Biometrics 000, 1–25 DOI: 000
000 0000

Assuming Independence in Spatial Latent Variable Models: Consequences and Implications of Misspecification

Francis K.C. Hui[1,*], Nicole A. Hill[2], and A.H. Welsh[1]

[1]Research School of Finance, Actuarial Studies & Statistics, Australian National University, Acton, ACT 2601, Australia

[2]Institute for Marine and Antarctic Studies, University of Tasmania, Hobart, TAS 7001, Australia

**email:* francis.hui@anu.edu.au

Summary: Multivariate spatial data, where multiple responses are simultaneously recorded across spatially indexed observational units, are routinely collected in a wide variety of disciplines. For example, the Southern Ocean Continuous Plankton Recorder survey collects records of zooplankton communities in the Indian sector of the Southern Ocean, with the aim of identifying and quantifying spatial patterns in biodiversity in response to environmental change. One increasingly popular method for modeling such data is spatial generalized linear latent variable models (GLLVMs), where the correlation across sites is captured by a spatial covariance function in the latent variables. However, little is known about the impact of misspecifying the latent variable correlation structure on inference of various parameters in such models. To address this gap in the literature, we investigate how misspecifying and assuming independence for the latent variables' correlation structure impacts estimation and inference in spatial GLLVMs. Through both theory and numerical studies, we show that performance of maximum likelihood estimation and inference on regression coefficients under misspecification depends on a combination of the response type, the magnitude of true regression coefficient and the corresponding loadings, and, most importantly, whether the corresponding covariate is (also) spatially correlated. On the other hand, estimation and inference of truly non-zero loadings and prediction of latent variables is consistently not robust to misspecification of the latent variable correlation structure.

Key words: Community Ecology; Factor Analysis; Loadings; Multivariate Abundance Data; Spatial; Spatio-temporal

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Multivariate spatial data, consisting of multiple responses recorded simultaneously across spatially indexed observational units, are routinely collected in a variety of disciplines. Such data are characterized by spatial correlations whose strength depends inversely on the geographic distance between units, and between response correlations for which we usually have little information *a priori* regarding their structure. This article is motivated by the Southern Ocean Continuous Plankton Recorder (SO-CPR) survey, an annual survey that collects count records of zooplankton communities using vessels traversing the Indian sector of the Southern Ocean (Hosie, 2020). Alongside species records, we have covariate information pertaining to the physical habitat and oceanographic environment, and one of the primary aims of the SO-CPR survey is to identify and quantify spatial patterns in biodiversity in response to environmental change.

In community ecology, one of the most popular emerging methods for analyzing multivariate spatial data is spatial generalized linear latent variable models (Thorson et al., 2015; Warton et al., 2016; Ovaskainen et al., 2016; Bjork et al., 2018; Shirota et al., 2019). This model is an extension of generalized linear latent variable models (GLLVMs, Skrondal and Rabe-Hesketh, 2004) for analyzing multi-response data, where the latent variables are given additional structure to model the spatial correlations, on top of their usual role in accounting for between species correlation. For instance, the latent variables are drawn from a multivariate normal distribution with zero mean vector and a correlation matrix parameterized using a Matérn correlation function based on some distance metric between sites. Outside of community ecology, such analyses are also referred to as spatial factor analysis (e.g., Zhu et al., 2005; Lopes et al., 2011).

Compared to the standard independence assumption for latent variables i.e., when the correlation matrix is set to the identity matrix, assuming a structured correlation matrix has two major implications. From a modeling perspective, it is evident that in the presence of spatial correlations, assuming independence in the latent variables results in clear model misspecification. There has been some research into the impacts of misspecifying the latent variables for estimation and inference in GLLVMs. However, the literature so far has focused on misspecification of the latent variable distribution rather than of the correlation structure, and the consequences on estimation and inference of the loadings and prediction of latent variables rather than the impact on regression coefficients (e.g., Ma and Genton, 2010; Irincheeva et al., 2012). Furthermore, the large sample theory developed in such previous research on misspecification in GLLVMs was motivated by non-spatial data. As such, assuming independence in the latent variables across units was a reasonable foundation on which to base developments. By contrast, with spatial data such theory does not carry over directly: one requires alternative asymptotic frameworks e.g., increasing domain, fixed domain infill, and it is not guaranteed that properties such as consistency can be necessarily achieved within all these frameworks The asymptotics of spatial and spatio-temporal modeling remains an active area of research (e.g., Lu and Tjostheim, 2014; Kurisu, 2019), and to our knowledge there has been no theoretical research into the large sample effects of misspecifying the correlation structure in spatial GLLVMs (although related empirical work has been done on this by Shirota et al., 2019).

From a computational perspective, and focusing on maximum likelihood estimation, spatial GLLVMs involve a high-dimensional integral and, usually, the inversion of a high-dimensional matrix during the estimation process. By contrast, estimation and inference assuming independence is substantially faster since the dimension of the integral in the likelihood is reduced and there are no covariance parameters to estimate. While considerable progress has been made into more efficient likelihood-based estimation and inference in GLLVMs and mixed models in general (Niku et al., 2019), as well as on computationally faster approaches for modeling spatial data in general, the substantial reduction in computation time brought about by assuming independence remains.

Motivated by the potential tradeoff between computation time and model misspecification, this article studies the consequences of assuming independence for the latent variable correlation structure on estimation and inference in GLLVMs applied to multivariate spatial data. We develop some large sample results for point estimation of regression coefficients in GLLVMs under misspecification of the correlation structure, and complement this with an extensive simulation study to examine how finite sample performance is affected by response type, the magnitude of the regression coefficient and corresponding loadings, and whether the measured environmental covariates are spatially correlated. The main findings and contributions of this article may be summarized as follows:

- For normally distributed responses, maximum likelihood estimates of all regression coefficients excluding the intercept are estimation consistent. For non-normally distributed responses, only a weaker zero consistency result can be obtained. The exception is responses that are uncorrelated with all other responses, for which we (also) obtain estimation consistency of all the regression coefficients (Section 3).
- For covariates that have little to no spatial correlation, empirical studies show that for normal, negative binomial, and Tweedie distributed responses, point estimation of regression coefficients is largely robust to misspecification, and sandwich-based confidence intervals are close to their the nominal significance level (Sections 4.2–4.3). For binary responses, point estimation and sandwich-based confidence intervals only perform well under misspecification when either the magnitude of the true regression coefficient or the magnitude of its corresponding loadings are not especially large (Sections 4.4).
- For covariates that have moderate to strong spatial correlation, assuming independence of the latent variables almost always leads to poor estimation and inferential performance

for regression coefficients, irrespective of the response type and the magnitude of the true coefficient. The one exception is coefficients corresponding to responses that are uncorrelated with all other responses, which remain robust to misspecification when the covariates are spatially correlated (Sections 4.2–4.4).

- Irrespective of the response type, estimation and inference of the non-zero loadings and prediction of latent variables in the GLLVM is not robust to misspecification of the latent variable correlation structure. The one exception is responses that have entirely zero loadings i.e, responses that are independent of all other responses, for which a zero consistency result can be obtained for the estimates of the loadings under misspecification (Sections 4.2–4.4).

