Supplementary materials to "A Computationally Efficient Projection-Based Approach for Spatial Generalized Linear Mixed Models" by Guan and Haran

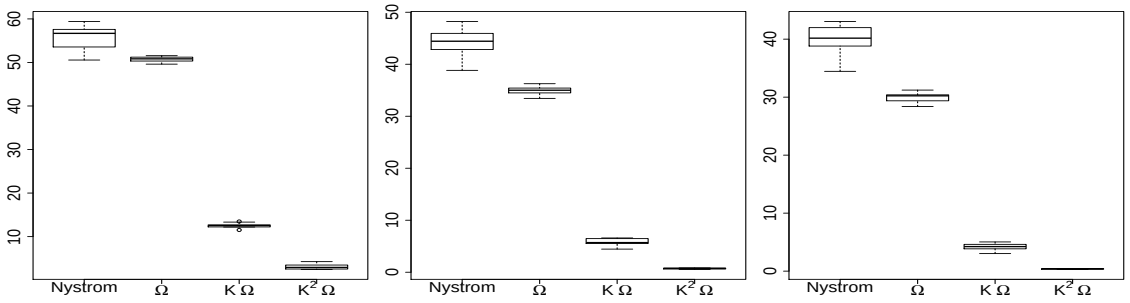
S.1 Eigencomponent Approximation Performance

Here we compare eigencomponent approximation performance for increasing smoothness $\nu = 0.5, 1.5, 2.5$ and increasing spatial dependence with effective range $r = 0.1, 0.3, 0.5, 0.7$. Figure 1 shows the distance between the subspaces generated by the first 100 approximated and true eigenvectors. Figure 2 shows the L^2 distance between the first 100 approximated and true eigenvalues. Our conclusion here is the same as in the manuscript. Introducing random matrix Ω improves the approximation. Taking Φ to be $K^{\alpha}\Omega$ further improves approximation, where in practice $\alpha = 1$ appears to be a good choice.

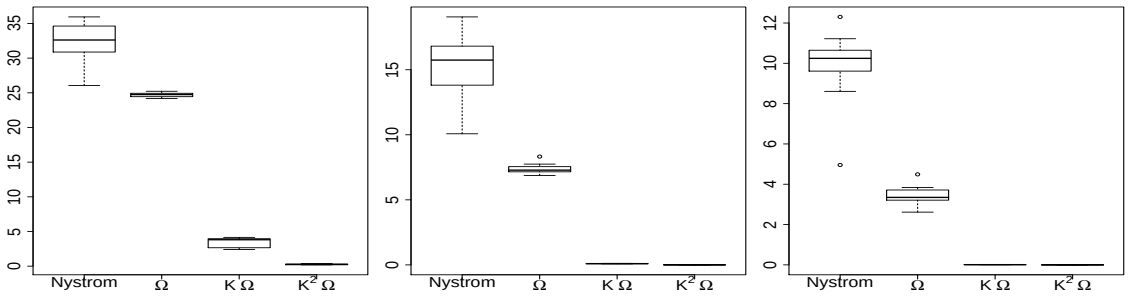
S.2 Full Conditionals for Projection-Based Approaches

The joint posterior distribution for the full model with random projection is $\pi(\boldsymbol{\delta}, \boldsymbol{\beta}, \sigma^2, \phi \mid \boldsymbol{Z}) \propto f(\boldsymbol{Z} \mid \boldsymbol{\delta}, \boldsymbol{\beta}, \sigma^2, \phi) \times f(\boldsymbol{\delta} \mid \sigma^2, \phi) \times \pi(\boldsymbol{\beta}) \times \pi(\sigma^2) \times \pi(\phi)$. From this we derive the full conditionals, shown below, which can be easily sampled using one-variable-at-a-time Metropolis-Hasting algorithm.

