorithm for state space models. *Applied Mathematics and Computation*, 241:222–232.
- Moura, A. C. M. and Fonseca, B. M. (2020). ESDA (exploratory spatial data analysis) of vegetation cover in urban areas—recognition of vulnerabilities for the management of resources in urban green infrastructure. *Sustainability*, 12(5):1933.
- Mukherjee, C., Kasibhatla, P. S., and West, M. (2014). Spatially varying SAR models and Bayesian inference for high-resolution lattice data. *Annals of the Institute of Statistical Mathematics*, 66(3):473–494.
- Ord, K. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, 70(349):120–126.
- Pace, R. K. (1997). Performing large spatial regressions and autoregressions. *Economics Letters*, 54(3):283–291.
- Pace, R. K. and Barry, R. (1997). Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297.
- Pace, R. K. and LeSage, J. P. (2004). Chebyshev approximation of log-determinants of spatial weight matrices. *Computational Statistics & Data Analysis*, 45(2):179–196.
- Petersen, K. B., Pedersen, M. S., et al. (2008). The Matrix Cookbook. *Technical University of Denmark*, 7(15):510.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.
- Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall/CRC.
- Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using Integrated Nested Laplace Approximations. *Journal of the Royal Statistical Society: Series B (statistical methodology)*, 71(2):319–392.
- Su, L. (2012). Semiparametric GMM estimation of spatial autoregressive models. *Journal of Econometrics*, 167(2):543–560.
- Suesse, T. (2018a). Estimation of spatial autoregressive models with measurement error for large data sets. *Computational Statistics*, 33(4):1627–1648.
- Suesse, T. (2018b). Marginal maximum likelihood estimation of SAR models with missing data. *Computational Statistics & Data Analysis*, 120:98–110.
- Suesse, T. and Zammit-Mangion, A. (2017). Computational aspects of the EM algorithm for spatial econometric models with missing data. *Journal of Statistical Computation and Simulation*, 87(9):1767–1786.
- Tognelli, M. F. and Kelt, D. A. (2004). Analysis of determinants of mammalian species richness in South America using spatial autoregressive models. *Ecography*, 27(4):427–436.

- Ver Hoef, J. M., Peterson, E. E., Hooten, M. B., Hanks, E. M., and Fortin, M.-J. (2018). Spatial autoregressive models for statistical inference from ecological data. *Ecological Monographs*, 88(1):36–59.
- Wang, W. and Lee, L.-F. (2013). Estimation of spatial autoregressive models with randomly missing data in the dependent variable. *The Econometrics Journal*, 16(1):73–102.

Online Supplement for Statistical Inference on Hierarchical Simultaneous Autoregressive Models with Missing Data

We use the following notation in the supplement. Eq. (1), Table 1, and Figure 1, etc, refer to the main paper, while Eq. (S1.1), Table S1.1, and Figure S1.1, etc, refer to the supplement.

S1 Computational complexity analysis

This section examines the computational complexities associated with computing computationally demanding individual terms for the two proposed likelihood approaches discussed in Sections 3.2 and 3.3 of the main paper. We explore these complexities through theoretical derivations and empirical estimations obtained from simulations.

S1.1 Common terms computed in both marginal ML and EM methods: calculating $\mathbf{A}\mathbf{A}^\top$

This section illustrates the complexity of the term $\mathbf{A}\mathbf{A}^\top$ defined in the rows of different tables in the main paper: (1) the first row of Table 4.1 in Section 4.1.1, (2) the second row of Table 4.2 in Section 4.1.1, and (3) the second row of Table 4.3 in Section 4.1.1.

We observed that $\mathbf{W}$ contains a local neighbourhood; see Section 4.1 of the main paper. As $\mathbf{A} = \mathbf{I}_n - \rho\mathbf{W}$, the structure of $\mathbf{A}$ also encompasses a local neighbourhood. Each row of $\mathbf{A}$ contains a maximum of 3, 4, or 5 non-zero elements, which is one additional non-zero element compared to $\mathbf{W}$. The additional non-zero element arises from the fact that while the diagonal of $\mathbf{W}$ is all zeros, the diagonal elements of $\mathbf{A}$ are 1.

To calculate a non-zero element in each row of $\mathbf{A}\mathbf{A}^\top$, it necessitates a maximum of 5 scalar multiplications and 4 scalar additions, resulting in a total of 9 flops. Simulation studies have shown that for any value of n , $\mathbf{A}\mathbf{A}^\top$ contains at most 13 non-zero entries in each row, indicating that $\mathbf{A}\mathbf{A}^\top$ constitutes a local neighbourhood. This implies that a total of 9×13 maximum number of flops are required to compute one row in $\mathbf{A}\mathbf{A}^\top$. With a total of n rows, the computation of all elements in $\mathbf{A}\mathbf{A}^\top$ demands $n \times 9 \times 13$ flops. As a result, the overall maximum computational complexity of this operation is $O(n)$. Through simulation studies, the time complexity of computing $\mathbf{A}\mathbf{A}^\top$ has been estimated to be approximately $O(n^{1.032}) \approx O(n)$. The method to estimate the computational complexity is discussed in Section 4.1 of the main document. However, in the implementation, this operation is not

conducted as a matrix operation. It can be simplified as a scalar multiplication, expressed as $\mathbf{A}\mathbf{A}^\top = \mathbf{I} - \rho(\mathbf{W} + \mathbf{W}^\top) + \rho^2\mathbf{W}\mathbf{W}^\top$. To maximise the concentrated log-likelihood in Equation (3.7) and the function $Q(\phi \mid \phi')$ in Equation (3.19) from the main document with respect to ρ and θ , we provide precomputed matrices $\mathbf{W} + \mathbf{W}^\top$ and $\mathbf{W}\mathbf{W}^\top$, along with the concentrated log-likelihood and $Q(\phi \mid \phi')$, as inputs to the `optim()` function for numerical maximisation; see Section 3 of the main paper.

S1.2 Computing terms in the marginal ML method

S1.2.1 Calculate $\mathbf{A}_{\mathbf{B}_o}^{-1}$

This section discusses the complexity of the term $\mathbf{A}_{\mathbf{B}_o}^{-1}$ found in the first row of Table 4.2 in Section 4.1.1 of the main paper.

In R programming language, the `solve()` function serves a dual purpose, allowing us to calculate the inverse of a matrix and solve systems of linear equations. This function takes two primary arguments. The command `solve(M1, M2)` solves the system $\mathbf{M}_1\mathbf{X} = \mathbf{M}_2$ for $\mathbf{X}$, with both $\mathbf{M}_2$ and $\mathbf{X}$ can be either a vector or a matrix, provided that they have compatible sizes. If direct inversion of $\mathbf{M}_1$ is desired, the command `solve(M1)` or `solve(M1, I)` can be employed, where $\mathbf{I}$ represents the identity matrix matched in size to $\mathbf{M}_1$.

It is important to highlight that when dealing with a sparse matrix $\mathbf{M}_1$, the `solve()` function in R's Matrix package utilises specialised algorithms designed for sparse matrices. These algorithms, including sparse Cholesky decomposition (see Section S1.3.1) and sparse LU decomposition (Davis, 2006), are explicitly used to optimise computations in scenarios involving sparse matrices. For more information, interested readers can refer to the documentation of the Matrix package (Bates et al., 2022).

In our case, the objective is to compute $\mathbf{A}_{\mathbf{B}_o}^{-1}$ with the provided matrices $\mathbf{A}^\top$ and $\mathbf{B}_o^\top$, where the calculation is expressed as

$$\mathbf{A}_{\mathbf{B}_o}^{-1} = (\mathbf{A}^\top)^{-1}\mathbf{B}_o^\top. \quad (\text{S1.1})$$

When the command `solve(A⊤, Bo⊤)` is executed, the solve function performs an LU decomposition on the sparse square matrix $\mathbf{A}^\top$ from the Csparse library (Davis, 2006), which is optimised for sparse matrices. The sparsity of $\mathbf{A}^\top$ allows the use of efficient sparse matrix LU decomposition algorithms, typically achieving computational complexities below $O(n^3)$. Let $\mathbf{L}_{\mathbf{A}^\top}$ and $\mathbf{U}_{\mathbf{A}^\top}$ represent the lower and upper triangular matrices obtained from the LU decomposition of $\mathbf{A}^\top$. The system to be solved in Equation (S1.1) can be expressed as:

$$\begin{aligned}
\mathbf{A}^\top \mathbf{A}_{\mathbf{B}_o}^{-1} &= \mathbf{B}_o^\top, \\
(\mathbf{L}_{\mathbf{A}^\top} \mathbf{U}_{\mathbf{A}^\top}) \mathbf{A}_{\mathbf{B}_o}^{-1} &= \mathbf{B}_o^\top, \\
\mathbf{L}_{\mathbf{A}^\top} (\mathbf{U}_{\mathbf{A}^\top} \mathbf{A}_{\mathbf{B}_o}^{-1}) &= \mathbf{B}_o^\top.
\end{aligned} \tag{S1.2}$$

To solve the system described in Equation (S1.2), a two-step process is employed. Firstly, forward substitution is used on the system $\mathbf{L}_{\mathbf{A}^\top} \mathbf{Z} = \mathbf{B}_o^\top$ to derive the solution for $\mathbf{Z}$. Subsequently, with $\mathbf{Z}$ determined, backward substitution is applied to the system $\mathbf{U}_{\mathbf{A}^\top} \mathbf{A}_{\mathbf{B}_o}^{-1} = \mathbf{Z}$ to obtain $\mathbf{A}_{\mathbf{B}_o}^{-1}$.

During the application of forward substitution to $\mathbf{L}_{\mathbf{A}^\top} \mathbf{Z} = \mathbf{B}_o^\top$, assuming all the lower triangular elements of $\mathbf{L}_{\mathbf{A}^\top}$ are non-zero, the computation of each row of the matrix $\mathbf{Z}$ requires n^2 flops per row (Ford, 2015). Given that $\mathbf{Z}$ has n_o columns, the total number of flops needed to compute it is $n^2 n_o$. Similarly, assuming all the upper triangular elements of $\mathbf{U}_{\mathbf{A}^\top}$ are non-zero, the application of backward substitution to $\mathbf{U}_{\mathbf{A}^\top} \mathbf{A}_{\mathbf{B}_o}^{-1} = \mathbf{Z}$ requires $n^2 n_o$ flops. Hence, the total number of flops required for both forward and backward substitutions is $2n^2 n_o$, and the complexity of these operations should not exceed $O(n^2 n_o)$.

In summary, the `solve()` function first computes the LU decomposition of $\mathbf{A}^\top$, with a complexity of less than $O(n^3)$. Subsequently, the application of forward and backward substitution involves a complexity of at most $O(n^2 n_o)$. As a result, the overall complexity of the `solve( $\mathbf{A}^\top, \mathbf{B}_o^\top$ )` operation should be less than $O(n^3)$. Through simulations, considering a constant value for n_o , the estimated time complexity for the `solve( $\mathbf{A}^\top, \mathbf{B}_o^\top$ )` operation, which includes both the LU decomposition of $\mathbf{A}^\top$ and the forward and backward substitutions steps is approximately $O(n^{1.522}) \approx O(n^{3/2})$. Moreover, a simulation study suggests that, when considering a constant value for n , the complexity of the `solve( $\mathbf{A}^\top, \mathbf{B}_o^\top$ )` operation can be estimated as $O(n_o^{0.973}) \approx O(n_o)$. This finding aligns with the overall complexity of the `solve( $\mathbf{A}^\top, \mathbf{B}_o^\top$ )` operation derived from its theoretical implementation.