The results we obtain bear some resemblance to existing research on estimation and inference for spatial data as well as correlated data more generally. One article of particular relevance here is Shirota et al. (2019), who empirically found that spatial GLLVMs consistently outperform independent GLLVMs when it comes to prediction of the linear predictor and estimation of the residual covariance matrix (formed from the loading matrix). The findings here are concordant with these previous results, although in this article we perform a broader investigation to develop large sample results as well as examine the critical issue of estimation and inference of the regression coefficients (which Shirota et al., 2019, did not focus on). In the more general spatial regression setting, several methods proposed for variable selection (usually assuming spatially uncorrelated covariates) have demonstrated that assuming independence across sites does not affect large sample selection consistency results, although in finite samples the loss of efficiency and overall consequences can be pronounced (Hoeting et al., 2006; Wang and Zhu, 2009; Xu et al., 2015). Also, previous research into generalized linear mixed models has largely shown that estimation and inference on fixed effects is fairly robust to misspecification of the random effects distribution, but that random effects inference is *not* robust to the same sort of misspecification (Hui et al., 2020). This article builds on such exisiting research, and presents several new findings and consequences. For instance, the lack of robustness under misspecification for inference on spatially correlated covariates has major implications on how multivariate spatial data should be analyzed. In most spatially indexed observational studies in ecology, it is either known *a priori* or evidence is uncovered empirically that both the responses and one or more of the environmental predictors are spatially correlated. When analyzing such data, our results demonstrate that it is imperative that spatial correlation in the latent variables be taken into account. Otherwise, one risks inference being incorrect for regression coefficients corresponding to the spatially correlated covariates e.g., coverage probabilities of confidence intervals substantially below the nominated level, as well as poor predictive performance of the model as a whole (see also Yoon and Welsh, 2020, for similar results in the context of linear mixed models). We conclude our comparison by applying both a spatial GLLVM and a standard GLLVM assuming independence of the latent variables to the SO-CPR survey. For both models, we obtained some but not entirely similar results with regards to identifying the effects of key environmental covariates on the distributions of 24 zooplankton species across 3,900 spatial locations.

2. Spatial GLLVMs

We establish the notation of spatial GLLVMs with ecological data in mind. Let $y_j(\boldsymbol{s}_i)$ denote the observed response for species $j = 1, \ldots, p$ at site (observational unit) $\boldsymbol{s}_i \in \mathcal{D}; i = 1, \ldots, n$ in some spatial domain $\mathcal{D}$. The spatial domain can be continuously (e.g., geostatistical data) or discretely (e.g., lattice data) spatially indexed, or both, and we make no explicit restrictions on this (subject to Assumption 1 discussed below).We also observe a q-vector of covariates at each site, $\boldsymbol{x}(\boldsymbol{s}_i) = (x_1(\boldsymbol{s}_i), \ldots, x_q(\boldsymbol{s}_i))^\top$, representing physical environment or habitat. We assume $x_1(\boldsymbol{s}_i) = 1$ for $i = 1, \ldots, n$ to represent an intercept term, while the other terms are centered to have expectation zero, $\mathrm{E}_X\{x_k(\boldsymbol{s})\} = 0$ for all $k = 2, \ldots, q$.

The responses are assumed to be generated from a spatial GLLVM as follows: conditional on a d-vector of latent variables $\boldsymbol{u}(\boldsymbol{s})$ with $d \ll p$, as well as the covariates, the elements of the response vector $\boldsymbol{y}(\boldsymbol{s}) = (y_1(\boldsymbol{s}), \ldots, y_p(\boldsymbol{s}))^\top$ are independent observations from the exponential family of distributions. That is, $f\{y_j(\boldsymbol{s})|\boldsymbol{x}(\boldsymbol{s}), \boldsymbol{u}(\boldsymbol{s})\} = \exp\{\phi_j^{-1}[y_j(\boldsymbol{s})\vartheta_j(\boldsymbol{s}) - a\{\vartheta_j(\boldsymbol{s})\}] + c(y_j(\boldsymbol{s}), \phi_j)\}$ for known functions $a(\cdot)$ and $c(\cdot,\cdot)$, where $\vartheta_j(\boldsymbol{s})$ is the canonical parameter and ϕ_j is species-specific dispersion parameter. The mean, $\mathrm{E}_{Y|U,X}\{y_j(\boldsymbol{s})|\boldsymbol{x}(\boldsymbol{s}), \boldsymbol{u}(\boldsymbol{s})\} = \mu_j(\boldsymbol{s}) = \mu_j(\boldsymbol{s}) = a'\{\vartheta_j(\boldsymbol{s})\}$, is modeled as $g\{\mu_j(\boldsymbol{s})\} = \eta_j(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})^\top\boldsymbol{\beta}_j + \boldsymbol{u}(\boldsymbol{s})^\top\boldsymbol{\lambda}_j$ for a known link function $g(\cdot)$, where $\boldsymbol{\beta}_j$ is the vector of species-specific regression coefficients and $\boldsymbol{\lambda}_j$ is the vector of species-specific loadings.

Write the full nd-vector of latent variables generically as $\boldsymbol{u}(S)$, where S indexes all sampled locations. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_p^\top)^\top$, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^\top$, and $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^\top, \ldots, \boldsymbol{\lambda}_p^\top)^\top$ denote the full vector of regression coefficients, dispersion parameters, and loadings respectively. The marginal likelihood for the spatial GLLVM is given by

$$L(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\theta}) = \int \prod_{i=1}^{n} \prod_{j=1}^{p} f\{y_j(\boldsymbol{s}_i)|\boldsymbol{x}(\boldsymbol{s}_i), \boldsymbol{u}(\boldsymbol{s}_i)\} f(\boldsymbol{u}(S)|\boldsymbol{\theta}) d\boldsymbol{u}(S), \tag{1}$$

where the latent variables are assumed to come from a multivariate normal distribution with zero mean vector and a correlation matrix parameterized by a vector $\boldsymbol{\theta}$, that is, $f(\boldsymbol{u}(S)|\boldsymbol{\theta}) = \mathcal{N}_{nd}\{\boldsymbol{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})\}$. Note $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is a correlation (as opposed to a covariance) matrix to avoid scale invariance in the GLLVM. It is the combination of covariance parameters $\boldsymbol{\theta}$ and loadings $\boldsymbol{\lambda}_j$ in (1) which models the between species and spatial correlations in the data; see the end of the following section for more discussion on the role of correlations between species versus spatial correlation. On the linear predictor scale, we have $\mathrm{Cov}\{\eta_j(\boldsymbol{s}_i), \eta_{j'}(\boldsymbol{s}_{i'})\} = \boldsymbol{\lambda}_j^\top \mathrm{Cov}\{\boldsymbol{u}(\boldsymbol{s}_i), \boldsymbol{u}(\boldsymbol{s}_{i'})\}\boldsymbol{\lambda}_{j'}$ for $i, i' = 1, \ldots, n$ and $j, j' = 1, \ldots, p$, where

$\text{Cov}\{\boldsymbol{u}(\boldsymbol{s}_i), \boldsymbol{u}(\boldsymbol{s}_{i'})\}$ depends on the precise structure of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. Thus we see that the between species correlation also depends on the two spatial locations of interest. One common choice for $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ in community ecology is to assume the d latent variables are independent and use a spatial covariance function for each latent variable (e.g., Thorson et al., 2015; Ovaskainen et al., 2016). If we let $u_l(\boldsymbol{s}_i)$ denote the l-th element in $\boldsymbol{u}(\boldsymbol{s}_i)$, then for $l = 1, \ldots, d$ we use the Matérn correlation function, $\text{Cov}\{u_l(\boldsymbol{s}_i), u_l(\boldsymbol{s}_{i'})\} = \rho(D(\boldsymbol{s}_i, \boldsymbol{s}_{i'}), \nu, \alpha_l)$, where $\rho(h, \nu, \alpha_l) = \{2^{\nu-1}\Gamma(\nu)\}^{-1}(h\alpha^{-1})^{\nu}\mathcal{K}_{\nu}(h\alpha_l^{-1})$ for smoothness $\nu > 0$ and spatial scale parameters $\alpha_l > 0$, where $\Gamma(\cdot)$ is the gamma function, $\mathcal{K}_{\nu}(\cdot)$ is the modified Bessel function of the second kind, and $D(\cdot, \cdot)$ is some distance metric. Note the scale parameter, but not the smoothness parameter, is allowed to be different across the latent variables. The vector of covariance parameters is given by $\boldsymbol{\theta} = (\nu, \alpha_1, \ldots, \alpha_d)^{\top}$, although the smoothness may be fixed *a priori* e.g., $\nu = 1$ is a recommended choice when $\mathcal{D} = \mathbb{R}^2$ (Lindgren and Rue, 2015). .