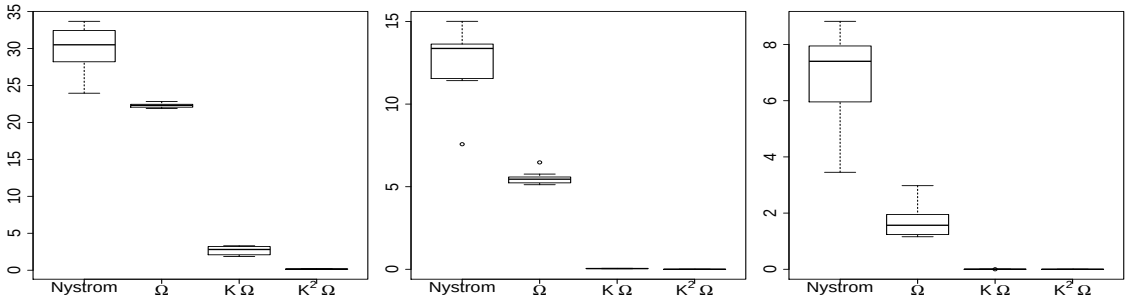
$$
\begin{aligned}
\boldsymbol{\beta} \mid \boldsymbol{\beta}_{-} &\propto \prod_{i=1}^{n} f\left(Z_i \mid \boldsymbol{\beta}, U_m, D_m, \boldsymbol{\delta}\right) \times \pi(\boldsymbol{\beta}), \\
\sigma^2 \mid \sigma^2_{-} &\propto \left(\sigma^2\right)^{-m/2} \exp\left(-\frac{1}{2\sigma^2}\boldsymbol{\delta}^T\boldsymbol{\delta}\right) \times \pi(\sigma^2), \\
\phi \mid \phi_{-} &\propto \prod_{i=1}^{n} f\left(Z_i \mid \boldsymbol{\beta}, U_m, D_m, \boldsymbol{\delta}\right) \times \exp\left(\frac{1}{2\sigma^2}\boldsymbol{\delta}^T\boldsymbol{\delta}\right) \times \pi(\phi), \\
\boldsymbol{\delta} \mid \boldsymbol{\delta}_{-} &\propto \prod_{i=1}^{n} f\left(Z_i \mid \boldsymbol{\beta}, U_m, D_m, \boldsymbol{\delta}\right) \times \exp\left(-\frac{1}{2\sigma^2}\boldsymbol{\delta}^T\boldsymbol{\delta}\right).
\end{aligned}
$$



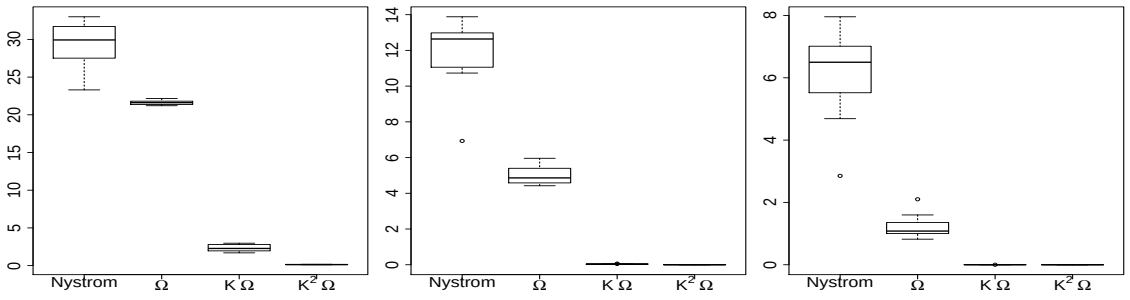
(a) ν=0.5, r=0.1 (b) ν=1.5, r=0.1 (c) ν=2.5, r=0.1