S1.2.2 Calculate $\mathbf{R}$

This section aims to demonstrate the complexity of the term $\mathbf{R}$ given in the third row of Table 4.2 in Section 4.1.1 of the main paper. The operation of calculating $\mathbf{R}$ given $(\mathbf{A}_{\mathbf{B}_o}^{-1})^\top$ and $(\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n)$ is a matrix multiplication:

$$\mathbf{R} = (\mathbf{A}_{\mathbf{B}_o}^{-1})^\top (\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n). \tag{S1.3}$$

It is clear that the structure of the matrix $\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n$ is similar to that of $\mathbf{A}\mathbf{A}^\top$. i.e. entries of zero and non-zero in $\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n$ equivalent to that of $\mathbf{A}\mathbf{A}^\top$. This implies that $\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n$

also comprises a local neighbourhood having at most 13 non-zero entries in each row.

We understand that $(\mathbf{A}_{\mathbf{B}_o}^{-1})^\top$ represents an $n_o \times n$ dense matrix, and $\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n$ is an $n \times n$ sparse matrix, with each row having a maximum of c non-zero elements, where $c = 13$. Since $\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n$ is symmetric, this property extends to its columns as well; each column also contains a maximum of c non-zero elements. The resulting matrix, $\mathbf{R}$, is an $n_o \times n$ dense matrix.

To compute a single entry in the $\mathbf{R}$ matrix, a maximum of $2c - 1$ flops are required, involving c scalar multiplications and $c - 1$ scalar additions. Subsequently, calculating an entire row in $\mathbf{R}$ necessitates evaluating n columns, leading to a total of $n(2c - 1)$ flops. Given that $\mathbf{R}$ has n_o rows, the total number of flops required to compute all rows in $\mathbf{R}$ is given by $n_on(2c - 1)$. As a result, the maximum time complexity for the operation described in Equation (S1.3) is $O(nn_o)$. Specifically, when n_o is held constant, the complexity reduces to $O(n)$, whereas, when n is fixed, the maximum complexity becomes $O(n_o)$.

Through simulation studies, we find that when n_o remains constant, the complexity of the operation described in Equation (S1.3) is $O(n^{1.298}) \approx O(n)$. Conversely, with a fixed n , the complexity is estimated to be $O(n_o^{1.215}) \approx O(n_o)$. These findings confirm the theoretical complexity analysis.

S1.2.3 Calculating $\mathbf{V}_{oo}$

This section aims to demonstrate the complexity of the term $\mathbf{V}_{oo}$ found in the fourth row of Table 4.2 in Section 4.1.1 of the main paper. Given matrices $\mathbf{R}$ and $\mathbf{A}_{\mathbf{B}_o}^{-1}$, the calculation of $\mathbf{V}_{oo}$ involves performing a matrix multiplication on two dense matrices. This operation can be expressed as:

$$\mathbf{V}_{oo} = \mathbf{R}\mathbf{A}_{\mathbf{B}_o}^{-1}. \quad (\text{S1.4})$$

The matrix dimensions for $\mathbf{R}$ and $\mathbf{A}_{\mathbf{B}_o}^{-1}$ are $n_o \times n$ and $n \times n_o$, respectively. Calculating a single entry in the $\mathbf{V}_{oo}$ matrix requires exactly $2n - 1$ flops. With $\mathbf{V}_{oo}$ containing n_o columns, computing an entire row demands exactly $n_o(2n - 1)$ flops. Since $\mathbf{V}_{oo}$ has n_o rows, the total flops for computing the entire $\mathbf{V}_{oo}$ matrix amount to exactly $n_o^2(2n - 1)$. Thus the complexity of the operation in Equation (S1.4) can be derived as $O(nn_o^2)$.

S1.2.4 Inversion of $\mathbf{A}^\top \mathbf{A}$

This section aims to demonstrate the complexity of computing the inverse of $\mathbf{A}^\top \mathbf{A}$ found in the second row of Table 4.1 in Section 4.1.1 of the main paper.

As described in Section S1.2.1, the `solve()` function is utilised to compute the inversion of $\mathbf{A}^\top \mathbf{A}$. In order to calculate inverse of $\mathbf{A}^\top \mathbf{A}$, we use the command `solve( $\mathbf{A}^\top \mathbf{A}$ ,  $\mathbf{I}$ )`. The system $(\mathbf{A}^\top \mathbf{A}) \mathbf{X} = \mathbf{I}$ requires solving for the matrix $\mathbf{X}$. Since $\mathbf{A}^\top \mathbf{A}$ is symmetric and sparse, The R function first calculates the Cholesky factors of $\mathbf{A}^\top \mathbf{A}$. Due to its local neighbourhood structure, the complexity of the Cholesky factorisation is $O(n^{3/2})$. Let us denote the Cholesky factor of $\mathbf{A}^\top \mathbf{A}$ as $\mathbf{L}_{\mathbf{A}^\top \mathbf{A}}$. The system to be solved can be written as:

$$\begin{aligned} (\mathbf{A}^\top \mathbf{A}) \mathbf{X} &= \mathbf{I}, \\ \left(\mathbf{L}_{\mathbf{A}^\top \mathbf{A}} (\mathbf{L}_{\mathbf{A}^\top \mathbf{A}})^\top \right) \mathbf{X} &= \mathbf{I}, \\ \mathbf{L}_{\mathbf{A}^\top \mathbf{A}} \left((\mathbf{L}_{\mathbf{A}^\top \mathbf{A}})^\top \mathbf{X} \right) &= \mathbf{I}. \end{aligned} \tag{S1.5}$$

To solve the system presented in Equation (S1.5), a two-step process is employed. Firstly, forward substitution is applied to the system $\mathbf{L}_{\mathbf{A}^\top \mathbf{A}} \mathbf{Z} = \mathbf{I}$, resulting in the derivation of $\mathbf{Z}$. Subsequently, with $\mathbf{Z}$ in hand, backward substitution is conducted on the system $(\mathbf{L}_{\mathbf{A}^\top \mathbf{A}})^\top \mathbf{X} = \mathbf{Z}$ to yield $\mathbf{X}$, which is the inverse of $\mathbf{A}^\top \mathbf{A}$.

During the application of forward substitution to $\mathbf{L}_{\mathbf{A}^\top \mathbf{A}} \mathbf{Z} = \mathbf{I}$ for obtaining $\mathbf{Z}$, it is notable that $\mathbf{L}_{\mathbf{A}^\top \mathbf{A}}$ forms a lower triangular matrix. Assuming all the lower triangular elements are non-zero, the computation of one row of the matrix $\mathbf{Z}$ requires n^2 flops. As $\mathbf{Z}$ possesses n columns, the total number of flops needed to compute its all entries is n^3 . Similarly, the application of backward substitution to $(\mathbf{L}_{\mathbf{A}^\top \mathbf{A}})^\top \mathbf{X} = \mathbf{Z}$ for obtaining $\mathbf{X}$ assuming all the upper triangular elements in $(\mathbf{L}_{\mathbf{A}^\top \mathbf{A}})^\top$ are non-zero demands n^3 flops.

However, it is crucial to emphasize that $\mathbf{A}^\top \mathbf{A}$ is a sparse matrix, meaning its Cholesky factors are also sparse. Consequently, a substantial portion of entries in the lower triangular matrix $\mathbf{L}_{\mathbf{A}^\top \mathbf{A}}$ is zero. This leads to a reduction in the number of flops required for both forward and backward substitutions, making it less than n^3 . This indicates that, the complexity of these operations is lower than $O(n^3)$.

In summary, the `solve()` function initially computes the Cholesky factors of $\mathbf{A}^\top \mathbf{A}$, which carries a complexity of $O(n^{3/2})$. Subsequently, the application of forward and backward substitution, benefiting from the sparsity of $\mathbf{A}^\top \mathbf{A}$, involves a complexity of less than $O(n^3)$. Consequently, the overall complexity of the `solve()` operation falls within the range of $O(n^{3/2})$ and $O(n^3)$. Based on a simulation study, the estimated computational complexity for the `solve( $\mathbf{A}^\top \mathbf{A}$ ,  $\mathbf{I}$ )` operation is $O(n^{2.057}) \approx O(n^2)$.

S1.3 Computing terms in the EM algorithm approach.

The computational complexity of the EM algorithm depends on computing submatrices $\mathbf{M}_{uu}$ and $\mathbf{M}_{ou}$, where $\mathbf{M}_{uu} = \mathbf{B}_u \mathbf{A}^\top (\mathbf{A} \mathbf{A}^\top + \mathbf{I}_n \theta)^{-1} \mathbf{A} \mathbf{B}_u^\top$ and $\mathbf{M}_{ou} = \mathbf{B}_o \mathbf{A}^\top (\mathbf{A} \mathbf{A}^\top + \mathbf{I}_n \theta)^{-1} \mathbf{A} \mathbf{B}_u^\top$. To investigate the total computational complexity of calculating these terms, we need to consider the individual complexity of each matrix operation involved in their computations.

S1.3.1 Calculating Cholesky factorisation of $\mathbf{A} \mathbf{A}^\top$

This section discusses the complexity of the term found in the third row of Table 4.3 in Section 4.1.2 of the main paper.

As discussed in Section S1.1, the matrix $\mathbf{A} \mathbf{A}^\top$ represents a local neighborhood. Rue and Held (2005) demonstrated that when performing Cholesky factorisation on a sparse matrix with a local neighborhood constructed on an $\sqrt{n} \times \sqrt{n}$ regular grid, the computational complexity is $O(n^{3/2})$. Consequently, the computational complexity of Cholesky factorisation for $\mathbf{A} \mathbf{A}^\top$ is $O(n^{3/2})$. Our simulation studies confirmed this by demonstrating an estimated computational complexity of $O(n^{1.659}) \approx O(n^{3/2})$.

S1.3.2 Calculating $\mathbf{A}_{\mathbf{B}_u}$

This section aims to demonstrate the complexity of the term $\mathbf{A}_{\mathbf{B}_u}$ found in the first row of Table 4.3 in Section 4.1.2 of the main paper. Given matrices $\mathbf{A}$ and $\mathbf{B}_u^\top$, the calculation of $\mathbf{A}_{\mathbf{B}_u}$ involves performing a matrix multiplication between two sparse matrices, as defined by the equation:

$$\mathbf{A}_{\mathbf{B}_u} = \mathbf{A} \mathbf{B}_u^\top \quad (\text{S1.6})$$

We previously established that the $n \times n$ matrix $\mathbf{A}$ represents a local neighbourhood, and $\mathbf{B}_u = [\mathbf{0} \mid \mathbf{I}_u]$ is an $n_u \times n$ sparse matrix, where $\mathbf{I}_u$ is the $n_u \times n_u$ identity matrix, and $\mathbf{0}$ is the $n_u \times n_o$ zero matrix. In $\mathbf{B}_u^\top$, each column has exactly one element with a value of 1, while all other elements are zero. This implies that to compute a single element in $\mathbf{A}_{\mathbf{B}_u}$, a maximum of 1 flop is required. Considering that each row in $\mathbf{A}_{\mathbf{B}_u}$ has n_u entries (corresponding to the n_u columns in $\mathbf{A}_{\mathbf{B}_u}$), the computation of an entire row demands at most n_u flops. Given that there are n rows in $\mathbf{A}_{\mathbf{B}_u}$, the total maximum flops required to compute all rows are $n_u \times n$. This results in a maximum total computational complexity of $O(nn_u)$. In a simulation study, it has been estimated that the computational complexity of the operation in Equation (S1.6) is $O(n^{0.7831}) \approx O(n)$ for a fixed n_u .