2.1 *Assuming Independence*

Of the various structures which can be imposed on the latent variables, the simplest one is to assume the correlation matrix is equal to the identity matrix, $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{I}_{nd}$. The latent variables are then assumed to be independently standard normally distributed, and the marginal log-likelihood for the GLLVM reduces to

$$L_{\text{Ind}}(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\lambda}) = \prod_{i=1}^{n} \left\{ \int \prod_{j=1}^{p} f\{y_j(\boldsymbol{s}_i)|\boldsymbol{x}(\boldsymbol{s}_i), \boldsymbol{u}(\boldsymbol{s}_i)\} h\{\boldsymbol{u}(\boldsymbol{s}_i)\} d\boldsymbol{u}(\boldsymbol{s}_i) \right\} = \prod_{i=1}^{n} L_{\text{Ind},i}(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\lambda}), \tag{2}$$

where $h\{\boldsymbol{u}(\boldsymbol{s}_i)\} = \mathcal{N}_d(\boldsymbol{0}, \boldsymbol{I}_d)$. Equation (2) is equivalent to the standard GLLVM advocated for use by Warton et al. (2016) for community ecology data, and is a common choice for modeling multivariate data in other disciplines (Skrondal and Rabe-Hesketh, 2004). With the independence structure, the latent variables can only model between species correlation: $\text{Cov}\{\eta_j(\boldsymbol{s}_i), \eta_{j'}(\boldsymbol{s}_{i'})\} = 0$, unless $i = i'$ in which case $\text{Cov}\{\eta_j(\boldsymbol{s}_i), \eta_{j'}(\boldsymbol{s}_i)\} = \boldsymbol{\lambda}_j^{\top}\boldsymbol{\lambda}_{j'}$.

It is clear that assuming spatial independence makes the likelihood function computationally much easier to maximize: the product over the n observational units is pulled outside the integral and there are no covariance parameters to estimate in (2). Of course the computational gain comes with a cost, and the presence of spatial correlation means that equation (2) is misspecified in the second moment. In fact, as seen above the independence assumption means that the expected counts of two species may be correlated at any particular site, but are otherwise uncorrelated across different sites. In Appendix A, we elaborate more on the two correlation structures present in a spatial GLLVM, noting that the focus of this article is misspecification of the spatial correlation structure, and not (also) on misspecification of the between species correlation.

3. Estimation Under Misspecification

For the set of observed data, consider $\log L(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\theta})$ in (1) as the log-likelihood corresponding to the true model. Define $\ell(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\theta}) = \log\{\int \prod_{j=1}^p f\{y_j(\boldsymbol{s})|\boldsymbol{x}(\boldsymbol{s}), \boldsymbol{u}(\boldsymbol{s})\} f(\boldsymbol{u}(\boldsymbol{s})|\boldsymbol{\theta}) d\boldsymbol{u}(\boldsymbol{s})\}$, and denote the full vector of true parameters as $(\boldsymbol{\beta}^0, \boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, \boldsymbol{\theta}^0)$ which satisfies $\mathrm{E}_{X,Y}\{\nabla \ell(\boldsymbol{\beta}^0, \boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, \boldsymbol{\theta}^0)\} = \mathbf{0}$. Next, consider $\log L_{\mathrm{Ind}}(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\lambda})$ in (2) as the log-likelihood corresponding to the misspecified GLLVM assuming independence, and let $\ell_{\mathrm{Ind}}(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\lambda}) = \log\{\int \prod_{j=1}^p f\{y_j(\boldsymbol{s})|\boldsymbol{x}(\boldsymbol{s}), \boldsymbol{u}(\boldsymbol{s})\} h\{\boldsymbol{u}(\boldsymbol{s})\} d\boldsymbol{u}(\boldsymbol{s})\}$. We define the Kullback-Leibler distance $D_{\mathrm{KL}}(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\lambda}) = \mathrm{E}_{X,Y}\{\ell(\boldsymbol{\beta}^0, \boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, \boldsymbol{\theta}^0) - \ell_{\mathrm{Ind}}(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\lambda})\}$, and consequently the vector of pseudo-true parameters $(\boldsymbol{\beta}^*, \boldsymbol{\phi}^*, \boldsymbol{\lambda}^*) = \arg\min_{\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\lambda}} D_{\mathrm{KL}}(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\lambda})$ which satisfy $\mathrm{E}_{X,Y}\{\nabla \ell_{\mathrm{Ind}}(\boldsymbol{\beta}^*, \boldsymbol{\phi}^*, \boldsymbol{\lambda}^*)\} = \mathbf{0}$.

Let $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\lambda}}) = \arg\max_{\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\lambda}} \log L_{\mathrm{Ind}}(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\lambda})$ denote the maximum likelihood estimator of the misspecified GLLVM. We base our large sample developments under the following general regularity condition.

Assumption 1: The appropriate conditions are assumed to be satisfied to ensure weak consistency of the maximum likelihood estimator under the misspecified GLLVM. That is, $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\lambda}}) \xrightarrow{p} (\boldsymbol{\beta}^*, \boldsymbol{\phi}^*, \boldsymbol{\lambda}^*)$ as $n \to \infty$.

The above assumption is rather general, and is chosen with reason. As reviewed in Section 1, large sample theory in spatial modeling remains an active area of research and is dependent both on the related asymptotic framework and the structure of the model itself (e.g., most theory has been developed assuming a univariate normal or continuous response with additive error structure, and some sort of expanding domain framework; Mardia and Marshall, 1984; Lu and Tjostheim, 2014). Our aim is to study misspecification in spatial GLLVMs *irrespective* of the framework selected, by starting from a general position where consistency of the maximum likelihood estimator towards a set of pseudo-true parameters (true parameters in the case of the true GLLVM) can be attained. By doing so, the results we develop will apply to any framework provided estimation consistency within that framework can be achieved. In Appendix B, we elaborate on this and provide a more concrete example under which Assumption 1 holds.

3.1 *Regression Coefficients*

We first consider the case of conditionally normally distributed responses and the identity link, $\mu_j(\boldsymbol{s}) = \eta_j(\boldsymbol{s})$.

THEOREM 1: *For conditionally normal responses, the regression coefficients satisfy $\boldsymbol{\beta}^* = \boldsymbol{\beta}^0$. Thus under Assumption 1, the maximum likelihood estimator of the misspecified GLLVM satisfies $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}^0$ as $n \to \infty$, irrespective of the values of $\boldsymbol{\lambda}^0$.*

All proofs are provided in Appendix B. With non-normal responses, a weaker zero consistency result can be obtained as follows. For a nonempty subset $\mathcal{A} \subseteq \{2, \ldots, q\}$, let $\boldsymbol{x}_{\mathcal{A}}(\boldsymbol{s})$ and $\boldsymbol{x}_{\mathcal{A}^c}(\boldsymbol{s})$ denote the vectors of covariates corresponding to $\mathcal{A}$ and its complement respectively.

THEOREM 2: *Suppose there exists a subset of covariates $\mathcal{A} \subseteq \{2, \ldots, q\}$ with $\mathcal{A} \neq \emptyset$, such* *that for all $j = 1, \ldots, p$ we can write $\boldsymbol{\beta}_j^0 = (\boldsymbol{\beta}_{j,\mathcal{A}}^0 = \mathbf{0}, \boldsymbol{\beta}_{j,\mathcal{A}^c}^0)^\top$. Assuming $\boldsymbol{x}_{\mathcal{A}}(\boldsymbol{s})$ and $\boldsymbol{x}_{\mathcal{A}^c}(\boldsymbol{s})$ are independent, then the corresponding subset of $\boldsymbol{\beta}_j^*$ will also equal zero i.e., $\boldsymbol{\beta}_{j,\mathcal{A}}^* = \mathbf{0}$. Thus under Assumption 1, the maximum likelihood estimator of the misspecified GLLVM satisfies $\hat{\boldsymbol{\beta}}_{j,\mathcal{A}} \xrightarrow{p} \mathbf{0}$ for all $j = 1, \ldots, p$, as $n \to \infty$, and irrespective of the values of $\boldsymbol{\lambda}^0$.*

The above result ensures that for any covariates which are truly uninformative for all species, the corresponding estimates of the misspecified GLLVM will consistently estimate these zero coefficients. While weaker than the full consistency result for conditionally normal responses, Theorem 2 may serve as a useful basis for studying variable selection in spatial GLLVMs e.g., establishing selection consistency of information criteria or penalized likelihood methods for choosing covariates driving the entire species community. Note that for the above result to hold, we require the strong assumption of independence between the truly informative and truly uninformative covariates. Such an assumption has been made previously in order to establish consistency when studying misspecification in generalized linear mixed models (e.g., Litire et al., 2007), and is similar to the partial orthogonality condition imposed in high-dimensional variable selection (e.g., Huang et al., 2008). In Section 4, we shall empirically assess the generality of Theorem 2 in settings with correlated covariates.