(d) ν=0.5, r=0.3 (e) ν=1.5, r=0.3 (f) ν=2.5, r=0.3

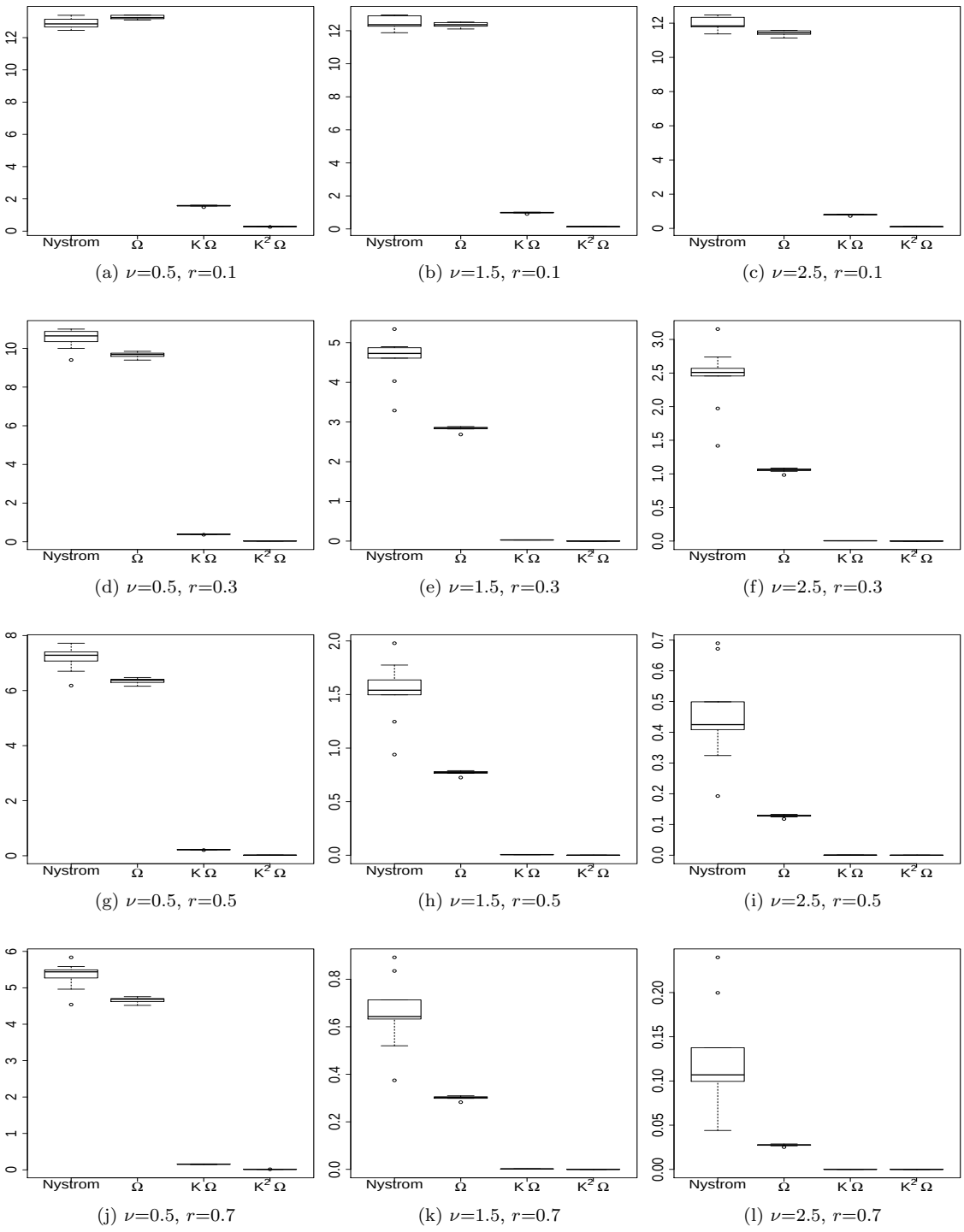


(g) ν=0.5, r=0.5 (h) ν=1.5, r=0.5 (i) ν=2.5, r=0.5



(j) ν=0.5, r=0.7 (k) ν=1.5, r=0.7 (l) ν=2.5, r=0.7

Figure 1: Distance between the subspaces generated by the fist 100 approximated and true eigenvectors under different smoothness ν and effective range r. Introducing randomness in $\Phi = K^{\alpha}\Omega$ improves the Nystöm approximation to the eigenvectors. Letting Φ to be $K^{\alpha}\Omega$ with small power $\alpha = 1, 2$ further improves approximation.



(a) ν=0.5, r=0.1 (b) ν=1.5, r=0.1 (c) ν=2.5, r=0.1

(d) ν=0.5, r=0.3 (e) ν=1.5, r=0.3 (f) ν=2.5, r=0.3

(g) ν=0.5, r=0.5 (h) ν=1.5, r=0.5 (i) ν=2.5, r=0.5

(j) ν=0.5, r=0.7 (k) ν=1.5, r=0.7 (l) ν=2.5, r=0.7

Figure 2: L^2 distance between the first 100 approximated and true eigenvalues under different smoothness ν and effective range r. Introducing randomness in $\Phi = K^{\alpha}\Omega$ improves the Nystöm approximation. Letting Φ to be $K^{\alpha}\Omega$ with small power $\alpha = 1, 2$ further improves approximation

The full conditionals for the restricted model with random projection is similar to the above except that U_m is replaced by $P^{\perp}_{[X]}U_m$.

S.3 Simulation Study Results

For the linear case, we simulate 100 data sets from the spatial linear mixed model (confounded simulation scheme) for data sizes of $n = 400$ and $n = 1000$. For the smaller data size, we fit both of our projection-based approaches, the spatial linear mixed model and restricted spatial regression model for overall comparisons. The distribution of $\boldsymbol{\beta}$ estimates all center around the true value and are comparable among all four models (Figure 3); inference and prediction provided by our projection-based approaches are similar to the original models they approximate (Table 1). For the larger data size we fit both FRP and RRP with rank $m = 50$, which is selected based on our heuristic described in the main text. Figure 4 shows the estimated random effects at the training locations and the predicted observations at the testing locations. We see that our projection-based approaches work well in recovering the spatial patterns.