S1.3.3 Calculating $\mathbf{A}_{\mathbf{B}_o}^\top = \mathbf{B}_o \mathbf{A}^\top$

This section aims to demonstrate the complexity of the term $\mathbf{A}_{\mathbf{B}_o}^\top$ found in the sixth row of Table 4.3 in Section 4.1.2 of the main paper. Given matrices $\mathbf{A}^\top$ and $\mathbf{B}_o$, the calculation of $\mathbf{A}_{\mathbf{B}_o}^\top$ involves performing a matrix multiplication between two sparse matrices, as defined by the equation:

$$\mathbf{A}_{\mathbf{B}_o}^\top = \mathbf{B}_o \mathbf{A}^\top. \quad (\text{S1.7})$$

The matrix $\mathbf{A}^\top$ represents a local neighbourhood with the dimension $n \times n$, and $\mathbf{B}_o = [\mathbf{I}_o | \mathbf{0}]$ is a sparse matrix of size $n_o \times n$, where $\mathbf{I}_o$ is the $n_o \times n_o$ identity matrix, and $\mathbf{0}$ is the $n_o \times n_u$ zero matrix. Consequently, $\mathbf{A}_{\mathbf{B}_o}^\top$ becomes an $n_o \times n$ sparse matrix. In $\mathbf{B}_o$, each row has exactly one element with a value of 1, while all other elements are zero. This implies that to compute a single element in $\mathbf{A}_{\mathbf{B}_o}^\top$, a maximum of 1 flop is required. Considering that each row in $\mathbf{A}_{\mathbf{B}_o}^\top$ has n entries (corresponding to the n columns in $\mathbf{A}^\top$), the computation of an entire row demands at most n flops. Finally, since there are n_o rows in $\mathbf{A}_{\mathbf{B}_o}^\top$, the total maximum flops required to compute all rows is $n_o \times n$. This results in a maximum total computational complexity of $O(nn_o)$. In a simulation study, it has been estimated that the computational complexity of the operation in Equation (S1.7) is $O(n^{0.9477}) \approx O(n)$ for a fixed n_o .

S1.3.4 Calculating $\mathbf{S} = (\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n)^{-1}\mathbf{A}_{\mathbf{B}_u}$

This section aims to demonstrate the complexity of the term $\mathbf{S}$ found in the fifth row of Table 4.3 in Section 4.1.2 of the main paper. Given matrices $\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n$ and $\mathbf{A}_{\mathbf{B}_u}$, the calculation to obtain $\mathbf{S}$ is

$$\mathbf{S} = (\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n)^{-1}\mathbf{A}_{\mathbf{B}_u}, \quad (\text{S1.8})$$

and this operation can be expressed as solving the following system for $\mathbf{S}$:

$$(\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n)\mathbf{S} = \mathbf{A}_{\mathbf{B}_u}. \quad (\text{S1.9})$$

The `solve()` function can be used to solve the system presented in Equation (S1.9) by employing the command `solve( $\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n$ ,  $\mathbf{A}_{\mathbf{B}_u}$ )`. However, the `solve()` function can operate more efficiently with the Cholesky factor of $\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n$, denoted as $\mathbf{L}_{\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n}$, and solving the system with the command `solve( $\mathbf{L}_{\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n}$ ,  $\mathbf{A}_{\mathbf{B}_u}$ )`. To do this, it is necessary to compute $\mathbf{L}_{\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n}$. We describe the calculation of $\mathbf{L}_{\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n}$ using the Cholesky factor of $\mathbf{A}\mathbf{A}^\top$ and R's `update()` function in Section 3.3 of the main paper. Consequently, $\mathbf{L}_{\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n}$ is readily

available and can be used to solve the system presented in Equation (S1.9) more efficiently.

Calculation of $\mathbf{S}$ (including calculation of $\mathbf{L}_{\mathbf{A}\mathbf{A}^\top + \theta \mathbf{I}_n}$) involves three steps. Firstly, we compute the Cholesky factor of $\mathbf{A}\mathbf{A}^\top$, which has a complexity of $O(n^{3/2})$ (refer to Section S1.3.1 for details). Subsequently, the Cholesky factorisation of $\mathbf{A}\mathbf{A}^\top + \theta \mathbf{I}_n$ is carried out by calling the `update()` function. The simulation studies carried out by Suesse (2018a) suggest that the computational complexity of this operation is approximately $O(n^{1.5})$. Finally, we execute the command `solve(LAA⊤+θIn, ABu)` to solve the original system presented in Equation (S1.9). The estimated complexities associated with this operation are: when n_u is held constant, the complexity is $O(n^{1.497}) \approx O(n^{3/2})$, and when n is held constant, the complexity becomes $O(n_u^{0.968}) \approx O(n_u)$.

S1.3.5 Calculating $\mathbf{M}_{uu} = (\mathbf{A}_{\mathbf{B}_u})^\top \mathbf{S}$

This section aims to demonstrate the complexity of the term $\mathbf{M}_{uu}$ found in the seventh row of Table 4.3 in Section 4.1.2 of the main paper. Given matrices $\mathbf{A}_{\mathbf{B}_u}$ and $\mathbf{S}$, the calculation of $\mathbf{M}_{uu}$ involves performing a matrix multiplication. This operation is carried out by multiplying a sparse matrix with a dense matrix, as specified by the following equation:

$$\mathbf{M}_{uu} = (\mathbf{A}_{\mathbf{B}_u})^\top \mathbf{S}. \quad (\text{S1.10})$$

The dimensions of the matrices are $n_u \times n$ for $(\mathbf{A}_{\mathbf{B}_u})^\top$ and $n \times n_u$ for $\mathbf{S}$. Notably, $\mathbf{S}$ is a dense matrix due to the involvement of a matrix inverse in its calculation (see Equation (S1.8)).

In contrast, while $(\mathbf{A}_{\mathbf{B}_u})^\top$ is a sparse matrix, its specific structure proves challenging to generalise. Consequently, simulation studies were undertaken to explore its structure. The results show that, for a fixed value of n and varying values of n_u , $(\mathbf{A}_{\mathbf{B}_u})^\top$ exhibits a local neighbourhood structure, with a maximum of 5 nonzero elements in each row. Similarly, with a fixed value of n_u and varying values of n , a corresponding local neighbourhood structure was observed in $(\mathbf{A}_{\mathbf{B}_u})^\top$, again with a maximum of 5 nonzero elements in each row.

Based on these findings, we demonstrate that computing a single entry of $\mathbf{M}_{uu}$ requires a maximum of 9 flops (5 scalar multiplications and 4 additions). As there are n_u entries in each row, the total number of maximum operations needed for computing a row is $9n_u$. Given that $\mathbf{M}_{uu}$ comprises n_u rows, the total maximum number of operations required to calculate all entries in $\mathbf{M}_{uu}$ is $9n_u^2$. This results in a maximum computational complexity of $O(n_u^2)$. The results from simulation studies indicate that the empirical complexity of this operation is estimated to be $O(n_u^{2.1683}) \approx O(n_u^2)$ when n is held constant.

S1.3.6 Calculating $\mathbf{M}_{ou} = \mathbf{A}_{\mathbf{B}_o}^\top \mathbf{S}$

This section aims to demonstrate the complexity of the term $\mathbf{M}_{ou}$ found in the eighth row of Table 4.3 in Section 4.1.2 of the main paper. Given the matrices $\mathbf{A}_{\mathbf{B}_o}^\top$ and $\mathbf{S}$, the calculation of $\mathbf{M}_{ou}$ involves performing a matrix multiplication. This operation is carried out by multiplying a sparse matrix with a dense matrix, as specified by the following equation:

$$\mathbf{M}_{ou} = \mathbf{A}_{\mathbf{B}_o}^\top \mathbf{S}. \quad (\text{S1.11})$$

Similar to the calculation of $\mathbf{M}_{uu}$ using $\mathbf{A}_{\mathbf{B}_o}^\top$ and $\mathbf{S}$ discussed in section S1.3.5, due to the difficulty in generalising the structure of $\mathbf{A}_{\mathbf{B}_o}^\top$, simulation studies were conducted to investigate its structure. The results revealed that $\mathbf{A}_{\mathbf{B}_o}^\top$ contains a local neighbourhood with a maximum of 5 nonzero elements. It can be shown that the maximum number of flops required to calculate all entries in $\mathbf{M}_{ou}$ is $9n_on_u$. The proof is very similar to the proof for calculating $\mathbf{M}_{uu}$, discussed in Section S1.3.5. This implies that the maximum computational complexity of this operation is $O(n_on_u)$.

S1.3.7 Calculating $\mathbf{T} = \mathbf{M}_{uu}^{-1} \mathbf{M}_{uo}(\mathbf{z}_o - \boldsymbol{\mu}_o)$

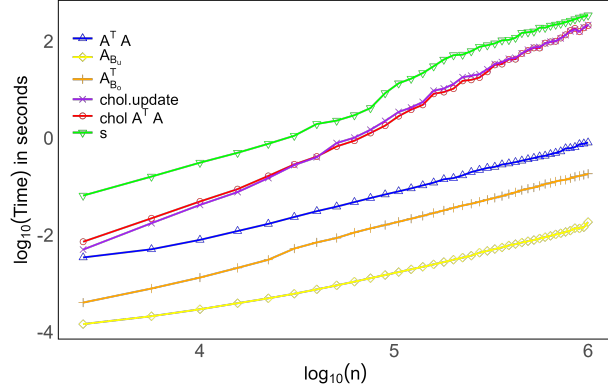
This section aims to demonstrate the complexity of the term $\mathbf{T}$ found in the last row of Table 4.3 in Section 4.1.2 of the main paper. Given matrices $\mathbf{M}_{uu}$ and $\mathbf{M}_{uo}(\mathbf{z}_o - \boldsymbol{\mu}_o)$, the calculation to obtain $\mathbf{T}$ is expressed as follows

$$\mathbf{T} = \mathbf{M}_{uu}^{-1} \mathbf{m}_{\mathbf{z}_u}, \quad (\text{S1.12})$$

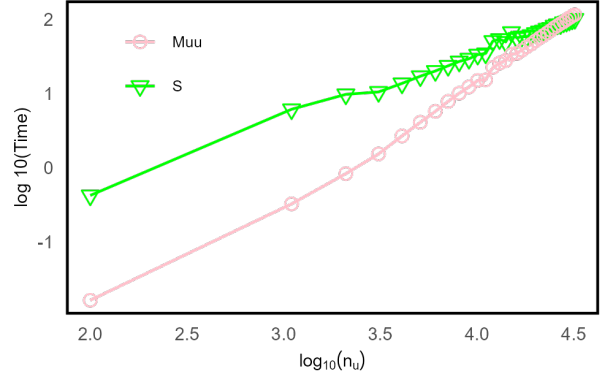
where $\mathbf{m}_{\mathbf{z}_u} = \mathbf{M}_{uo}(\mathbf{z}_o - \boldsymbol{\mu}_o)$, and this operation can be expressed as solving the following system for $\mathbf{T}$:

$$\mathbf{M}_{uu} \mathbf{T} = \mathbf{m}_{\mathbf{z}_u} \quad (\text{S1.13})$$

As discussed in Section S1.2.1, the `solve` function can be employed to solve the system presented in Equation (S1.13) using the command `solve( $\mathbf{M}_{uu}$ ,  $\mathbf{m}_{\mathbf{z}_u}$ )`. It is important to note that while $\mathbf{M}_{uu}$ is an $n_u \times n_u$ dense matrix, $\mathbf{m}_{\mathbf{z}_u}$ is a vector of length n_u . Consequently, the `solve()` function cannot leverage sparse matrix operations, resulting in computational complexity for solving a system with a dense matrix, which is $O(n_u^3)$.