It is important to note that Theorem 2 demonstrates zero consistency only for completely uninformative covariates: for a partly informative covariate that is important for some but not all species, zero consistency cannot be guaranteed for species whose responses are not related to this covariate. Intuitively, this is because there remains an indirect dependence on this covariate through the combination of the direct dependence on the covariate for some species and the correlation between species induced by the latent variables. In Section 4, we shall empirically study the effects of misspecification on estimation and inference for partly informative covariates.

3.2 *Loadings*

Since the loadings directly relate to the correlation structures in the GLLVM, then one would expect that misspecifying the correlation structure in the latent variables should adversely affect their estimation and inference. This is largely the case, although a zero consistency result can be achieved.

THEOREM 3: *Suppose there exists a subset* $\mathcal{B} \subseteq \{1, \ldots, p\}$ *with* $\mathcal{B} \neq \emptyset$*, such that* $\boldsymbol{\lambda}_j^0 = \mathbf{0}$ *for all* $j \in \mathcal{B}$*. Then the corresponding subset of* $\boldsymbol{\lambda}_j^*$ *will also equal zero i.e.,* $\boldsymbol{\lambda}_j^* = \mathbf{0}$*. Thus under Assumption 1, the maximum likelihood estimator of the misspecified GLLVM satisfies* $\hat{\boldsymbol{\lambda}}_j \xrightarrow{p} \mathbf{0}$ *for all* $j \in \mathcal{B}$*, as* $n \to \infty$*.*

While Theorem 3 may not be important in community ecology, since we do not expect many (if any at all) species to be completely uncorrelated to other species, it may still play a useful role in other applications of GLLVMs where such independence can arise (e.g., Hirose and Konishi, 2012). Intuitively, any species that is uncorrelated with all others in the model remains uncorrelated under misspecification of the latent variable correlation structure. Therefore, we cannot borrow strength from other species to improve estimation and inference of the parameters specific to this species. This leads to the following result.

COROLLARY 1: *For any species* $j \in \mathcal{B}$ *that is independent of all other species on the linear predictor scale of the model, as defined in Theorem 3, the regression coefficients satisfy* $\boldsymbol{\beta}_j^* = \boldsymbol{\beta}_j^0$*. Thus under Assumption 1, the maximum likelihood estimator of the misspecified GLLVM satisfies* $\hat{\boldsymbol{\beta}}_j \xrightarrow{p} \boldsymbol{\beta}^0$ *for all* $j \in \mathcal{B}$*, as* $n \to \infty$*.*

4. Simulation Study

We conducted a simulation study to assess the finite sample performance of misspecified GLLVMs, relative to GLLVMs where the true spatial correlation structure for the latent variables was known. Specifically, we simulated data from a spatial GLLVM with $p = 20$ species, $q = 7$ covariates including the intercept, and $d = 3$ latent variables. The full details of the simulation design are provided in Appendix C, and we only provide some of the defining features below. Specifically, the n sites were arranged in a square lattice to reflect an expanding domain framework. Excluding the intercept, the covariates $\boldsymbol{x}(\boldsymbol{s}_i)$ included three spatially structured covariates and three spatially independent (but correlated to each other) covariates. The covariates were also generated in a way such that the conditions required for Theorem 2 are not satisfied, thereby allowing us to assess performance under a more realistic setting. The regression coefficients $\boldsymbol{\beta}_j$ varied both in magnitude and sign, and were constructed so that two covariates were informative for all species, two were informative for half of the species, and two were uninformative for all species. Similarly, the loadings $\boldsymbol{\lambda}_j$ in the spatial GLLVM were constructed such that species 16 to 20 were independent of all other species, while the remaining fifteen species had loadings varying in magnitude and size.

We generated multivariate spatial data with four possible response types i.e., four possible choices for the conditional distribution $f\{y_j(\boldsymbol{s}_i)|\boldsymbol{x}(\boldsymbol{s}_i), \boldsymbol{u}(\boldsymbol{s}_i)\}$: 1) continuous responses from the normal distribution; 2) overdispersed counts from the negative binomial distribution; 3) non-negative continuous responses from the Tweedie distribution; 4) presence-absence responses from the Bernoulli distribution. We considered grids of dimension $n^{1/2} = 7, 10, 14, 22, 32$, and simulated 400 datasets for each value of n. The grid sizes were chosen such that the total number of sites approximately doubled with each grid.

4.1 *Model Fitting and Performance Assessment*

We fitted two models to each simulated dataset: a spatial GLLVM with marginal likelihood given by (1) and assuming the true spatial correlation structure for the latent variables is known, and a misspecified independent GLLVM with marginal likelihood given by (2). In addition to point estimates, we calculated 95% Wald confidence intervals for the regression coefficients and loadings using both models. For the true model, this was based on standard errors obtained from the inverse of the observed information matrix, $\hat{\boldsymbol{I}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}}) = -\nabla^2 \log L(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}})$ where $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}})^\top$ generically denotes the corresponding maximum likelihood estimates. For the misspecified model, we calculated standard errors based on the inverse of the sandwich information matrix. Let $L_{\text{Ind},i}(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\lambda}) = \int \prod_{j=1}^{p} f\{y_j(\boldsymbol{s}_i)|\boldsymbol{x}(\boldsymbol{s}_i), \boldsymbol{u}(\boldsymbol{s}_i)\} h\{\boldsymbol{u}(\boldsymbol{s}_i)\} d\boldsymbol{u}(\boldsymbol{s}_i)$ denote the likelihood for the misspecified GLLVM for site i. Then we define $\hat{\boldsymbol{I}}_{\text{sand}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\lambda}}) = \hat{\boldsymbol{H}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\lambda}}) \hat{\boldsymbol{J}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\lambda}})^{-1} \hat{\boldsymbol{H}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\lambda}})$ where

$$
\begin{aligned}
\hat{\boldsymbol{H}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\lambda}}) &= -\frac{1}{n} \sum_{i=1}^{n} \nabla^2 \log L_{\text{Ind},i}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\lambda}}) \\
\hat{\boldsymbol{J}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\lambda}}) &= \frac{1}{n} \sum_{i=1}^{n} \nabla \log L_{\text{Ind},i}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\lambda}}) \{\nabla \log L_{\text{Ind},i}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\lambda}})\}^\top.
\end{aligned}
$$

The sandwich information matrix is generally used as the basis for quantifying uncertainty for maximum likelihood estimates under misspecified models, and we refer the reader to (White, 1982; Verbeke and Lesaffre, 1997) on its use in misspecified models.

All GLLVMs were estimated using Template Model Builder (TMB, Kristensen et al., 2015), which uses a combination of automatic differentiation and the Laplace approximation to produce functions for efficiently calculating the gradient and Hessian of the (Laplace approximated) marginal log-likelihood. For the spatial GLLVM, we made use of `R-INLA` (Lindgren and Rue, 2015) to construct a stochastic partial differential equation approximation to the distribution of the spatial latent variables, which ensured the computation remained feasible for large n (see also Thorson et al., 2015). To assess performance of the regression coefficients, we calculated the empirical bias and root mean squared error (RMSE) averaged across the simulated datasets, and the empirical coverage probability and mean interval width of the 95% confidence intervals. To assess performance of the loadings and latent variables predictions, we calculated the Procrustes errors between the estimated and true loading matrix, and between the estimated and true latent variable matrix (which is formed by stacking the n sets of vectors $\boldsymbol{u}(\boldsymbol{s}_i)$ as rows). Finally, we recorded the computation time (in minutes) used to fit the GLLVM and calculate the associated 95% confidence intervals. We present these results in Appendix C since, as to be expected, the simpler misspecified independent GLLVM was always faster and scaled better with increasing sample size across all response types.