Table 1: Model comparisons for linear case with $n = 400$.

	SLMM	FRP	RSR	RRP	A-RRP
β_1 (coverage)	1.01 (0.99)	0.98 (0.97)	1.00 (0.07)	1.00 (0.07)	1.00 (0.97)
β_1 mse	0.39	0.46	0.79	0.79	0.79
β_2 (coverage)	1.02 (0.95)	1.01 (0.95)	1.02 (0.03)	1.02 (0.03)	1.02 (0.94)
β_2 mse	0.60	0.59	1.06	1.06	1.06
ϕ	0.21	0.22	0.21	0.21	NA
ϕ mse	0.62	0.61	0.62	0.63	NA
σ^2	1.25	1.34	1.24	1.20	NA
σ^2 mse	1.32	1.54	1.26	1.18	NA
pmse	0.13	0.13	0.13	0.13	NA

For the Poisson case, we simulate 100 data sets from the spatial linear mixed model (confounded scheme) and restricted spatial regression model (orthogonal schemes) for data sizes of $n = 1000$. Under the confounded simulation scheme, FRP and RRP have similar distributions for point estimates (Figure 5); however, the credible interval(CI) of RRP is

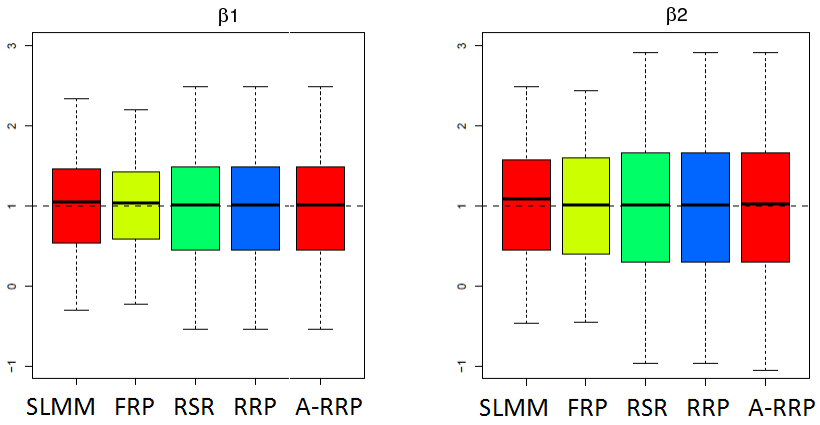


Figure 3: Distribution of posterior mean estimates of $\boldsymbol{\beta}$ among four models and with adjustments. The distributions all center around the true value and are comparable. Random projection models FRP and RRP with rank=50 produce results that are similar to the models they approximate.

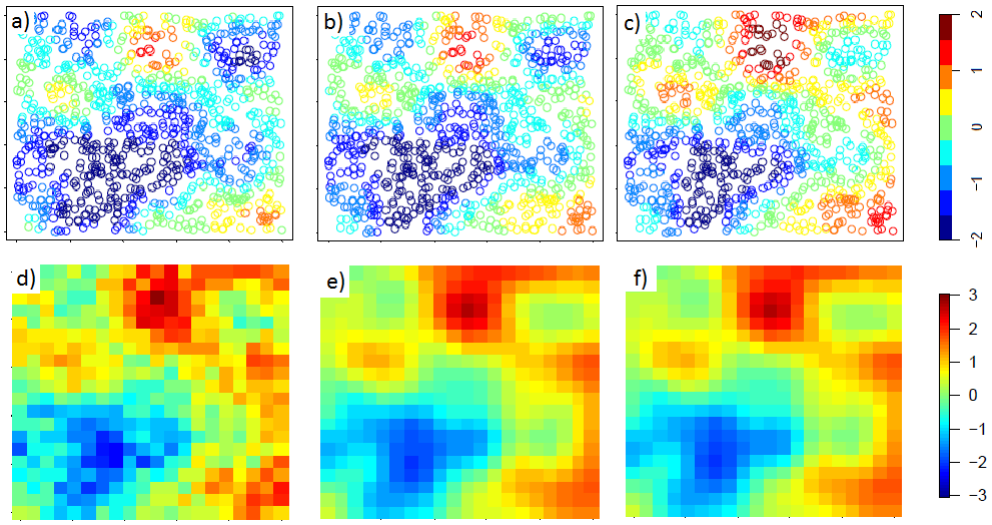


Figure 4: Linear case with $n = 1000$. First row shows the random effects estimate at training locations. Second row shows the prediction on a 20×20 grid using random projection models. Left column is simulated data, middle column shows the results from FRP, and right column shows the results from RRP. Random projections approach works well in recovering the true random effects.

inappropriately narrow with length 0.246(0.166, 0.367) and a coverage of 14 % compared to the FRP, the CI of which has length 3.123(1.664, 5.743) and a coverage of 91%. Under the orthogonal simulation scheme, RRP performs much better than FRP; its point estimates are closely centered around the true value (5)), its CI is 0.225(0.176, 0.299), much narrower compared to 2.985(1.666, 4.995) of the FRP, and both RRP and FRP have coverages that are comparable to the nominal rate. Under both simulation schemes, the adjusted inference A-RRP provides similar results to FRP. Hence, we can fit only the RRP model in practice for its computational efficiency, then apply the adjustment to recover inference results for the full model.