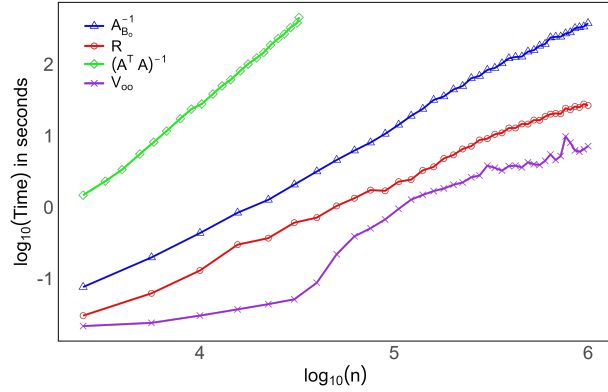


(a) Fixed n_u and n_o

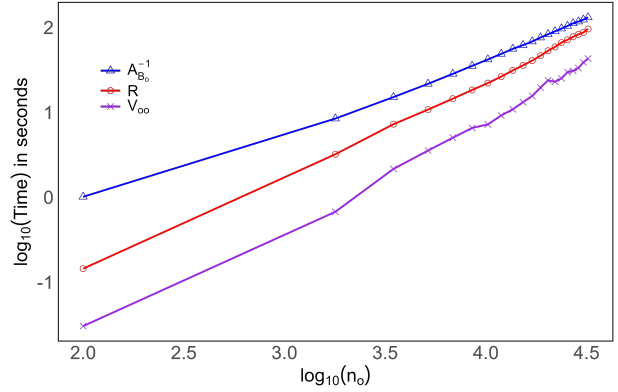


(b) Fixed n

Figure S1.1: Time needed of various matrix operations in the EM algorithm in $\log_{10}(\text{time})$ in seconds versus $\log_{10}(n)$ in the left panel and $\log_{10}(n_u)$ in the right panel



(a) Fixed n_u and n_o



(b) Fixed n

Figure S1.2: Time needed of various matrix operations in the marginal ML method in $\log_{10}(\text{time})$ in seconds versus $\log_{10}(n)$ in the left panel and $\log_{10}(n_o)$ in the right panel.

Figure S1.1(a) illustrates the relationship between the log of the average computing time needed to compute individual terms in the EM algorithm and the log of the total number of units (n). Figure S1.1(b) shows the relationship between the log of the average computing time and the log of the number of unobserved units (n_u). The estimated complexities (empirical complexities), discussed in Section S1.3, are derived by estimating the slopes of these graphs. See Section 4.1 in the main paper for further details. Similarly, Figure S1.2(a) displays the logarithm of the average computing time of individual terms in the marginal ML method against the logarithm of n . Figure S1.2(b) illustrates the relationship between the logarithm of average computing time and the logarithm of the number of observed units,

n_o . The estimated complexities, discussed in Section S1.2, are determined by estimating the slopes of these graphs.

S1.4 Complexity summary

Table S1.1: Theoretical and empirical (estimated) computational complexities of calculating $\mathbf{V}_{oo}$ using the Direct method.

Term	Description	Theoretical complexity	Empirical complexity
$\mathbf{A}^\top \mathbf{A}$	two $n \times n$ sparse matrix multiplication	$O(n)$	$O(n^{1.032})$
$(\mathbf{A}^\top \mathbf{A})^{-1}$	Direct inversion of $n \times n$ sparse matrix	-	$O(n^{2.057})$
$\mathbf{V}_{oo} = (\mathbf{I}_n + \theta(\mathbf{A}^\top \mathbf{A})^{-1})_{oo}$	No matrix operations	-	-

Table S1.1 presents a summary of the computational complexities associated with each operation necessary for calculating $\mathbf{V}_{oo}$ in the marginal ML method using the direct method. Notably, the computational complexity primarily relies on the inversion of $\mathbf{A}^\top \mathbf{A}$. However, it is important to note that evaluating this complexity theoretically is challenging. As a result, we approximate its complexity using simulations as $O(n^{2.057})$, which can be expressed as approximately $O(n^2)$.

Table S1.2: Theoretical and empirical (estimated) computational complexities of calculating $\mathbf{V}_{oo}$ using the parameterisation method.

Term	Description	Theoretical complexity	Empirical complexity	
			fixed n_o	fixed n
$\mathbf{A}_{\mathbf{B}_o}^{-1} = \text{solve}((\mathbf{A}^\top)^{-1}, \mathbf{B}_o^\top)$	solving a system of two sparse matrices	-	$O(n^{1.522})$	$O(n_o^{0.973})$
$\mathbf{A}\mathbf{A}^\top$	Two $n \times n$ sparse matrix multiplication	$O(n)$	$O(n^{1.032})$	constant
$\mathbf{R} = (\mathbf{A}_{\mathbf{B}_o}^{-1})^\top (\mathbf{A}\mathbf{A}^\top + \theta \mathbf{I})$	$n_o \times n$ times $n \times n$	$O(nn_o)$	$O(n^{1.298})$	$O(n_o^{1.215})$
$\mathbf{V}_{oo} = \mathbf{R}\mathbf{A}_{\mathbf{B}_o}^{-1}$	$n_o \times n$ times $n \times n_o$	$O(nn_o^2)$	$O(n^{1.045})$	$O(n_o^{1.795})$

Table S1.2 provides a summary of the computational complexity associated with each operation required to calculate $\mathbf{V}_{oo}$ in the marginal ML method using the parameterisation method. With the exception of the first operation, we have access to theoretical complexity values for all other operations, and these values align with the estimated complexities. For the first operation, the computational complexity estimate, based on simulations, is

$O(n^{1.522}n_o^{0.973})$, which can be further simplified to approximately $O(n^{1.5}n_o)$. In Section 4.1.1 of the main paper, we provide a comprehensive discussion on the overall complexity of calculating $\mathbf{V}_{oo}$, both in a general context and in the context of two specific real-world scenarios.

Table S1.3: Theoretical and empirical (estimated) computational complexities of calculating $\mathbf{T}$ in the EM algorithm.

Term	Description	Theoretical complexity	Empirical complexity	
			fixed n_u	fixed n
$\mathbf{A}_{\mathbf{B}_u} = \mathbf{A}\mathbf{B}_u^\top$	Two sparse matrix multiplication	$O(nn_u)$	$O(n^{0.7831})$	
$\mathbf{A}\mathbf{A}^\top$	Two $n \times n$ sparse matrix multiplication	$O(n)$	$O(n^{1.032})$	constant
$\mathbf{L}_{\mathbf{A}\mathbf{A}^\top}$	Cholesky of local neighbourhood	$O(n^{1.5})$		constant
$\mathbf{L}_{\mathbf{A}\mathbf{A}^\top + \theta \mathbf{I}_n}$	Cholesky update	-		constant
$\mathbf{S} = \text{solve}(\mathbf{L}_{\mathbf{A}\mathbf{A}^\top + \theta \mathbf{I}_n}, \mathbf{A}_{\mathbf{B}_u})$	solving a system of two sparse matrices using Cholesky factors	-	$O(n^{1.497})$	$O(n_u^{0.9098})$
$\mathbf{A}_{\mathbf{B}_o}^\top = \mathbf{B}_o \mathbf{A}^\top$	$n_o \times n$ times $n \times n$	$O(nn_o)$	$O(n^{0.9477})$	
$\mathbf{M}_{uu} = \mathbf{A}_{\mathbf{B}_u}^\top \mathbf{S}$	$n_u \times n$ times $n \times n_u$	$O(n_u^2)$	constant	$O(n_u^{2.1683})$
$\mathbf{M}_{ou} = \mathbf{A}_{\mathbf{B}_o}^\top \mathbf{S}$	$n_o \times n$ times $n \times n_u$	$O(n_u n_o)$	constant	
$\mathbf{T} = \text{solve}(\mathbf{M}_{uu}, \mathbf{M}_{uo}(\mathbf{z}_o - \boldsymbol{\mu}_o))$ $= \mathbf{M}_{uu}^{-1} \mathbf{M}_{uo}(\mathbf{z}_o - \boldsymbol{\mu}_o)$	solving a system with a dense matrix and a vector	$O(n_u^3)$	constant	

Similarly, Table S1.3 summarises the computational complexities attached to individual operations in calculating the conditional mean $\boldsymbol{\mu}_{u|o}$ for the proposed EM algorithm. Except for the fourth and fifth operations (calculating the Cholesky update and $\mathbf{S}$), we have theoretical complexity values. We can approximate the complexity of $\mathbf{S}$ as $O(n^{1.497}n_u^{0.9098})$, which can be expressed as approximately $O(n^{1.5}n_o)$. Moreover, when performing the Cholesky update operation, we rely on the complexity findings presented by Suesse (2018a). These findings suggest an approximate computational complexity of $O(n^{1.5})$. A comprehensive overview of the overall computational complexity involved in calculating $\boldsymbol{\mu}_{u|o}$ is available in Section 4.1.2 of the main paper.

S2 Extended simulation results

Tables S2.1 to S2.12 present the empirical means, mean squared errors of parameter estimates and their coverages obtained from 2500 simulated datasets (each with $n = 625$ units) for the

H-SEM and H-SAM. The synthetic data for these simulations are generated under settings similar to those employed in the simulation study outlined in Section 4.2 of the main paper. The coverage calculation is determined using 95% Wald-type confidence intervals, employing the formula $\hat{\delta} \pm 1.96\sqrt{\widehat{\text{Var}}(\hat{\delta})}$, where $\delta \in \phi$ and $\widehat{\text{Var}}(\hat{\delta})$ is the estimated variance of the parameter δ . The estimated variances for the parameters are computed based on the formula given in Section S6 of the online supplement.

Table S2.1: H-SEM mean estimates (est.), mean squared error (m.s.e.), and coverage (cov.) of estimated parameters by the ODM, the marginal ML method, and the EM algorithm - 10% of missing data. These attributes are computed from 2500 simulated datasets, each with $n = 625$ units.

Parameter	True value	ODM			CLM					
		ML			Marginal ML			EM algorithm		
		est.	m.s.e	cov.	est.	m.s.e	cov.	est.	m.s.e	cov.
β_0	1	0.9929	0.0415	0.9329	0.9932	0.0411	0.9399	0.9932	0.0411	0.9399
β_1	5	4.9962	0.0050	0.9717	4.9961	0.0050	0.9717	4.9961	0.0050	0.9717
ρ	0.8	0.7631	0.0080	0.9399	0.7779	0.0060	0.9470	0.7779	0.0060	0.9505
σ_ϵ^2	2	2.0128	0.1807	0.9364	1.8777	0.1995	0.9682	1.8777	0.1994	0.9717
σ_y^2	1	1.0337	0.2190	0.9011	1.1342	0.2360	0.9399	1.1342	0.2359	0.9435

For H-SEM, when estimating parameters with a small percentage of missing data, specifically 10%, 20%, and 30% (see Tables S2.1, S2.3, and S2.5), we observe that, the mean estimates obtained from all three approaches (the ODM specification, marginal ML method, and EM algorithm) are relatively close to the true values. However, the mean squared errors (MSE) of estimates from the ODM specification tend to be generally higher compared to those from the CLM.

Table S2.2: H-SAM mean estimates (est.), mean squared error (m.s.e.), and coverage (cov.) of estimated parameters by the ODM, the marginal ML method, and the EM algorithm - 10% of missing data. These attributes are computed from 2500 simulated datasets, each with $n = 625$ units.

Parameter	True value	ODM			CLM					
		ML			Marginal ML			EM algorithm		
		est.	m.s.e	cov.	est.	m.s.e	cov.	est.	m.s.e	cov.
β_0	1	1.2403	0.0746	0.3307	0.9945	0.0049	0.9402	0.9945	0.0049	0.9402
β_1	5	5.0602	0.0138	0.8924	4.9970	0.0070	0.9323	4.9970	0.0070	0.9323
ρ	0.8	0.7519	0.0026	0.0737	0.8009	0.0001	0.9422	0.8009	0.0001	0.9422
σ_ϵ^2	2	1.8826	0.1091	0.9402	2.0182	0.0686	0.9502	2.0182	0.0686	0.9502
σ_y^2	1	2.1226	1.3899	0.0319	0.9621	0.0385	0.9183	0.9621	0.0384	0.9183

For H-SAM, when dealing with 10%, 20%, 30%, 70%, and 80% missing data, the marginal ML method and EM algorithm consistently provide nearly identical estimates with a coverage probability of approximately 0.95 (i.e., 95%) and exhibit low mean squared errors (MSE). In contrast, the estimates obtained from the observed data model specification deviate significantly from the true values, displaying higher MSE compared to the marginal ML method and EM algorithm. Some parameters even exhibit 0 coverage in certain cases. See Tables S2.2, S2.4, S2.6, S2.8, and S2.10 for further details.