4.2 *Setting 1: Normal Responses*

With regards to the regression coefficients, the biggest factor determining performance for misspecified GLLVMs was whether the corresponding covariate was spatially correlated or not (Figure 1). Thus we shall discuss our results based on these two cases separately.

For spatially uncorrelated covariates, both the bias and RMSE of the regression coefficients tended to zero for misspecified GLLVMs as sample size increased (Figure 1 left and middle columns, solid circles). However the RMSE for misspecified GLLVMs was consistently higher than for true GLLVMs, suggesting more variability in the estimates under misspecification. There was also little difference between the true and misspecified GLLVMs in terms of coverage probability for coefficients corresponding to spatially uncorrelated covariates, with the intervals for both models tending to the nominal level irrespective of the size of the regression coefficient and the norm of the corresponding loadings (Figure 1 right column). The sandwich based confidence intervals from the misspecified GLLVM tended to be wider than the intervals based on the true GLLVM, especially for small to medium sized regression coefficients (see Appendix C).

It is startling to see how much the performance deteriorated for coefficients corresponding to spatially correlated covariates under misspecification: the empirical bias and RMSE were substantially larger compared to both estimates from the true GLLVM, as well as estimates corresponding to spatially uncorrelated covariates (Figure 1 left and middle columns, empty circles). The sandwich based confidence interval suffered severe undercoverage, with performance becoming worse with increasing sample size. This poor performance occurred irrespective of the true magnitude of the regression coefficient, but did depend on the norm of the corresponding loadings.

The difference in performance between spatially correlated versus uncorrelated covariates, under misspecification, can be attributed to the former exhibiting a strong correlation to the spatially structured latent variables. Specifically, a non-negligible proportion of the variation in each spatially correlated covariate can be explained by a linear combination of the eigenvectors of the spatial covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. Under misspecification, because the latent variables are assumed to be independent, the independent GLLVM then erroneously attributes part of this spatial correlation in the latent variables to the covariates. This causes estimation and inference for the corresponding elements of $\boldsymbol{\beta}_j$ to deteriorate as the estimated coefficients become affected by the values of $\boldsymbol{\lambda}_j$. By contrast, spatially uncorrelated covariates (in general) exhibit little correlation to spatially structured latent variables, and so the erroneous attribution does not occur and estimation and inference for the corresponding coefficients are largely unaffected. This difference in performance has close connections to the issue of spatial confounding, and we expand upon this issue in Section 6 and in Appendix F.

[Figure 1 about here.]

Neither estimation of the loadings nor prediction of the latent variables was robust to misspecification of the correlation structure, with the Procrustes errors for the misspecified GLLVM being consistently higher than those from the true GLLVM (Figure 2 top row). Additional analyses (not shown) of the results for loadings reveal that, consistent with the zero consistency result of Theorem 3, it was the non-zero loadings (from species 1 to 15) that were poorly estimated under misspecification, while the truly zero loadings (from species 16 to 20) were estimated well under misspecification of the latent variable correlation structure.

Of course, given the above results a natural question to ask is whether performance of the misspecified independent GLLVM would improve if we were to fit a model with a larger number of latent variables. In Appendix C, we perform such additional simulations where we fit independent GLLVMs with $d = 5, 7, 9$, noting that true number of latent variables in the simulation is $d = 3$. In summary, results for normal responses show that estimation and inference on the regression coefficients is hardly affected by increasing the number of latent variables, while prediction performance actually deteriorates (as quantified by the Frobenius norm between the matrix of estimated and true linear predictor component $\boldsymbol{u}(\boldsymbol{s}_i)^\top \boldsymbol{\lambda}_j$).

[Figure 2 about here.]

4.3 *Setting 2: Negative Binomial Counts*

The overall patterns seen for the normal response case largely carry over to this setting. For spatially uncorrelated covariates, the empirical bias and RMSE of the regression coefficients tended to zero for misspecified GLLVMs (Figure 3 left and middle columns, solid circles). This finding expands the zero consistency result of Theorem 2 to the case of covariates that are correlated with each other, and to truly zero coefficients coming from partly informative as well as completely uninformative covariates. There was also little difference between the true and misspecified GLLVMs in terms of the coverage probability of the confidence intervals for coefficients corresponding to spatially uncorrelated covariates, with intervals for both models tending to the nominal level irrespective of the size of the coefficient and the corresponding norm of the loadings (Figure 3 right column). By contrast, for spatially correlated covariates, estimation and inference of regression coefficients suffered severely under misspecification, with considerably higher RMSE and major undercoverage especially for coefficients that also had a large corresponding norm of the loadings. Again, the poor performance can be explained by the spatially correlated covariates exhibiting a strong correlation to the spatially structured latent variables, and so the misspecified GLLVM erroneously attributes part of the latent variable component of the model to the covariates, leading to poor estimation and inference on the corresponding coefficients.

[Figure 3 about here.]

Estimation of the loadings and prediction of the latent variables both suffered under misspecification of the correlation structure for negative binomial GLLVMs (Figure 2 bottom row), with additional analyses again revealing that it was the non-zero loadings that were particularly poorly estimated under misspecification, while the truly zero loadings were estimated relatively well.

Simulation results for the Tweedie response GLLVM presented similar trends to those for the negative binomial counts and normal responses seen above, and the results for these are presented in Appendix C. In particular, estimation and inference on the regression coefficients was largely robust to misspecification for spatially uncorrelated variables, but suffered severely misspecification for spatially correlated variables. Estimation of the loadings and prediction of the latent variables performed consistently poorly under misspecification of the latent variable correlation structure.

4.4 *Setting 3: Binary Responses*

In contrast to previous settings where estimation and inference of coefficients corresponding to spatially uncorrelated covariates was largely robust to misspecification, for binary responses we observe that the degree of robustness also depended heavily on whether the true value of the coefficient was itself close to or exactly equal to zero. In the left and middle columns of Figure 4, we observe eight notably outlying regression coefficients (four corresponding to spatially uncorrelated covariates and four corresponding to spatially correlated covariates) that were both large in their true magnitude and had large corresponding norm of the loadings, which performed especially poorly under misspecification. These eight coefficients also had coverage probabilities for the corresponding sandwich based confidence intervals that were notably below the nominal level (Figure 4 right column). This is despite the fact that the sandwich based confidence intervals tended to be relatively wide in this setting overall (see Appendix C).

[Figure 4 about here.]

Comparing against results from previous simulating settings, this suggests that for binary responses, misspecification of the latent variable correlation structure can have a major impact on inference for non-zero $\boldsymbol{\beta}_j$'s (even for spatially uncorrelated covariates), *if* the corresponding norm of the loadings is also relatively large. For coefficients corresponding to spatially correlated covariates, misspecification again led to poor performance overall (Figure 4 empty circles).

The results for estimation of the loadings and prediction of the latent variables were similar to those seen with the other response types in Sections 4.2 and 4.3, and so for brevity we present them in the Appendix C.

5. Application to SO-CPR Survey

We fitted negative binomial GLLVMs to data collected in season 2007–2008 of the SO-CPR survey, with the aim being to quantify the relationship between the species responses and select environmental covariates. The data consisted of $n = 3,875$ sampling locations irregularly spaced across the Southern Ocean. Furthermore, we focused on $p = 24$ species detected at more than 5% of the sites (see Appendix D for a figure of the sampling locations and the species included in the analyses). We included two environmental covariates that were *a priori* considered to be important in influencing one or more of the species distributions: salinity (which is highly correlated with temperature) and photosynthetically active radiation (PAR). The latter is a measure of amount of light available for photosynthesis. Using orthogonal polynomials, we entered both covariates into the GLLVMs as linear and quadratic terms, and also considered a pairwise interaction term between them. Along with an intercept term, this produced a total of $q = 6$ covariates in $\boldsymbol{x}(\boldsymbol{s}_i)$ and $pq = 144$ coefficients in $\boldsymbol{\beta}$. We fitted GLLVMs with $d = 3$ latent variables and considered two possible spatial correlation structures: 1) a Matérn covariance function with $\nu = 1$; 2) an independence correlation structure.