S.4 A Comparison with an Existing Method for Areal Data

Here we compare our approach with an existing method for lattice/areal data (Hughes and Haran, 2013). We simulate a count data set with $n = 900, \tau = 1$ from:

$$\begin{aligned} g\left\{E(\boldsymbol{Z} \mid \boldsymbol{\beta})\right\} &= \boldsymbol{x_1} + \boldsymbol{x_2} + \boldsymbol{W}, \\ p(\boldsymbol{W} \mid \tau) &\propto \tau^{rank(Q)/2} \exp\left(-\frac{\tau}{2}\boldsymbol{W}^T Q \boldsymbol{W}\right). \end{aligned} \tag{1}$$

The ICAR model has improper prior, meaning its precision matrix is rank deficient; therefore, direct simulation from (1) is not feasible. Hence, the spatial random effects is simulated using the eigencomponents of the precision matirx Q. Let (λ_i, e_i) denote the eigenpairs of Q, we simulate $\delta_i \sim N(0, \lambda_i^{(-1)})$ for $\lambda_i \neq 0$. Then $W = \sum_i \delta_i e_i$ has the desired distribution. To reduce the dimension of $\boldsymbol{W}$ using RRP, we will first invert Q using generalized inverse, then approximate Q^{-1} using Algorithm 1 from the main text. The full conditionals of RRP for this reparameterized model can be easily derived. We then fit both RRP and HH to the simulated data set for comparison. Figure 6 shows that the marginal posterior density plot are similar from the two models.

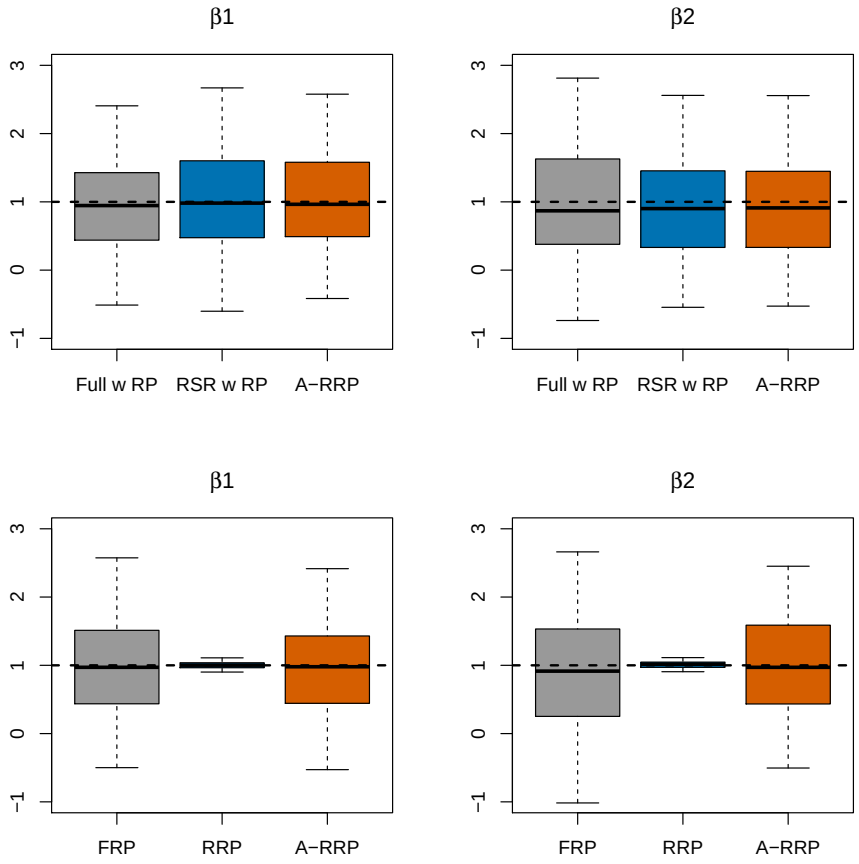


Figure 5: Poisson simulation study: distribution of $\boldsymbol{\beta}$ posterior mean estimates for RP models and after adjustment. First row for the confounded case, and second row for the orthogonal case. All distributions center around the true value. For the confounded case (top row), FRP and RRP have similar results; while under the orthogonal case (bottom row), RRP produce much tighter distribution.