Table S2.3: H-SEM mean estimates (est.), mean squared error (m.s.e.), and coverage (cov.) of estimated parameters by the ODM, the marginal ML method, and the EM algorithm - 20% of missing data. These attributes are computed from 2500 simulated datasets, each with $n = 625$ units.

Parameter	True value	ODM			CLM					
		ML			Marginal ML			EM algorithm		
		est.	m.s.e	cov.	est.	m.s.e	cov.	est.	m.s.e	cov.
β_0	1	1.0422	0.0494	0.8774	1.0429	0.0489	0.9151	1.0429	0.0489	0.9151
β_1	5	5.0062	0.0066	0.9623	5.0060	0.0066	0.9623	5.0060	0.0066	0.9623
ρ	0.8	0.7400	0.0138	0.9214	0.7743	0.0077	0.9308	0.7743	0.0077	0.9245
σ_ϵ^2	2	2.1273	0.2788	0.8679	1.8771	0.2538	0.9434	1.8770	0.2537	0.9403
σ_y^2	1	0.9719	0.2932	0.8270	1.1424	0.2631	0.9308	1.1424	0.2630	0.9245

Table S2.4: H-SAM mean estimates (est.), mean squared error (m.s.e.), and coverage (cov.) of estimated parameters by the ODM, the marginal ML method, and the EM algorithm - 20% of missing data. These attributes are computed from 2500 simulated datasets, each with $n = 625$ units.

Parameter	True value	ODM			CLM					
		ML			Marginal ML			EM algorithm		
		est.	m.s.e	cov.	est.	m.s.e	cov.	est.	m.s.e	cov.
β_0	1	1.5631	0.3570	0.0351	1.0041	0.0045	0.9396	1.0041	0.0045	0.9396
β_1	5	5.1861	0.0489	0.6316	5.0077	0.0071	0.9688	5.0077	0.0071	0.9688
ρ	0.8	0.6892	0.0129	0.0000	0.7994	0.0001	0.9376	0.7994	0.0001	0.9376
σ_ϵ^2	2	1.8075	0.2354	0.9240	2.0213	0.0833	0.9532	2.0213	0.0832	0.9532
σ_y^2	1	3.4897	6.5525	0.0000	0.9658	0.0454	0.9181	0.9659	0.0454	0.9181

Table S2.5: H-SEM mean estimates (est.), mean squared error (m.s.e.), and coverage (cov.) of estimated parameters by the ODM, the marginal ML method, and the EM algorithm - 30% of missing data. These attributes are computed from 2500 simulated datasets, each with $n = 625$ units.

Parameter	True value	ODM			CLM					
		ML			Marginal ML			EM algorithm		
		est.	m.s.e	cov.	est.	m.s.e	cov.	est.	m.s.e	cov.
β_0	1	0.9690	0.0457	0.8806	0.9659	0.0454	0.9463	0.9659	0.0454	0.9463
β_1	5	4.9927	0.0083	0.9522	4.9926	0.0081	0.9493	4.9926	0.0081	0.9493
ρ_0	0.8	0.7139	0.0182	0.9373	0.7817	0.0067	0.9612	0.7817	0.0067	0.9582
σ_ϵ^2	2	2.2285	0.2808	0.8627	1.8757	0.2600	0.9612	1.8757	0.2598	0.9582
σ_y^2	1	0.9166	0.2770	0.8507	1.1203	0.2700	0.9403	1.1203	0.2699	0.9403

Table S2.6: H-SAM mean estimates (est.), mean squared error (m.s.e.), and coverage (cov.) of estimated parameters by the ODM, the marginal ML method, and the EM algorithm - 30% of missing data. These attributes are computed from 2500 simulated datasets, each with $n = 625$ units.

Parameter	True value	ODM			CLM					
		ML			Marginal ML			EM algorithm		
		est.	m.s.e	cov.	est.	m.s.e	cov.	est.	m.s.e	cov.
β_0	1	1.9794	1.0589	0.0054	1.0039	0.0046	0.9458	1.0039	0.0046	0.9458
β_1	5	5.3251	0.1262	0.3469	4.9994	0.0079	0.9485	4.9994	0.0079	0.9485
ρ	0.8	0.6136	0.0360	0.0000	0.7998	0.0001	0.9512	0.7998	0.0001	0.9512
σ_ϵ^2	2	1.5454	0.6575	0.9079	2.0069	0.1144	0.9377	2.0069	0.1142	0.9377
σ_y^2	1	5.3717	20.1868	0.0000	0.9734	0.0495	0.9350	0.9734	0.0494	0.9350

Table S2.7: H-SEM mean estimates (est.), mean squared error (m.s.e.), and coverage (cov.) of estimated parameters by the ODM, the marginal ML method, and the EM algorithm - 70% of missing data. These attributes are computed from 2500 simulated datasets, each with $n = 625$ units.

Parameter	True value	ODM			CLM					
		ML			Marginal ML			EM algorithm		
		est.	m.s.e	cov.	est.	m.s.e	cov.	est.	m.s.e	cov.
β_0	1	0.9863	0.0809	0.7903	0.9797	0.0820	0.9032	0.9796	0.0820	0.9032
β_1	5	4.9916	0.0147	0.9886	4.9897	0.0153	0.9839	4.9897	0.0153	0.9839
ρ	0.8	0.5302	0.1331	0.6613	0.7726	0.0136	0.9032	0.7726	0.0135	0.9032
σ_ϵ^2	2	2.0334	1.8126	0.8065	1.7241	1.0984	0.9677	1.7248	1.0915	0.9677
σ_y^2	1	1.6499	2.4837	0.8065	1.2912	0.8625	0.8548	1.2901	0.8574	0.8548

In the case of fitting the H-SEM with 70% missing values, the MSEs are lower, and the coverage is generally higher for estimates derived from the CLM specification compared to estimates from the ODM specification. See Table S2.7 for further details. In contrast to scenarios with lower missing value percentages such as 10%, 20%, and 30%, the estimate for ρ from ODM considerably deviates from the true value.

Table S2.8: H-SAM mean estimates (est.), mean squared error (m.s.e.), and coverage (cov.) of estimated parameters by the ODM, the marginal ML method, and the EM algorithm - 70% of missing data. These attributes are computed from 2500 simulated datasets, each with $n = 625$ units.

Parameter	True value	ODM			CLM					
		ML			Marginal ML			EM algorithm		
		est.	m.s.e	cov.	est.	m.s.e	cov.	est.	m.s.e	cov.
β_0	1	3.6927	7.5528	0.0000	1.0044	0.0073	0.9386	1.0044	0.0073	0.9386
β_1	5	5.9984	1.0796	0.0702	5.0053	0.0212	0.9474	5.0053	0.0212	0.9474
ρ	0.8	0.3444	0.2095	0.0000	0.7987	0.0002	0.9561	0.7987	0.0002	0.9561
σ_ϵ^2	2	0.0387	3.8995	0.9754	2.0980	0.4133	0.9649	2.0970	0.4096	0.9649
σ_y^2	1	15.1225	205.5219	0.0526	0.9440	0.1491	0.9123	0.9444	0.1482	0.9123

Table S2.9: H-SEM mean estimates (est.), mean squared error (m.s.e.), and coverage (cov.) of estimated parameters by the ODM, the marginal ML method, and the EM algorithm - 80% of missing data. These attributes are computed from 2500 simulated datasets, each with $n = 625$ units.

Parameter	True value	ODM			CLM					
		ML			Marginal ML			EM algorithm		
		est.	m.s.e	cov.	est.	m.s.e	cov.	est.	m.s.e	cov.
β_0	1	0.9424	0.0583	0.9000	0.9522	0.0514	0.9734	0.9522	0.0514	0.9734
β_1	5	4.9346	0.0306	0.9333	4.9342	0.0323	0.9333	4.9343	0.0323	0.9333
ρ	0.8	0.5012	0.1742	0.8000	0.8162	0.0105	0.8333	0.8159	0.0105	0.8333
σ_ϵ^2	2	2.1295	1.3991	0.8333	1.9895	0.8954	0.8667	1.9879	0.8903	0.8667
σ_y^2	1	1.3452	2.1345	0.8000	0.8016	0.4482	0.8000	0.8021	0.4469	0.8000

Fitting the H-SEM with 80% missing values using the ODM specification leads to inaccurate estimates compared to the CLM specification. The ODM specification has higher MSE values compared to those of CLM for the parameters ρ , σ_ϵ^2 and σ_y^2 , as displayed in Table S2.9.

Table S2.10: H-SAM mean estimates (est.), mean squared error (m.s.e.), and coverage (cov.) of estimated parameters by the ODM, the marginal ML method, and the EM algorithm - 80% of missing data. These attributes are computed from 2500 simulated datasets, each with $n = 625$ units.

Parameter	True value	ODM			CLM					
		ML			Marginal ML			EM algorithm		
		est.	m.s.e	cov.	est.	m.s.e	cov.	est.	m.s.e	cov.
β_0	1	4.0162	9.5202	0.0000	0.9838	0.0117	0.8605	0.9838	0.0117	0.8605
β_1	5	6.0322	1.2216	0.2093	4.9542	0.0420	0.9302	4.9543	0.0420	0.9302
ρ	0.8	0.2737	0.2799	0.0000	0.8017	0.0003	0.9070	0.8017	0.0003	0.9070
σ_ϵ^2	2	0.3197	3.7892	0.9877	2.2445	0.7745	0.9535	2.2406	0.7614	0.9535
σ_ϵ^2	1	16.2810	241.6783	0.6047	0.8367	0.2287	0.8837	0.8384	0.2257	0.9070

In Section 3.3 of the main paper, we discuss the limitations associated with employing the EM algorithm when dealing with large values of n_u . When dealing with datasets with 90% missing values, the EM algorithm exhibits slow convergence. As a result, we only compare the estimates from the ODM with the estimates from the marginal ML method.

Tables S2.11 and S2.12 present the mean estimates, the MSE, and the coverage of estimates obtained from the ODM and CLM estimated using the marginal ML method with 90% missing values.

Table S2.11: H-SEM mean estimates (est.), mean squared error (m.s.e.), and coverage (cov.) of estimated parameters by the ODM, the marginal ML method, and the EM algorithm - 90% of missing data. These attributes are computed from 2500 simulated datasets, each with $n = 625$ units. Entries marked with 'NC' indicate that the EM algorithm failed to converge for a significant number of simulations, preventing the reliable calculation of mean estimates, MSE, and coverages.

Parameter	True value	ODM			CLM					
		ML			Marginal ML			EM algorithm		
		est.	m.s.e	cov.	est.	m.s.e	cov.	est.	m.s.e	cov.
β_0	1	0.9966	0.1101	0.8848	0.9939	0.1077	0.9168	NC	NC	NC
β_1	5	5.0103	0.0738	0.8542	5.0141	0.0706	0.9424	NC	NC	NC
ρ	0.8	0.2781	0.4004	0.6112	0.5547	0.2621	0.8384	NC	NC	NC
σ_ϵ^2	2	1.1193	3.2489	0.9360	0.9749	2.5725	0.9808	NC	NC	NC
σ_y^2	1	2.7998	6.2493	0.8588	1.9075	2.4003	0.9040	NC	NC	NC

The MSEs of estimates obtained from the CLM specifications are consistently lower than those from the ODM specification for the H-SEM with 90% missing values, as illustrated in Table S2.11. These results imply that the CLM specification displays greater robustness to high percentages of missing data, resulting in more accurate estimates for the H-SEM.