The results are presented in Appendix D, from which we observe that many of the estimated coefficients from the two fitted models were similar in magnitude and sign, revealing that salinity and PAR, along with the quadratic and pairwise interactions terms, were important factors in driving many of the species responses. There were however some notable differences: focusing on whether the 95% confidence intervals contained zero or not, each covariate had several species that produced differing conclusions across the two models (5 for the linear effect of salinity, 3 for the quadratic effect of salinity, 3 for the linear effect of PAR, 6 for the quadratic effect of PAR, and 5 for the interaction between salinity and PAR). The majority of species had a negative coefficient of the quadratic effect of salinity, thus supporting the idea of a "niche" in this environment space and consequently of particular water masses between Tasmania and Antarctica. By contrast, many species had significant negative linear effects and significant positive quadratic effects of PAR, suggesting almost all marine species for analysis tended to prefer low light environments. Regarding computation time, the independent GLLVM took 3.70 hours to complete (on an Intel Xeon E5-2680 V3 at 2.50 GHz with 5 CPUs), while the spatial GLLVM took 15.5 hours to fit.

6. Discussion

We explored the consequences of misspecifying the spatial correlation structure of the latent variables in GLLVMs. If the aim of the analysis centers on estimation and inference for the regression coefficients, our findings show that by far the strongest determinant of performance is whether the covariate is spatially correlated or not. For spatially uncorrelated covariates, estimation and inference for the coefficients is largely robust to misspecification, with the exception of binary responses. But for spatially correlated covariates, which arise quite frequently in observational studies in ecology, assuming independence of the latent variables leads to poor performance for the corresponding regression coefficients. It is also important to account for potential spatial correlation if the aim of the analysis centers on estimation and inference for loadings, and for prediction from the GLLVM as a whole.

In our simulation study, we only generated covariates that were either spatially independent or covariates with a fixed spatial scale. This raises the question of what happens if we were to consider covariates with varying degrees of spatial correlation. To answer this question, we conducted a new set of simulations where multivariate spatial data were generated from a spatial GLLVM where the covariates possessed varying strengths of spatial strengths. The full details are presented in Appendix E, but in brief, results exhibited a general and consistent trend across all four response types, such that estimation and inference on the regression under the misspecified independent GLLVM gradually worsened as the corresponding covariate became more strongly spatially correlated. By contrast, estimation and inference from fitting the true spatial GLLVM were largely unaffected by the strength of spatial correlation for the covariates. These results provide further evidence that, if there is *a priori* information covariates are spatially correlated, then it is imperative to account for the spatial correlation when specifying the latent variable structure.

Comparing across empirical and theoretical results, there may appear to be some disagreement in terms of point estimation of the regression coefficients: Theorems 1 and 2 state that consistency is achievable under misspecification for all or some of the coefficients respectively, yet our simulations suggest issues with consistency for coefficients corresponding to spatially correlated covariates. While not definitive, and we encourage future research into the large sample properties of maximum likelihood estimation for misspecified spatial GLLVMs, this discordance may be because spatially correlated covariates possibly do not entirely satisfy the conditions underlying Assumption 1. As discussed in Section 4.2, there may be potential issues surrounding the identifiability of coefficients for spatially correlated covariates when a misspecified independent GLLVM is fitted, and the problems bears similarity to the issue of spatial confounding. We expand upon this connection in Appendix F, noting that while spatial confounding has been investigated in some detail for univariate spatial regression, there is very little research on its impact for spatial GLLVMs (see Shirota et al., 2019, for a recent exception that presented an interesting discussion on this topic), let alone what it means to "correct" for spatial confounding in misspecified independent GLLVM where the latent variables are, by definition, spatially independent.

While the above results have focused on the most extreme type of misspecification in the latent variable correlation structure (by assuming independence), a natural avenue of research is to examine cases of "slight misspecification" e.g., we include spatially correlated latent variables but the form of the spatial covariance is misspecified, or the true spatial correlation structures differ across the latent variables and but we misspecify and assume the same structure. Of course, we must be mindful that there is a limitless number of ways one can "mix and match" different types of spatially structured latent variables (e.g., we could have latent variables included at different scales to reflect a nested design, similar to Ovaskainen et al., 2016, and misspecify the correlation structure at one scale but not another), let alone the issue of what happens if we misspecify any combination or set of combinations of these. In Appendix G we present results from an additional simulation study where we generate data from a spatial GLLVM, and compare the performance of an independent GLLVM versus a "slightly misspecified" spatial GLLVM where the wrong smoothness parameter ν in the Matérn correlation function is assumed. Overall, and not surprisingly, the performance of the slightly misspecified spatial GLLVM is generally better than the independent GLLVM.

However there is effectively zero computational gain from fitting this slightly misspecified spatial GLLVM, and so in some sense is against the motivation and spirit of this article i.e., our aim is to examine the consequences on estimation and inference when we are far away from the true spatial structure, where this is traded off against substantial reductions in computation time.

Finally, in our simulations we assumed that the true number of latent variables is known, when in practice this also has been chosen (e.g., Hui et al., 2018). How misspecifying the latent variable correlation structures affects the number of latent variables chosen is an avenue of future research to pursue, along with the more general topic of how misspecification of the latent variable correlation structure affects variable selection as well as other aspects of inference for GLLVMs as a whole (see Hoeting et al., 2006; Xu et al., 2015, for examples of related research on variable selection in geostatistical models when the spatial correlation is misspecified).

Acknowledgements

Both FKCH and AHW were supported by Australia Research Council Disovery Grants. Thanks to Noel Cressie, Anders Nielsen, and Graham Hosie for useful discussions.

Data Availability Statement

The data that support the findings in this paper are openly available as part of the "Australian Antarctic Data Centre, at `http://dx.doi.org/doi:10.26179/5ee84f77cc4ec`, cited as Hosie (2020).

References

Bjork, J. R., Hui, F. K. C., O'Hara, R. B., and Montoya, J. M. (2018). Uncovering the drivers of host-associated microbiota with joint species distribution modeling. *Molecular* *ecology* **27,** 2714–2724.

Hirose, K. and Konishi, S. (2012). Variable selection via the weighted group lasso for factor analysis models. *Canadian Journal of Statistics* **40,** 345–361.

Hoeting, J. A., Davis, R. A., Merton, A. A., and Thompson, S. E. (2006). Model selection for geostatistical models. *Ecological Applications* **16,** 87–98.

Hosie, G. (2020). Southern Ocean Continuous Plankton Recorder Zooplankton Records. *Australian Antarctic Data Centre.* Version 8. 10.26179/5ee84f77cc4ec .

Huang, J., Horowitz, J. L., and Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *The Annals of Statistics* **36,** 587–613.

Hui, F. K. C., Mueller, S., and Welsh, A. H. (2020). Random effects misspecification can have severe consequences for random effects inference in linear mixed models. *International Statistical Review* **In Press,**.

Hui, F. K. C., Tanaka, E., and Warton, D. I. (2018). Order selection and sparsity in latent variable models via the ordered factor LASSO. *Biometrics* **74,** 1311–1319.

Irincheeva, I., Cantoni, E., and Genton, M. G. (2012). Generalized linear latent variable models with flexible distribution of latent variables. *Scandinavian Journal of Statistics* **39,** 663–680.

Kristensen, K., Nielsen, A., Berg, C., Skaug, H., and Bell, B. (2015). TMB: Automatic Differentiation and Laplace Approximation. *Journal of Statistical Software* **70,** 1–21.

Kurisu, D. (2019). On nonparametric inference for spatial regression models under domain expanding and infill asymptotics. *Statistics & Probability Letters* **154,** 108543.

Lindgren, F. and Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software* **63,** 1–25.

Litire, S., Alonso, A., and Molenberghs, G. (2007). Type I and Type II Error Under Random-Effects Misspecification in Generalized Linear Mixed Models. *Biometrics* **63,** 1038–1044.

Lopes, H. F., Gamerman, D., and Salazar, E. (2011). Generalized spatial dynamic factor models. *Computational Statistics & Data Analysis* **55,** 1319–1330.