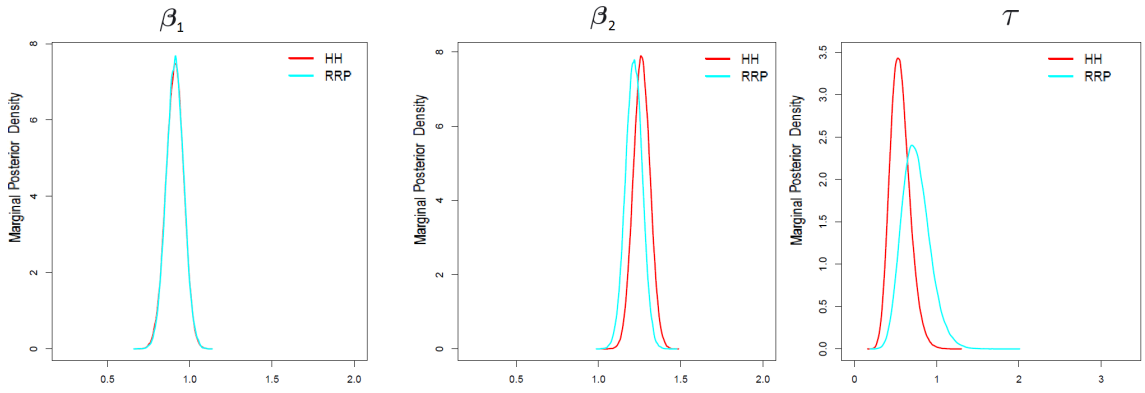


Figure 6: Marginal posterior plots for HH and RRP models. Results from the two models are comparable. RRP = restricted model with random projection

S.5 A Comparison with Predictive Process for Point-Referenced Data

To compare the performance of our projection-based approaches with the predictive process, we simulate 100 Poisson data sets from the traditional SGLMM. We fit both FRP and RRP with rank $m = 50$ to the datasets, and compare their results with the predictive process with reference points on a 7×7 grid. In this simulation study, our projection-based approaches provide comparable inference and smaller mean prediction square error (MPSE) (Figure 7).

S.6 SGLMMs with small-scale (nugget) spatial effect

For SGLMMs where inclusion of small scale, non-spatial heterogeneity is appropriate, the model becomes,

$$g\left\{E(Z(s) \mid \beta, W(s))\right\} = X(s)\beta + w(s) + \epsilon(s), \tag{2}$$

where $\epsilon(s) \overset{iid}{\sim} N(0, \tau^2)$. We provide implementations of our method for two cases: (1) when Gibbs sampling of the latent variables is available, and (2) when it is not. Examples for case (1) are the spatial binary model with probit link (considered by Berrett and Calder, 2016) and spatial probit model for correlated ordinal data (Schliep and Hoeting, 2015); examples for case (2) are already considered in this manuscript.

We begin by redefining some notation. Let $\boldsymbol{W} = (W_1, \ldots, W_n)^T$ denote the latent variable, $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^T$ the observed spatial binary data and X the $n \times p$ design matrix.

Case (1): We first consider the case where Gibbs sampling is available for the latent variables, for example when using SGLMM with a probit link for binary data. The model is defined as

$$Z_i = \begin{cases} 1, & Y_i \geq 0 \\ 0, & Y_i < 0 \end{cases} \tag{3}$$

where $Y_i = X_i\beta + W_i + \epsilon_i$. $\boldsymbol{W} \sim MVN(0, \sigma^2 R_\phi)$ captures large-scale spatial variation and $\epsilon_i \overset{i.i.d}{\sim} N(0, \tau^2)$ captures small-scale variation. The conditional distribution for $Y \mid \boldsymbol{\beta}, \sigma^2, \phi, \tau^2$

