

# A MATRIX-FREE METHOD FOR SPATIAL- TEMPORAL GAUSSIAN STATE-SPACE MODELS

*Oregon State University*

## Supplementary Material

The supplement has four sections. The first section illustrates the effect of preconditioning while the second section provides a connection between REML estimation and non-linear gamma regression. The second section also provides new insight into the spectral decomposition of the precision matrix. The third section provides an analysis of monthly soil moistures across North America. Finally, the last section discusses possible extensions of the proposed method.

### S1 Effect of Preconditioning

To illustrate the effect of preconditioning in Section 4, we next focus on the eigenvalues for  $\lambda_2 RF^T FR^T + M$  and  $(\lambda_2 I + M)^{-1/2}(\lambda_2 RF^T FR^T + M)(\lambda_2 I + M)^{-1/2}$ . In Section 6.1 we sampled  $y_1, y_2, \dots, y_{10}$  from the state-space model (2.3)–(2.5) on a  $128 \times 128$  array at times  $t = 1, 2, \dots, 10$  with  $\lambda_0 = 1, \lambda_1 = 2, \lambda_2 = 0.01$ , and  $\lambda_3 = 1$ . However, most available softwares such as Matlab, Python and R failed to accurately compute the eigenvalues decomposition of the  $163840 \times 163840$  matrices as used in the simulation study in Section 6.1.

Instead, we choose  $r = 64$ ,  $c = 64$ , and  $s = 3$ , i.e., we sample  $y_1, y_2, y_3$  from the state-space model (2.3)–(2.5) on a  $64 \times 64$  array at time  $t = 1, 2, 3$  with  $\lambda_0 = 1$ ,  $\lambda_1 = 2$ ,  $\lambda_2 = 0.01$ , and  $\lambda_3 = 1$ . The data generation follows the same sampling steps in Section 6.1. The dimension of matrices  $\lambda_2 RF^T FR^T + M$  and  $(\lambda_2 I + M)^{-1/2}(\lambda_2 RF^T FR^T + M)(\lambda_2 I + M)^{-1/2}$  reduces to  $12288 \times 12288$ . Figure S1 provides the eigenvalues plot for these matrices. For the latter, we see that a large proportion of eigenvalues are clusters at 1 and others are below 1 but strictly about from 0, which results in a much smaller conditioning number compared with

$$\lambda_2 RF^T FR^T + M$$

and the original  $A$  matrix. As a consequence of this preconditioning, we get a speed up in the convergence of the Lanczos algorithm.

## S2 Non-convexity in REML estimation

To characterize the non-convexity in the REML estimation, we apply circulant embedding and rewrite spatial-temporal autoregressions (2.5) as

$$\psi_1 = G\psi_{2s} + \nu_1, \psi_i = G\psi_{i-1} + \nu_i, \quad i = 2, \dots, 2s.$$

Let  $\psi_E^T = (\psi_1^T, \dots, \psi_{2s}^T)$ ,  $\psi_E$  is normally distributed with mean 0 and a precision matrix  $\Gamma_E$ , whose spectral decomposition takes the form of  $\Gamma_E =$

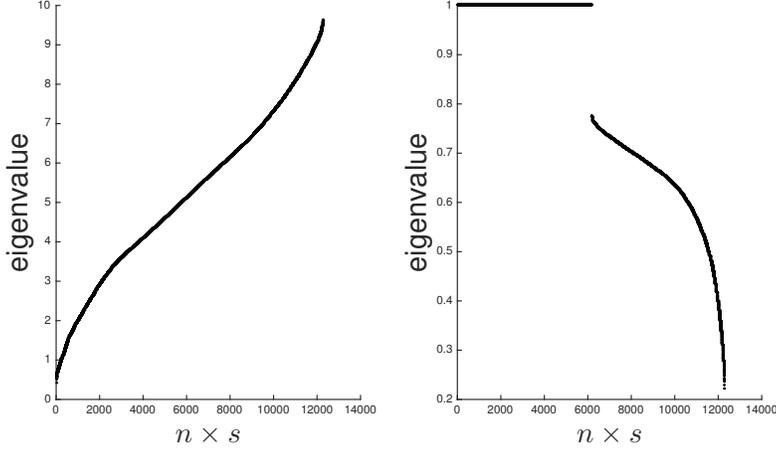


Figure S1: The left plot shows the eigenvalues of  $\lambda_2 R F^T F R^T + M$ . The right plot shows eigenvalues of  $(\lambda_2 I + M)^{-1/2} (\lambda_2 R F^T F R^T + M) (\lambda_2 I + M)^{-1/2}$