Lu, Z. and Tjostheim, D. (2014). Nonparametric estimation of probability density functions for irregularly observed spatial data. *Journal of the American Statistical Association* **109,** 1546–1564.

Ma, Y. and Genton, M. G. (2010). Explicit estimating equations for semiparametric generalized linear latent variable models. *Journal of the Royal Statistical Society Series B* **72,** 475–495.

Mardia, K. V. and Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71,** 135–146.

Niku, J., Brooks, W., Herliansyah, R., Hui, F. K. C., Taskinen, S., and Warton, D. I. (2019). Efficient estimation of generalized linear latent variable models. *PloS one* **14,** e0216129.

Ovaskainen, O., Roy, D. B., Fox, R., and Anderson, B. J. (2016). Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. *Methods in Ecology and Evolution* **7,** 428–436.

Shirota, S., Gelfand, A. E., and Banerjee, S. (2019). Spatial joint species distribution modeling using Dirichlet processes. *Statistica Sinica* **29,** 1127–1154.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models.* CRC Press.

Thorson, J. T., Scheuerell, M. D., Shelton, A. O., See, K. E., Skaug, H. J., and Kristensen, K. (2015). Spatial factor analysis: a new tool for estimating joint species distributions and correlations in species range. *Methods in Ecology and Evolution* **6,** 627–637.

Verbeke, G. and Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis* **23,** 541–556.

Wang, H. and Zhu, J. (2009). Variable selection in spatial regression via penalized least squares. *Canadian Journal of Statistics* **37,** 607–624.

Warton, D. I., Blanchet, F. G., OHara, R., Ovaskainen, O., Taskinen, S., Walker, S. C., and Hui, F. K. C. (2016). Extending Joint Models in Community Ecology: A Response to Beissinger et al. *Trends in Ecology & Evolution* **31,** 737–738.

White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica* **50,** 1–26.

Xu, L., Wang, Y.-G., Zheng, S., and Shi, N.-Z. (2015). Model selection with misspecified spatial covariance structure. *Journal of Statistical Computation and Simulation* **85,** 2276–2294.

Yoon, H.-J. and Welsh, A. H. (2020). On the effect of ignoring correlation in the covariates when fitting linear mixed models. *Journal of Statistical Planning and Inference* **204,** 18–34.

Zhu, J., Eickhoff, J., and Yan, P. (2005). Generalized linear latent variable models for repeated measures of spatially correlated multivariate data. *Biometrics* **61,** 674–683.

Supporting Information

Web Appendices, figures, and additional details and discussion referenced in Sections 3 to 5 are available with this paper at the Biometrics website on Wiley Online Library. Additionally, template `R` scripts for estimating GLLVMs and for performing all simulations are provided.

Figure 1. Simulation results for empirical bias (left column), root mean squared error (middle column), and coverage probability (right column) of regression coefficients $\boldsymbol{\beta}_j$, for normal response GLLVMs. Points are differentiated by shape ('•' for misspecified GLLVM coefficients corresponding to spatially uncorrelated covariates; '∘' for misspecified GLLVM coefficients corresponding to spatially correlated covariates; '△' for true GLLVM coefficients) and shading (darker shades are coefficients with larger corresponding norms of the loadings).

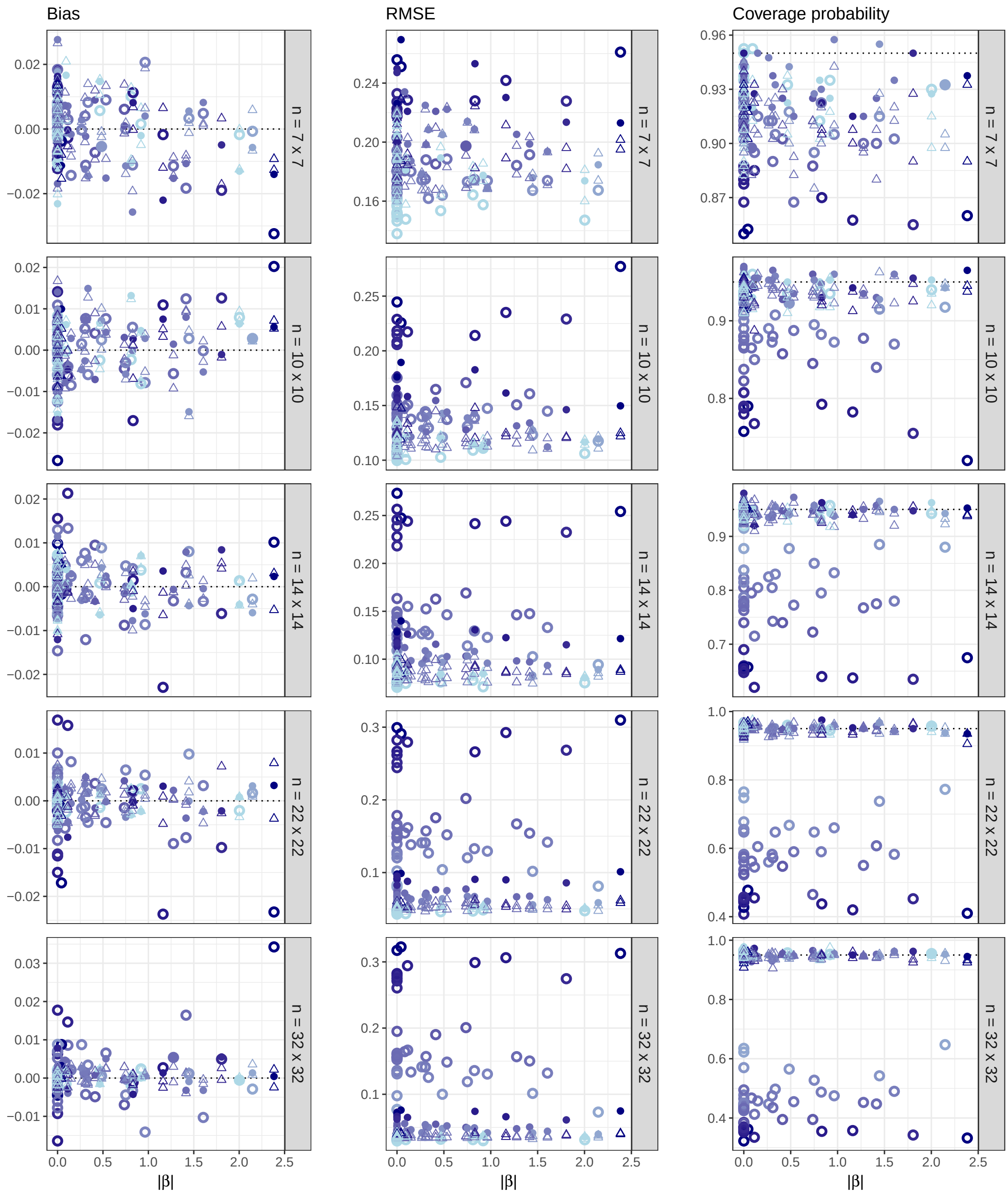

Figure 2. Simulation results showing comparative boxplots (left and red for misspecified GLLVM, right and blue for true GLLVM) for the Procrustes error of the estimated loadings $\mathbf{\Lambda}$(left column) and Procrustes error of the predicted latent variables $\boldsymbol{u}(\boldsymbol{s}_i)$ (middle right column), for GLLVMs with normal response (top row) and negative binomial response (botom row).