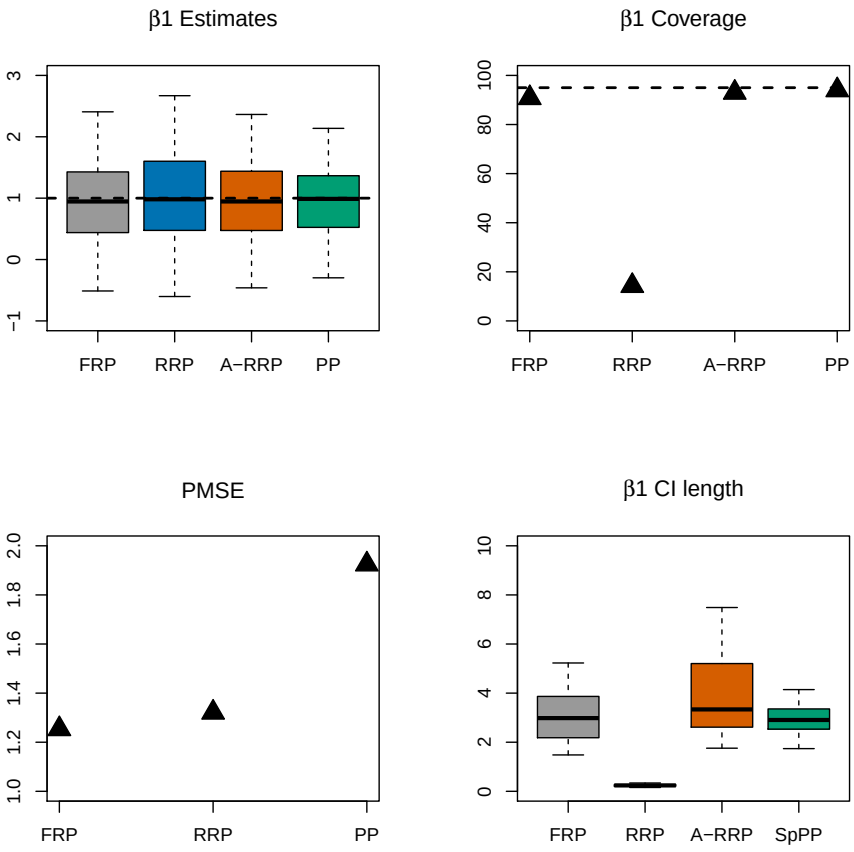


Figure 7: Poisson simulation study: compare projection-based models with predictive process on gridded knots. The point estimate distributions for β_1 are comparable (top left); while the coverage for RRP is much lower than the others, however after adjustment, the coverage for A-RRP is corrected and is comparable to FRP and PP (top right). Both FRP and RRP have better prediction performance than PP (bottom left). The length of the CIs for FRP and PP are comparable, while RRP produce much narrower CI; but after adjustment the CI gets much wider (bottom right). FRP = full model with random projection, RRP = restricted model with random projection, A-RRP = adjusted inference for RRP.

is therefore multivariate normal with mean $X\boldsymbol{\beta}$ and variance $\sigma^2 R_\phi + \tau^2 I$. **Our method can be used to facilitate model fitting in this case as follows:** We approximate the eigen-components of R_ϕ using random projections and obtain its first m eigenvectors $U_\phi = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m]$ and eigenvalues $D_\phi = \text{diag}(\lambda_1, \ldots, \lambda_m)$. Let $M_\phi = U_\phi D_\phi^{1/2}$ be the projection matrix, then we reduce the dimension of the latent variables by approximating $\boldsymbol{W}$ with $M_\phi \boldsymbol{\delta}$. For a specific value of ϕ, we can treat M_ϕ as fixed spatial covariates and $\boldsymbol{\delta}$ the corresponding coefficients. Write $X_\phi = [X, M_\phi]$ and $\boldsymbol{\beta}_\phi = (\boldsymbol{\beta}^T, \boldsymbol{\delta}^T)^T$ as the reparameterized design matrix and coefficients, respectively, then Y_i is approximated by $X_i\boldsymbol{\beta} + M_{\phi,i}\boldsymbol{\delta} + \epsilon_i$ and can be rewritten as $X_{\phi,i}\boldsymbol{\beta}_\phi + \epsilon_i$. We use a normal conjugate prior for $\boldsymbol{\beta}$, inverse gamma conjugate priors for σ^2 and τ^2, and a uniform prior for ϕ. Then, fitting the reduced-rank Bayesian probit model involves the following steps.

At the t^{th} iteration of the algorithm,

Step 1: Gibbs update for latent variables. Sample $\boldsymbol{Y}^{(t)}$ from $\boldsymbol{Y}|\boldsymbol{Z}, \boldsymbol{\beta}^{(t-1)}, \sigma^{2(t-1)}, \phi^{(t-1)}, \tau^{2(t-1)}$

(a) Compute projection matrix M_ϕ for $\phi^{(t-1)}$. Form X_ϕ and $\boldsymbol{\beta}_\phi$.