$R_E^T M_E R_E$ . The  $2ns \times 2ns$  matrix  $R_E$  is a block diagonal matrix with all  $n \times n$  diagonal blocks equal to  $P$ . The matrix  $M_E$  is block circulant matrix with non-zero blocks  $M_{E,(i,i)}$  (diagonal),  $M_{E,(i,i+1)}$  (upper diagonal),  $M_{E,(i-1,i)}$  (lower diagonal),  $M_{E,(1,2s)}$  (last block in first row) and  $M_{E,(2s,1)}$  (first block in last row) are such that

$$M_{E,(i,i)} = \Lambda_1^{-1} (I + e^{-\lambda_0 \Lambda}), \quad i = 1, \dots, 2s,$$

$$M_{E,(1,2s)} = M_{E,(2s,1)} = -e^{-\lambda_0 \Lambda / 2} \Lambda_1^{-1},$$

$$M_{E,(j,j+1)} = M_{E,(k-1,k)} = -e^{-\lambda_0 \Lambda / 2} \Lambda_1^{-1}, \quad j = 1, \dots, 2s - 1, \quad k = 2, \dots, 2s.$$

Furthermore, symmetric permutation of rows and columns on  $M_E$  results in a block diagonal matrix with circulant blocks. Let  $\mathcal{P}$  be the permutation

matrix with  $\mathcal{P}^T \mathcal{P} = I$ , then  $M_E = \mathcal{P}^T \mathcal{K} \mathcal{P}$ . The matrix  $\mathcal{K}$  is a block diagonal matrix with  $2s \times 2s$  diagonal blocks  $\mathcal{K}_i$ ,  $i = 1, \dots, n$ . Each  $\mathcal{K}_i$  is circulant matrix with spectral decomposition denoted as  $\mathcal{K}_i = \Phi^T \mathcal{T}_i \Phi$ , where  $\Phi$  corresponds to the discrete Fourier transform and  $\mathcal{T}_i$  is a diagonal matrix with  $j$ th element

$$\mathcal{T}_{i,j} = \rho_i(1+e^{-\lambda_0\rho_i})/(1-e^{-\lambda_0\rho_i}) - 2\{\rho_i e^{-\lambda_0\rho_i/2}/(1-e^{-\lambda_0\rho_i})\} \cos\{\pi(j-1)/(2s)\} \quad (\text{S2.1})$$

for  $i = 1, \dots, n$  and  $j = 1, \dots, 2s$ , where  $\rho_i$  is the  $i$ th diagonal element of  $\Lambda$ .

Let  $\Phi_E$  denote a  $2ns \times 2ns$  block diagonal matrix with all  $2s \times 2s$  diagonal blocks equal to  $\Phi$ . Then,  $\mathcal{K} = \Phi_E^T \mathcal{T} \Phi$ , where  $\mathcal{T}$  is a diagonal matrix with diagonal blocks  $\mathcal{T}_1, \dots, \mathcal{T}_n$ . Therefore, the spectral decomposition of the original precision matrix  $\Gamma_E$  is

$$\Gamma_E = \Phi_E^T \mathcal{P}^T R_E^T \mathcal{T} R_E \mathcal{P} \Phi_E. \quad (\text{S2.2})$$

Now we follow the settings in Section 3.1. Let  $\eta_E^T = (\eta_1^T, \dots, \eta_{2s}^T)$ . It is immediate that  $\eta_E \sim N(0, M_E^{-1})$ . Next, let  $n_+ \times 2ns$  matrix  $F_E$  be the column binding of  $F$  and a  $n_+ \times 2ns$  0 matrix, and assume that

$$u_E = \begin{pmatrix} y \\ 0 \end{pmatrix}, X_E = \begin{pmatrix} F_E \\ R_E \mathcal{P} \Phi_E \end{pmatrix}, \zeta_E = \begin{pmatrix} \epsilon \\ \eta_E \end{pmatrix}, Q_E = \begin{pmatrix} \lambda_3 I_{n_+} & 0 \\ 0 & \mathcal{T} \end{pmatrix},$$