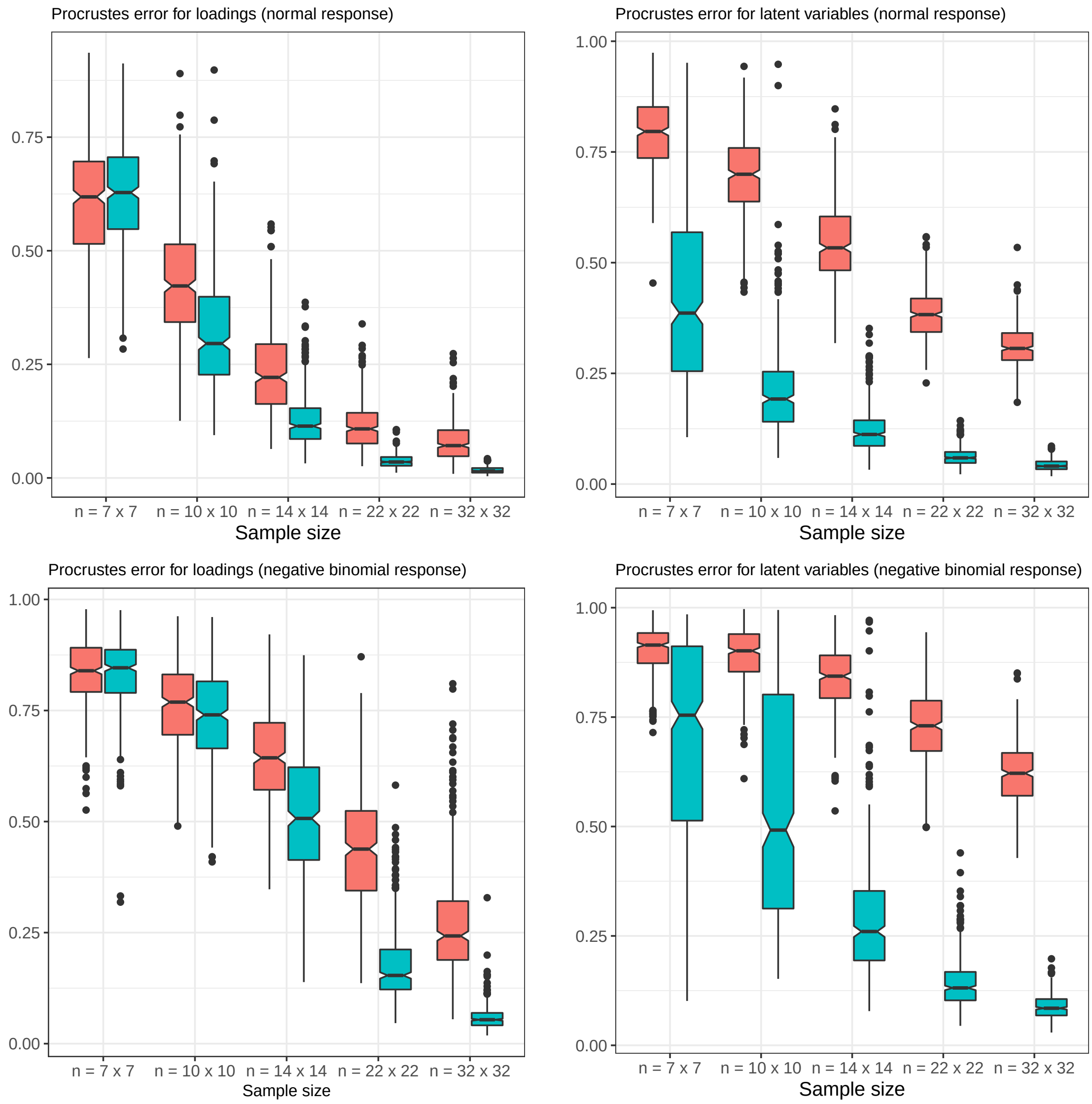

Figure 3. Simulation results for empirical bias (left column), root mean squared error (middle column), and coverage probability (right column) of regression coefficients $\boldsymbol{\beta}_j$, for negative binomial GLLVMs. Points are differentiated by shape ('•' for misspecified GLLVM coefficients corresponding to spatially uncorrelated covariates; '∘' for misspecified GLLVM coefficients corresponding to spatially correlated covariates; '△' for true GLLVM coefficients) and shading (darker shades are coefficients with larger corresponding norms of the loadings).

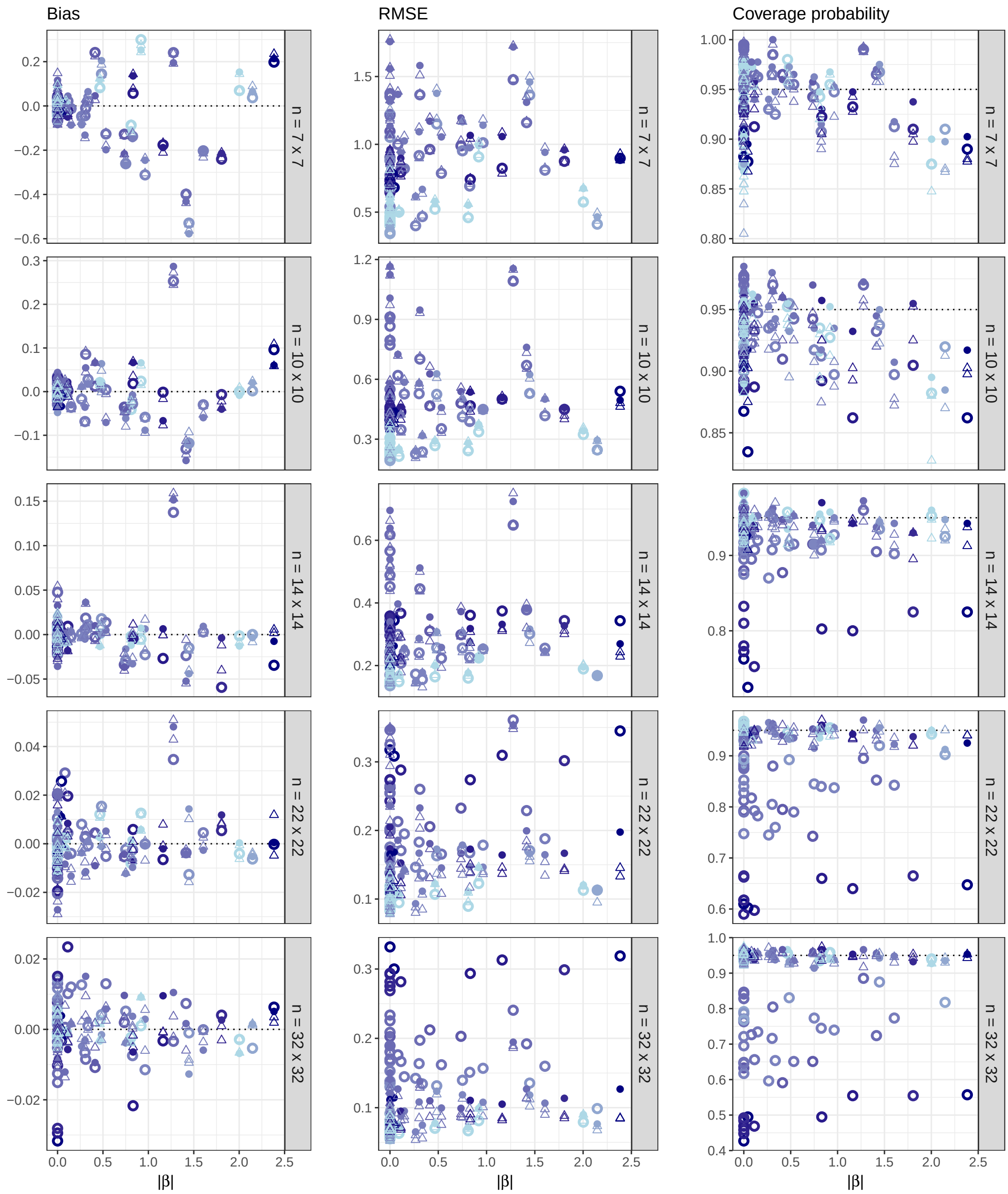

Figure 4. Simulation results for empirical bias (left column), root mean squared error (middle column), and coverage probability (right column) of regression coefficients $\boldsymbol{\beta}_j$, for Setting 3 with binary GLLVMs. Points are differentiated by shape ('•' for misspecified GLLVM coefficients corresponding to spatially uncorrelated covariates; '∘' for misspecified GLLVM coefficients corresponding to spatially correlated covariates; '△' for true GLLVM coefficients) and shading (darker shades are coefficients with larger corresponding norms of the loadings).