Table S2.12: H-SAM mean estimates (est.), mean squared error (m.s.e.), and coverage (cov.) of estimated parameters by the ODM, the marginal ML method, and the EM algorithm - 90% of missing data. These attributes are computed from 2500 simulated datasets, each with $n = 625$ units. Entries marked with 'NC' indicate that the EM algorithm failed to converge for a significant number of simulations, preventing the reliable calculation of mean estimates, MSE, and coverages.

Parameter	True value	ODM			CLM					
		ML			Marginal ML			EM algorithm		
		est.	m.s.e	cov.	est.	m.s.e	cov.	est.	m.s.e	cov.
β_0	1	4.4343	12.6305	0.0045	1.0019	0.0209	0.9189	NC	NC	NC
β_1	5	6.2834	2.0270	0.4009	5.0123	0.0812	0.9279	NC	NC	NC
ρ	0.8	0.2849	0.2750	0.0180	0.7991	0.0007	0.9009	NC	NC	NC
σ_ϵ^2	2	3.7648	47.2668	0.9775	2.2564	1.5172	0.9910	NC	NC	NC
σ_y^2	1	13.8161	213.6280	0.9640	0.7381	0.4297	0.9595	NC	NC	NC

Similar to the scenarios with missing percentages of 10%, 20%, 30%, 70%, and 80%, the CLM specification consistently yields estimates that are closer to the true parameters, exhibiting lower MSEs and higher coverages compared to ODM for H-SAM with 90% of missing values.

In summary, we find that estimating the H-SAM with ODM specification results in inaccurate parameter estimates for any level of missing value percentage, whereas CLM consistently produces highly accurate parameter estimates. Consequently, we recommend adopting the CLM specification for estimating H-SAM regardless of the missing value percentage. On the other hand, for H-SEM, when dealing with a low percentage of missing values (10%, 20%, 30%), both ODM and CLM provide accurate parameter estimates. However, as the missing value percentage increases to 70%, 80%, and 90%, ODM produces estimates that are considerably different from the true values. In contrast, CLM yields estimates closer to the true values, demonstrating lower MSE, particularly for parameters such as ρ , σ_ϵ^2 and σ_y^2 . Therefore, we recommend utilising CLM for H-SEM, especially in scenarios with higher percentages of missing values.

S3 Extended Real data analysis results

Table S3.1 displays estimates and standard errors for fixed effects (β 's), ρ , σ_ϵ^2 , and σ_y^2 for the H-SAM with 10% missing values obtained using FDM, ODM, and CLM estimated using the marginal ML method and the EM algorithm. The table includes the marginal log-likelihood values for the marginal ML and the EM methods, the computing time, the time taken to compute the standard errors of the parameter estimates, and the number of iterations for the EM algorithm.

According to the results presented in the complexity analysis in Section 4.1 of the main paper, when dealing with low percentages of missing values, the EM algorithm is generally expected to offer computational advantages over the marginal ML method for both H-SEM and H-SAM. However, contrary to this expectation, as indicated in the third last row of Table S3.1, the EM algorithm takes longer to converge than the marginal ML method. This observation can be attributed to the high number of iterations (10) in the EM algorithm. It is important to note that despite the difference in computation time, both marginal ML and EM algorithm methods produce identical parameter estimates. Furthermore, parameter estimates obtained from CLM are found to be closer to the estimates obtained from the FDM than those from the ODM, which relies solely on observed data.

Table S3.4 displays estimates and standard errors for fixed effects (β 's), ρ , σ_ϵ^2 , and σ_y^2 for the H-SEM with 90% missing values obtained using FDM, ODM, and CLM estimated using the marginal ML method and the EM algorithm. The table includes the marginal log-likelihood values for the marginal ML and the EM methods, the computing time, the time taken to compute the standard errors of the parameter estimates, and the number of iterations for the EM algorithm.

It is important to highlight that, similar to the case of fitting H-SAM with 90% missing data as discussed in Section 5 of the main paper, we terminated the EM algorithm for H-SEM with 90% missing data after 100 iterations (after about 49 hours and 53 minutes) due to computational constraints. At this point, the EM algorithm had not reached convergence, and its execution time was significantly longer in comparison to the marginal ML method. This observation aligns with the conclusions drawn from our computational complexity study in Section 4.1 of the main paper, emphasising that when faced with a substantial proportion of missing data, the marginal ML method demonstrates superior computational efficiency compared to the EM algorithm. Note that, in this example, the computational cost of the marginal ML and EM algorithms is large because the total number of observations (n) is large.

Table S3.1: Parameter estimates (est.) with their standard errors (S.E.), marginal log-likelihood values for the marginal ML and the EM algorithms, estimation time and the time to calculate the standard errors of the parameters (Time std.error) in seconds for fitting H-SAM using FDM-ML for the full data set, and ODM-ML, CLM-Marginal ML, and CLM-EM for a partially observed data set with 10% of missing data, and the number of EM iterations.

	FDM-ML		ODM-ML		CLM-Marginal ML		CLM-EM	
	est.	S.E.	est.	S.E.	est.	S.E.	est.	S.E.
Intercept	-0.1124	0.0507	2.7726	0.0895	-0.0962	0.0538	-0.0962	0.0538
age	0.9565	0.0429	1.9633	0.0805	0.9512	0.0453	0.9512	0.0453
age ²	-1.5790	0.0797	-4.0116	0.1458	-1.5702	0.0848	-1.5701	0.0848
age ³	0.3697	0.0440	1.2107	0.0789	0.3714	0.0458	0.3714	0.0458
log(lotsize)	0.0413	0.0022	0.1724	0.0043	0.0400	0.0024	0.0400	0.0024
rooms	-0.0052	0.0026	0.0090	0.0044	-0.0045	0.0028	-0.0045	0.0028
log(TLA)	0.4454	0.0083	0.8930	0.0141	0.4351	0.0098	0.4351	0.0098
beds	0.0129	0.0039	-0.0174	0.0065	0.0111	0.0041	0.0111	0.0041
syear1994	0.0357	0.0066	0.0461	0.0106	0.0333	0.0069	0.0333	0.0069
syear1995	0.0710	0.0064	0.0835	0.0103	0.0670	0.0067	0.0670	0.0067
syear1996	0.0864	0.0063	0.1083	0.0100	0.0810	0.0065	0.0810	0.0065
syear1997	0.1191	0.0062	0.1451	0.0099	0.1134	0.0065	0.1134	0.0065
syear1998	0.1675	0.0064	0.2045	0.0102	0.1635	0.0067	0.1635	0.0067
ρ	0.6727	0.0001	0.0220	0.0300	0.6794	0.0051	0.6794	0.0051
σ_y^2	0.0399	0.0008	0.1745	0.0062	0.0399	0.0013	0.0399	0.0013
σ_ϵ^2	0.042	0.0009	0.0001	0.0016	0.0418	0.0012	0.0418	0.0012
number of EM iterations	—		—		—		10	
estimation time (s)	6.319		20.091		642.4235		1438.0863	
Time std.error (s)	5.926		9.0800		93.8317		85.8197	
marginal log-likelihood	—		—		-6838.894		-6838.894	

Table S3.2: Parameter estimates (est.) with their standard errors (S.E.), marginal log-likelihood values for the marginal ML and the EM algorithms, estimation time and the time to calculate the standard errors of the parameters (Time std.error) in seconds for fitting H-SAM using FDM-ML for the full data set, and ODM-ML, CLM-Marginal ML, and CLM-EM for a partially observed data set with 90% of missing data, and the number of EM iterations. Entries labeled as 'NC' indicate instances where the EM algorithm did not successfully converge.

	FDM-ML		ODM-ML		CLM-Marginal ML		CLM-EM	
	est.	S.E.	est.	S.E.	est.	S.E.	est.	S.E.
Intercept	-0.1124	0.0507	3.2539	0.2683	0.1863	0.1511	NC	
age	0.9565	0.0429	1.6800	0.2551	0.8215	0.1219	NC	
age ²	-1.5790	0.0797	-3.7745	0.4711	-1.2960	0.2327	NC	
age ³	0.3697	0.0440	1.1916	0.2583	0.0989	0.1264	NC	
log(lotsize)	0.0413	0.0022	0.1701	0.0124	0.0414	0.0064	NC	
rooms	-0.0052	0.0026	0.0296	0.0134	-0.0042	0.0082	NC	
log(TLA)	0.4454	0.0083	0.8581	0.0422	0.5234	0.0324	NC	
beds	0.0129	0.0039	-0.0312	0.0194	-0.0143	0.0125	NC	
syear1994	0.0357	0.0066	0.0620	0.0311	0.0631	0.0213	NC	
syear1995	0.0710	0.0064	0.0900	0.0305	0.0816	0.0207	NC	
syear1996	0.0864	0.0063	0.1005	0.0297	0.1010	0.0201	NC	
syear1997	0.1191	0.0062	0.1341	0.0290	0.1288	0.0200	NC	
syear1998	0.1675	0.0064	0.1629	0.0304	0.1518	0.0212	NC	
ρ	0.6727	0.0001	0.0027	0.0311	0.6046	0.0201	NC	
σ_y^2	0.0399	0.0008	0.0882	0.0114	0.0682	0.0070	NC	
σ_ϵ^2	0.042	0.0009	0.0882	0.0234	0.0203	0.0080	NC	
number of EM iterations	—		—		—		100	
estimation time (s)	6.319		0.216		18.7909		101916.2943	
Time std.error (s)	5.926		0.836		46.3919		41.6773	
marginal log-likelihood	—		—		-1120.645		NC	

Table S3.3: Parameter estimates (est.) with their standard errors (S.E.), marginal log-likelihood values for the marginal ML and the EM algorithms, estimation time and the time to calculate the standard errors of the parameters (Time std.error) in seconds for fitting H-SEM using FDM-ML for the full data set, and ODM-ML, CLM-Marginal ML, and CLM-EM for a partially observed data set with 10% of missing data, and the number of EM iterations.

	FDM-ML		ODM-ML		CLM-Marginal ML		CLM-EM	
	est.	S.E.	est.	S.E.	est.	S.E.	est.	S.E.
Intercept	5.2578	0.0748	4.3654	0.0831	5.2025	0.0793	5.2024	0.079
age	0.6994	0.0793	1.2398	0.0848	0.8220	0.0837	0.8203	0.0837
age ²	-1.7558	0.1321	-2.9150	0.1440	-1.9681	0.1398	-1.9681	0.1398
age ³	0.6355	0.0659	1.0907	0.0733	0.7360	0.0700	0.7348	0.0700
log(lotsize)	0.1458	0.0046	0.2008	0.0048	0.1482	0.0049	0.1481	0.0049
rooms	0.0056	0.0029	0.0069	0.0034	0.0062	0.0031	0.0062	0.0031
log(TLA)	0.6038	0.0103	0.6606	0.0119	0.6059	0.0109	0.6059	0.0109
beds	0.0164	0.0043	0.0131	0.0050	0.0164	0.0045	0.0164	0.0045
syear1994	0.0365	0.0067	0.0400	0.0079	0.0339	0.0072	0.0339	0.0072
syear1995	0.0799	0.0066	0.0805	0.0077	0.0781	0.0070	0.0781	0.0070
syear1996	0.0962	0.0064	0.1006	0.0075	0.0940	0.0068	0.0940	0.0068
syear1997	0.1413	0.0063	0.1438	0.0074	0.1393	0.0067	0.1393	0.0068
syear1998	0.1937	0.0065	0.1954	0.0076	0.1937	0.0069	0.1937	0.0069
ρ	0.9866	0.0002	0.8654	0.0049	0.9868	0.0002	0.9868	0.0002
σ_y^2	0.0004	0.0001	0.0226	0.0011	0.0004	0.0001	0.0004	0.0001
σ_ϵ^2	0.0685	0.0007	0.0603	0.0012	0.0691	0.0007	0.0691	0.0007
number of EM iterations	—		—		—		8	
estimation time (s)	14.249		4.963		1803.9157		1546.3783	
Time std.error (s)	5.849		5.52		59.2792		65.7387	
marginal log-likelihood	—		—		-5885.136		-5885.098	

Table S3.4: Parameter estimates (est.) with their standard errors (S.E.), marginal log-likelihood values for the marginal ML and the EM algorithms, estimation time and the time to calculate the standard errors of the parameters (Time std.error) in seconds for fitting H-SEM using FDM-ML for the full data set, and ODM-ML, CLM-Marginal ML, and CLM-EM for a partially observed data set with 90% of missing data, and the number of EM iterations. Entries labeled as 'NC' indicate instances where the EM algorithm did not successfully converge.