(b) For $i = 1, \ldots, n$, draw Y_i from

$$Y_i|\boldsymbol{Z}, \boldsymbol{Y_{-i}}, \boldsymbol{\beta}, \sigma^2, \phi, \tau^2 \sim \begin{cases} TN(X_{\phi,i}\boldsymbol{\beta}_\phi, \tau^2, 0, \infty), & \text{if } Z_i = 1 \\ TN(X_{\phi,i}\boldsymbol{\beta}_\phi, \tau^2, -\infty, 0), & \text{if } Z_i = 0, \end{cases}$$

where $TN(\mu_{Y_i}, \sigma^2_{Y_i}, 0, \infty)$ is a truncated normal distribution with lower bound 0, upper bound ∞, mean $X_{\phi,i}\boldsymbol{\beta}_\phi$ and variance τ^2.

Step 2: Gibbs update for $\boldsymbol{\beta}_\phi$.

Sample from $\boldsymbol{\beta}_\phi \mid \boldsymbol{Z}, \boldsymbol{Y}^{(t)}, \sigma^{2(t-1)}, \phi^{(t-1)}, \tau^{2(t-1)} \sim \texttt{MVN}\left(\hat{\boldsymbol{\beta}}_\phi, (\frac{1}{\tau^{2(t-1)}} X_\phi^T X_\phi + \Sigma_\beta^{-1})^{-1}\right)$,

where $\hat{\boldsymbol{\beta}}_\phi = (\frac{1}{\tau^{2(t-1)}} X_\phi^T X_\phi + \Sigma_\beta^{-1})^{-1} \frac{1}{\tau^{2(t-1)}} X_\phi^T \boldsymbol{Y}^{(t)}$, and $\Sigma_\beta = \begin{bmatrix} \Sigma_0 & 0 \\ 0 & \sigma^{2(t-1)} I_{m\times m} \end{bmatrix}$ with Σ_0 denotes the normal prior variance.

Step 3: Gibbs update for τ^2.

Step 4: Gibbs update for σ^2.

Step 5: Metropolis-Hastings update for ϕ.

We have not provided details for steps 3-5 since they remain the same as when fitting SGLMMs in general. Furthermore, techniques for dealing with non-identifiable parameters (Berrett and Calder, 2012, 2016) can also be used.

Case (2): We now consider the case where Gibbs sampling from the latent variable is not available. We first explain why the reparameterization for Case (1) is not suitable here, and then provide an alternative strategy. In Case (1) above, $\boldsymbol{W}$ is reparameterized with a low-rank representation, however, the dimension of latent variable $\boldsymbol{Y}$ remains high; $\boldsymbol{Y}$ is approximated by $X\boldsymbol{\beta} + M_\phi \boldsymbol{\delta} + \boldsymbol{\epsilon}$, and has a normal distribution with mean $X\boldsymbol{\beta}$ and covariance $\sigma^2 M_\phi M_\phi^T + \tau^2 I$. Constructing efficient MCMC to sample $\boldsymbol{Y}$ from its full conditional distribution is not easy due to its high dimensions. Hence, we propose an alternative: reduce the dimension of $\boldsymbol{Y}$ by approximating $\boldsymbol{W} + \boldsymbol{\epsilon}$ with $U_\theta D_\theta^{1/2} \boldsymbol{\delta}$, where U_θ and D_θ are eigenvectors and eigenvalues of $\sigma^2 R_\phi + \tau^2 I$, respectively. Hence, the eigencomponents here depend on all parameters $\boldsymbol{\theta} = (\sigma^2, \phi, \tau^2)^T$ of the covariance function. In fact U_θ is identical to U_ϕ from Case (1), and D_θ is identical to $\sigma^2 D_\phi + \tau^2 I_{m\times m}$. This alternative reparameterization provides some computational gains. The latent variable $\boldsymbol{Y}$ is now approximated by $X\boldsymbol{\beta} + M_\theta \boldsymbol{\delta} = [X, M_\theta](\boldsymbol{\beta}^T, \boldsymbol{\delta}^T)^T$ whose full conditional distribution has $m + p$ dimensions. Reducing the dimension of the posterior distribution allows for easier construction of efficient MCMC.

References

Berrett, C. and Calder, C. A. (2012). Data augmentation strategies for the Bayesian spatial probit regression model. *Computational Statistics and Data Analysis*, 56(3):478 – 490.

Berrett, C. and Calder, C. A. (2016). Bayesian spatial binary classification. *Spatial Statistics*, 16:72 – 102.

Hughes, J. and Haran, M. (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):139–159.

Schliep, E. M. and Hoeting, J. A. (2015). Data augmentation and parameter expansion for independent or spatially correlated ordinal data. *Computational Statistics and Data Analysis*, 90:1 – 14.