	FDM-ML		ODM-ML		CLM-Marginal ML		CLM-EM	
	est.	S.E.	est.	S.E.	est.	S.E.	est.	S.E.
Intercept	5.2578	0.0748	3.3498	0.2640	3.4952	0.2446	NC	
age	0.6994	0.0793	1.5135	0.2553	1.1073	0.2472	NC	
age ²	-1.7558	0.1321	-3.4992	0.4678	-2.6208	0.4332	NC	
age ³	0.6355	0.0659	1.1014	0.2548	0.7277	0.2287	NC	
log(lotsize)	0.1458	0.0046	0.1806	0.0125	0.1743	0.0133	NC	
rooms	0.0056	0.0029	0.0264	0.0128	0.0021	0.0110	NC	
log(TLA)	0.6038	0.0103	0.8340	0.0411	0.8300	0.0365	NC	
beds	0.0164	0.0043	-0.0241	0.0187	-0.0049	0.0165	NC	
syear1994	0.0365	0.0067	0.0660	0.0299	0.0770	0.0258	NC	
syear1995	0.0799	0.0066	0.0862	0.0294	0.1206	0.0253	NC	
syear1996	0.0962	0.0064	0.0977	0.0287	0.1384	0.0247	NC	
syear1997	0.1413	0.0063	0.1386	0.0277	0.1744	0.0242	NC	
syear1998	0.1937	0.0065	0.1575	0.0293	0.2148	0.0253	NC	
ρ	0.9866	0.0002	0.6866	0.0012	0.9936	0.0006	NC	
σ_y^2	0.0004	0.0001	0.1643	0.0030	0.0001	0.0002	NC	
σ_ϵ^2	0.0685	0.0007	0.0001	0.0081	0.0837	0.0035	NC	
number of EM iterations	—		—		—		100	
estimation time (s)	14.249		0.968		52.02428		179518.2548	
Time std.error (s)	5.849		0.363		42.44297		44.4597	
marginal log-likelihood	—		—		-1159.595		NC	

S4 Additional proofs for the marginal ML method

This section derives the analytical forms of ML estimators for β and ω for the proposed marginal ML method presented in Section 3.2 of the main paper. By differentiating the

marginal log-likelihood for $\mathbf{z}_o$ in Equation (3.3) in the main paper with respect to $\boldsymbol{\beta}$ we get

$$\frac{\partial \log f(\mathbf{z}_o; \omega, \boldsymbol{\theta}, \rho, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\frac{1}{2\omega} \frac{\partial (\mathbf{r}_o^\top \mathbf{V}_{oo}^{-1} \mathbf{r}_o)}{\partial \boldsymbol{\beta}}, \quad (\text{S4.1})$$

by substituting $\mathbf{r}_o = \mathbf{z}_o - \tilde{\mathbf{X}}_o \boldsymbol{\beta}$,

$$\begin{aligned} \frac{\partial \log f(\mathbf{z}_o; \omega, \boldsymbol{\theta}, \rho, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= -\frac{1}{2\omega} \frac{\partial (\mathbf{z}_o - \tilde{\mathbf{X}}_o \boldsymbol{\beta})^\top \mathbf{V}_{oo}^{-1} (\mathbf{z}_o - \tilde{\mathbf{X}}_o \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\ &= -\frac{2}{2\omega} (-\tilde{\mathbf{X}}_o)^\top \mathbf{V}_{oo}^{-1} (\mathbf{z}_o - \tilde{\mathbf{X}}_o \boldsymbol{\beta}) \\ &= \frac{1}{\omega} (\tilde{\mathbf{X}}_o^\top \mathbf{V}_{oo}^{-1} \mathbf{z}_o - \tilde{\mathbf{X}}_o^\top \mathbf{V}_{oo}^{-1} \tilde{\mathbf{X}}_o \boldsymbol{\beta}). \end{aligned} \quad (\text{S4.2})$$

Then, by setting Equation (S4.2) to zero, we obtain the ML estimator for $\boldsymbol{\beta}$

$$\begin{aligned} 0 &= \frac{1}{\omega} (\tilde{\mathbf{X}}_o^\top \mathbf{V}_{oo}^{-1} \mathbf{z}_o - \tilde{\mathbf{X}}_o^\top \mathbf{V}_{oo}^{-1} \tilde{\mathbf{X}}_o \boldsymbol{\beta}) \\ \tilde{\mathbf{X}}_o^\top \mathbf{V}_{oo}^{-1} \tilde{\mathbf{X}}_o \boldsymbol{\beta} &= \tilde{\mathbf{X}}_o^\top \mathbf{V}_{oo}^{-1} \mathbf{z}_o \\ \hat{\boldsymbol{\beta}}(\rho, \theta) &= \left(\tilde{\mathbf{X}}_o^\top \mathbf{V}_{oo}^{-1} \tilde{\mathbf{X}}_o \right)^{-1} \tilde{\mathbf{X}}_o^\top \mathbf{V}_{oo}^{-1} \mathbf{z}_o. \end{aligned} \quad (\text{S4.3})$$

Similarly, the ML estimator for ω is derived by differentiating (3.3) with respect to ω as

$$\frac{\partial \log f(\mathbf{z}_o; \omega, \boldsymbol{\theta}, \rho, \boldsymbol{\beta})}{\partial \omega} = -\frac{n_o}{2\omega} + \frac{1}{2\omega^2} \mathbf{r}_o^\top \mathbf{V}_{oo}^{-1} \mathbf{r}_o. \quad (\text{S4.4})$$

By setting Equation (S4.4) to zero, the ML estimator for $\hat{\omega}$ is

$$\hat{\omega}(\rho, \theta) = \frac{\mathbf{r}_o^\top \mathbf{V}_{oo}^{-1} \mathbf{r}_o}{n_o}. \quad (\text{S4.5})$$

S5 Additional proof for the EM algorithm

S5.1 Maximum Likelihood estimators for the parameters $\boldsymbol{\beta}$ and ω

This section derives the analytical forms of ML estimators for $\boldsymbol{\beta}$ and ω for the proposed EM algorithm presented in Section 3.3 of the main paper. By differentiating $Q(\boldsymbol{\phi} \mid \boldsymbol{\phi}')$ in Equation (3.15) in the main paper with respect to $\boldsymbol{\beta}$, we obtain

$$\frac{\partial Q(\boldsymbol{\phi} \mid \boldsymbol{\phi}')}{\partial \boldsymbol{\beta}} = -\frac{1}{2\omega} \frac{\partial (\mathbf{r}_{u|o}^\top \mathbf{M} \mathbf{r}_{u|o})}{\partial \boldsymbol{\beta}}, \quad (\text{S5.1})$$

and by substituting $\mathbf{r}_{u|o} = E_{u|o}(\mathbf{z}) - \boldsymbol{\mu}$, we obtain

$$\begin{aligned}\frac{\partial Q(\phi | \phi')}{\partial \boldsymbol{\beta}} &= -\frac{1}{2\omega} \frac{\partial (E_{u|o}(\mathbf{z}) - \boldsymbol{\mu})^\top \mathbf{M} (E_{u|o}(\mathbf{z}) - \boldsymbol{\mu})}{\partial \boldsymbol{\beta}} \\ &= -\frac{2}{2\omega} (-\tilde{\mathbf{X}})^\top \mathbf{M}^\top (E_{u|o}(\mathbf{z}) - \boldsymbol{\mu}) \\ &= \frac{1}{\omega} (\tilde{\mathbf{X}}^\top \mathbf{M} E_{u|o}(\mathbf{z}) - \tilde{\mathbf{X}}^\top \mathbf{M} \tilde{\mathbf{X}} \boldsymbol{\beta})\end{aligned}\tag{S5.2}$$

Then, by setting Equation (S5.2) to zero, we obtain the ML estimator for $\boldsymbol{\beta}$,

$$\begin{aligned}0 &= \frac{1}{\omega} (\tilde{\mathbf{X}}^\top \mathbf{M} E_{u|o}(\mathbf{z}) - \tilde{\mathbf{X}}^\top \mathbf{M} \tilde{\mathbf{X}} \boldsymbol{\beta}) \\ \tilde{\mathbf{X}}^\top \mathbf{M} \tilde{\mathbf{X}} \boldsymbol{\beta} &= \tilde{\mathbf{X}}^\top \mathbf{M} E_{u|o}(\mathbf{z}) \\ \hat{\boldsymbol{\beta}}(\rho, \theta) &= \left(\tilde{\mathbf{X}}^\top \mathbf{M} \tilde{\mathbf{X}} \right)^{-1} \tilde{\mathbf{X}}^\top \mathbf{M} E_{u|o}(\mathbf{z}).\end{aligned}\tag{S5.3}$$

Similarly, the ML estimator for ω is derived by differentiating the conditional expectation in Equation (3.15) in the main paper with respect to ω as

$$\frac{\partial Q(\phi | \phi')}{\partial \omega} = -\frac{n}{2\omega} + \frac{1}{2\omega^2} \mathbf{r}_{u|o}^\top \mathbf{M} \mathbf{r}_{u|o}.\tag{S5.4}$$

By setting Equation (S5.4) to zero, the ML estimator for $\hat{\omega}$ is,

$$\hat{\omega}(\rho, \theta) = \frac{\mathbf{r}_{u|o}^\top \mathbf{M} \mathbf{r}_{u|o}}{n}.\tag{S5.5}$$

S5.2 Simplifying other terms in the EM algorithm.

This section simplifies the expressions for the terms $\mathbf{M}$, $\log|\mathbf{M}|$, $\tilde{\mathbf{X}}^\top \mathbf{M} \tilde{\mathbf{X}}$, and $\mathbf{r}_{u|o}^\top \mathbf{M} \mathbf{r}_{u|o}$ found in the EM algorithm presented in Section 3.3 of the main paper. This simplification aims to reduce the computational burden associated with direct calculations. First, the matrix $\mathbf{M}$ can be expressed as

$$\begin{aligned}\mathbf{M} = \mathbf{V}^{-1} &= (\mathbf{I}_n + \theta(\mathbf{A}^\top \mathbf{A})^{-1})^{-1} \\ &= \mathbf{A}^\top (\mathbf{A}^\top)^{-1} (\mathbf{I}_n + \theta \mathbf{A}^{-1} (\mathbf{A}^\top)^{-1})^{-1} \mathbf{A}^{-1} \mathbf{A} \\ &= \mathbf{A}^\top (\mathbf{A} (\mathbf{I}_n + \theta \mathbf{A}^{-1} (\mathbf{A}^\top)^{-1}) \mathbf{A}^\top)^{-1} \mathbf{A} \\ &= \mathbf{A}^\top (\mathbf{A} \mathbf{A}^\top + \mathbf{I}_n \theta)^{-1} \mathbf{A}.\end{aligned}\tag{S5.6}$$

Second, the determinant of the $\log|\mathbf{M}|$ is expressed as

$$\begin{aligned}
|\mathbf{M}| &= |\mathbf{A}^\top (\mathbf{A}\mathbf{A}^\top + \mathbf{I}_n\theta)^{-1} \mathbf{A}| \\
|\mathbf{M}| &= |\mathbf{A}^\top| |(\mathbf{A}\mathbf{A}^\top + \mathbf{I}_n\theta)^{-1}| |\mathbf{A}| \\
\log|\mathbf{M}| &= \log|\mathbf{A}^\top| - \log|\mathbf{A}\mathbf{A}^\top + \mathbf{I}_n\theta| + \log|\mathbf{A}| \\
\log|\mathbf{M}| &= -\log|\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n| + \log|\mathbf{A}\mathbf{A}^\top|.
\end{aligned} \tag{S5.7}$$

We define the following two terms:

$$\mathbf{C} = (\mathbf{L}_{\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n})^{-1} \tilde{\mathbf{X}}_{\mathbf{A}}, \tag{S5.8}$$

and

$$\mathbf{c} = (\mathbf{L}_{\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n})^{-1} \mathbf{z}_{\mathbf{A}_{u|o}}, \tag{S5.9}$$

where $\mathbf{L}_{\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n}$ is the Cholesky factor of $\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n$, $\tilde{\mathbf{X}}_{\mathbf{A}} = \mathbf{A}\tilde{\mathbf{X}}$, and $\mathbf{z}_{\mathbf{A}_{u|o}} = \mathbf{A}E_{u|o}(\mathbf{z})$. Then, we express $\tilde{\mathbf{X}}^\top \mathbf{M} \tilde{\mathbf{X}}$ and $\mathbf{r}_{u|o}^\top \mathbf{M} \mathbf{r}_{u|o}$ as

$$\begin{aligned}
\tilde{\mathbf{X}}^\top \mathbf{M} \tilde{\mathbf{X}} &= \tilde{\mathbf{X}}^\top \mathbf{A}^\top (\mathbf{A}\mathbf{A}^\top + \mathbf{I}_n\theta)^{-1} \mathbf{A} \tilde{\mathbf{X}} \\
&= (\mathbf{A} \tilde{\mathbf{X}})^\top (\mathbf{A}\mathbf{A}^\top + \mathbf{I}_n\theta)^{-1} \mathbf{A} \tilde{\mathbf{X}} \\
&= \tilde{\mathbf{X}}_{\mathbf{A}}^\top (\mathbf{L}_{\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n} (\mathbf{L}_{\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n})^\top)^{-1} \tilde{\mathbf{X}}_{\mathbf{A}} \\
&= \tilde{\mathbf{X}}_{\mathbf{A}}^\top (\mathbf{L}_{\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n})^{-1\top} (\mathbf{L}_{\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n})^{-1} \tilde{\mathbf{X}}_{\mathbf{A}} \\
&= \left((\mathbf{L}_{\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n})^{-1} \tilde{\mathbf{X}}_{\mathbf{A}} \right)^\top (\mathbf{L}_{\mathbf{A}\mathbf{A}^\top + \theta\mathbf{I}_n})^{-1} \tilde{\mathbf{X}}_{\mathbf{A}} \\
&= \mathbf{C}^\top \mathbf{C}.
\end{aligned} \tag{S5.10}$$

and

$$\begin{aligned}
\mathbf{r}_{u|o}^\top \mathbf{M} \mathbf{r}_{u|o} &= (E_{u|o}(\mathbf{z}) - \boldsymbol{\mu})^\top \mathbf{M} (E_{u|o}(\mathbf{z}) - \boldsymbol{\mu}) \\
&= E_{u|o}(\mathbf{z})^\top \mathbf{M} E_{u|o}(\mathbf{z}) - 2E_{u|o}(\mathbf{z})^\top \mathbf{M} \tilde{\mathbf{X}} \boldsymbol{\beta} + \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{M} \tilde{\mathbf{X}} \boldsymbol{\beta}.
\end{aligned} \tag{S5.11}$$

To calculate the quadratic form presented in Equation (S5.11), we need to calculate $E_{u|o}(\mathbf{z})^\top \mathbf{M} E_{u|o}(\mathbf{z})$,

and $E_{u|o}(\mathbf{z})^\top \mathbf{M} \tilde{\mathbf{X}}$ as follows:

$$\begin{aligned}
E_{u|o}(\mathbf{z})^\top \mathbf{M} E_{u|o}(\mathbf{z}) &= E_{u|o}(\mathbf{z})^\top \mathbf{A}^\top (\mathbf{A} \mathbf{A}^\top + \mathbf{I}_n \theta)^{-1} \mathbf{A} E_{u|o}(\mathbf{z}) \\
&= (\mathbf{A} E_{u|o}(\mathbf{z}))^\top (\mathbf{A} \mathbf{A}^\top + \mathbf{I}_n \theta)^{-1} \mathbf{A} E_{u|o}(\mathbf{z}) \\
&= \left((\mathbf{L}_{\mathbf{A} \mathbf{A}^\top + \theta \mathbf{I}_n})^{-1} \mathbf{z}_{\mathbf{A}_{u|o}} \right)^\top \left((\mathbf{L}_{\mathbf{A} \mathbf{A}^\top + \theta \mathbf{I}_n})^{-1} \mathbf{z}_{\mathbf{A}_{u|o}} \right) \\
&= \mathbf{c}^\top \mathbf{c}.
\end{aligned} \tag{S5.12}$$

and,

$$\begin{aligned}
E_{u|o}(\mathbf{z})^\top \mathbf{M} \tilde{\mathbf{X}} &= E_{u|o}(\mathbf{z})^\top \mathbf{A}^\top (\mathbf{A} \mathbf{A}^\top + \mathbf{I}_n \theta)^{-1} \mathbf{A} \tilde{\mathbf{X}} \\
&= (\mathbf{A} E_{u|o}(\mathbf{z}))^\top (\mathbf{A} \mathbf{A}^\top + \mathbf{I}_n \theta)^{-1} \mathbf{A} \tilde{\mathbf{X}} \\
&= \left((\mathbf{L}_{\mathbf{A} \mathbf{A}^\top + \theta \mathbf{I}_n})^{-1} \mathbf{z}_{\mathbf{A}_{u|o}} \right)^\top (\mathbf{L}_{\mathbf{A} \mathbf{A}^\top + \theta \mathbf{I}_n})^{-1} \tilde{\mathbf{X}} \mathbf{A} \\
&= \mathbf{c}^\top \mathbf{C}.
\end{aligned} \tag{S5.13}$$

S6 Computing Information matrix

To obtain variances for the parameters of interest $\beta, \rho, \sigma_\epsilon^2$ and σ_y^2 , the standard procedure requires the calculation of either the expected or the observed information matrix. They are obtained by first calculating the second derivatives of the negative of marginal log-likelihood. The negative of marginal log-likelihood is

$$-L_o = -\log f(\mathbf{z}_o; \sigma_\epsilon^2, \sigma_y^2, \rho, \beta) = \frac{n_o}{2} \log(2\pi) + \frac{n_o}{2} \log(\sigma_\epsilon^2) - \frac{1}{2} \log |\mathbf{V}_{oo}^{-1}| + \frac{1}{2\sigma_\epsilon^2} \mathbf{r}_o^\top \mathbf{V}_{oo}^{-1} \mathbf{r}_o, \tag{S6.1}$$

and for the convenience of notation, we denote the negative marginal log-likelihood $-L_o = \bar{L}_o$.

First, we discuss the H-SEM. The second derivative of $\bar{L}_o$ with respect to β

$$\begin{aligned}
\frac{\partial \bar{L}_o}{\partial \beta} &= \frac{1}{2\sigma_\epsilon^2} 2(-\mathbf{X}_o^\top) \mathbf{V}_{oo}^{-1} (\mathbf{z}_o - \mathbf{X}_o \beta) \\
&= -\frac{1}{\sigma_\epsilon^2} \mathbf{X}_o^\top \mathbf{V}_{oo}^{-1} \mathbf{r}_o,
\end{aligned} \tag{S6.2}$$

and since $E(\mathbf{r}) = E(\mathbf{z}_o - \mathbf{X}_o \beta) = 0$, it is straightforward to show that

$$E \left(\frac{\partial^2 \bar{L}_o}{\partial \beta \partial \rho} \right) = 0, \quad E \left(\frac{\partial^2 \bar{L}_o}{\partial \beta \partial \sigma_\epsilon^2} \right) = 0, \quad E \left(\frac{\partial^2 \bar{L}_o}{\partial \beta \partial \sigma_y^2} \right) = 0,$$

meaning that $\hat{\beta}$ is asymptotically independent from $\zeta = (\rho, \sigma_\epsilon^2, \sigma_y^2)^\top$. Further,

$$E \left(\frac{\partial \bar{L}_o}{\partial \beta \partial \beta^\top} \right) = \frac{1}{\sigma_\epsilon^2} \mathbf{X}_o^\top \mathbf{V}_{oo}^{-1} \mathbf{X}_o. \quad (\text{S6.3})$$

These results leads to $\sigma_\epsilon^2 \text{Cov}(\hat{\beta}) = (\mathbf{X}_o^\top \mathbf{V}_{oo}^{-1} \mathbf{X}_o)^{-1}$, and $\text{Cov}(\hat{\zeta}) = \left[E \left(\frac{\partial^2 \bar{L}_o}{\partial \zeta \partial \zeta^\top} \right) \right]^{-1}$. To calculate $\text{Cov}(\hat{\zeta})$, we recommend utilising the observed information matrix through numerical differentiation. This approach helps prevent the need for frequent and expensive inversions. Now, we consider the H-SAM, which is more complicated because $\mathbf{r}_o$ depends on ρ . The first derivative of $\bar{L}_o$ with respect to β is

$$\begin{aligned} \frac{\partial \bar{L}_o}{\partial \beta} &= \frac{1}{2\omega} 2(-(\mathbf{A}^{-1} \mathbf{X})_o^\top) \mathbf{V}_{oo}^{-1} (\mathbf{z}_o - (\mathbf{A}^{-1} \mathbf{X})_o \beta) \\ &= -\frac{1}{\sigma_\epsilon^2} (\mathbf{A}^{-1} \mathbf{X})_o^\top \mathbf{V}_{oo}^{-1} \mathbf{r}_o. \end{aligned} \quad (\text{S6.4})$$

Then, it is straightforward to show that

$$E \left(\frac{\partial \bar{L}_o}{\partial \beta \partial \beta^\top} \right) = \frac{1}{\sigma_\epsilon^2} (\mathbf{A}^{-1} \mathbf{X})_o^\top \mathbf{V}_{oo}^{-1} (\mathbf{A}^{-1} \mathbf{X})_o. \quad (\text{S6.5})$$

In addition, while $E \left(\frac{\partial^2 \bar{L}_o}{\partial \beta \partial \rho} \right) \neq 0$, it can be shown that $E \left(\frac{\partial^2 \bar{L}_o}{\partial \beta \partial \sigma_\epsilon^2} \right) = E \left(\frac{\partial^2 \bar{L}_o}{\partial \beta \partial \sigma_y^2} \right) = 0$, with

$$E \left(\frac{\partial^2 \bar{L}_o}{\partial \beta \partial \rho} \right) = \frac{1}{\sigma_\epsilon^2} \left((\mathbf{A}^{-1} \mathbf{X})_o^\top \mathbf{V}_{oo}^{-1} \frac{\partial (\mathbf{A}^{-1} \mathbf{X})_o}{\partial \rho} \beta \right), \quad (\text{S6.6})$$

where $\frac{\partial (\mathbf{A}^{-1} \mathbf{X})_o}{\partial \rho} = (\mathbf{A}^{-1} \mathbf{W} \mathbf{A}^{-1} \mathbf{X})_o$. We recommend employing numerical differentiation to compute the corresponding second derivative for the term $E \left(\frac{\partial^2 \bar{L}_o}{\partial \zeta \partial \zeta^\top} \right)$. This allows us to derive the observed information matrix without utilising expensive matrix computations.