where  $u_E$  is a  $(n_+ + 2ns)$  column vector and  $Q_E$  is a  $(n_+ + 2ns) \times (n_+ + 2ns)$

diagonal matrix. The state-space model with circulant embedding takes the

regression form

$$u_E = X_E \psi_E + \zeta_E,$$

where  $\zeta_E \sim N(0, Q_E^{-1})$ , which is very similar to (3.1) and the state vectors can be reestimated as

$$X_E^T Q_E X_E \hat{\psi}_E = X_E^T Q_E u_E.$$

The log-residual likelihood function in Section 3.2 now can be rewritten as

$$2l_E(\lambda) = \log |Q_E| - \log |X_E^T Q_E X_E| - (u_E - X_E \hat{\psi}_E)^T Q_E (u_E - X_E \hat{\psi}_E).$$

Next, denote by  $H_E = X_E (X_E^T Q_E X_E)^{-1} X_E^T Q_E$  the hat matrix. Let  $\delta_i$  be the square of the  $i$ th element of the residual vector  $u_E - X_E \hat{\psi}_E$  and let  $q_{E,i}$  be the  $i$ th diagonal element of  $Q_E$ . Take  $\delta_i^* = \delta_i / (1 - h_{E,i})$ , where  $h_{E,i}$  the  $i$ th diagonal element of the hat matrix  $H_E$ . The score equations in (3.3) are then same as

$$\begin{aligned} \partial l_E / \partial \lambda_3 &= (1/2) \sum_{i=1}^{2ns} (1 - h_{E,i}) (\delta_i^* - \lambda_3) / \lambda_3^2, \\ \partial l_E / \partial \lambda_j &= (1/2) \sum_{i=n_++1}^{n_++2ns} (\partial q_{E,i} / \partial \lambda_j) (1 - h_{E,i}) (\delta_i^* - q_{E,i}) / q_{E,i}^2, \quad j = 0, 1, 2. \end{aligned}$$

These score equations coincide with estimating equations of a gamma regression where the response variables are adjusted squared residuals  $\delta_i^*$  and follow independent gamma distribution. Furthermore, we have inverse link, nonlinear predictors  $q_{E,i}$  as a non-linear function of  $\lambda_0, \dots, \lambda_3$  and prior

weights  $(1 - h_{E,i})$ . The non-linearity in the gamma regression specifies the exact nature of the non-convexity in the REML estimation.

### S3 Analysis of soil moisture data

The Climate Prediction Center of the National Weather Service provides monthly mean soil moistures at  $0.5^\circ \times 0.5^\circ$  spatial resolution for the time period 1948 to 2014. For further references, see Fan and van den Dool (2004). Here, we consider a subset of the data with latitudes between  $40^\circ\text{N}$  and  $50^\circ\text{N}$ , longitudes between  $95^\circ\text{W}$  and  $75^\circ\text{W}$ , for the time period from January, 2009 to December, 2009. The subset constitutes  $s = 12$  months of data, and is spatially embedded into a  $40 \times 20$  array. Due to presence of lakes and water bodies, there are 36 array cells with no observations. Let  $y_1, \dots, y_{12}$  be the observed monthly mean soil moisture in this subset. Next, we consider the state-space model in (2.4)–(2.5). We consider two different scenarios. First, the underlying state vectors  $\psi_t$ ,  $t = 1, \dots, 12$ , follow spatial-temporal autoregressions as in equation (2.4) at the original spatial resolution  $0.5^\circ \times 0.5^\circ$ . Second, the state vectors follow spatial-temporal autoregressions at a finer spatial resolution  $0.125^\circ \times 0.125^\circ$ .

In Scenario 2, we split each array cell into  $4 \times 4$  sub-cells so that  $\psi_t$ ,  $t = 1, \dots, 12$  lie on a  $160 \times 80$  spatial array, and  $rcs = 153600$ . Accordingly,

Table S1: REML estimates of precision parameters for soil moisture data under no splitting (Scenario 1) and  $4 \times 4$  splitting (Scenario 2) of the original array. The standard errors are in parentheses.

Parameters	$\lambda_0$	$\lambda_1$	$\lambda_2$	$\lambda_3$
Scenario 1	0.325 (0.005)	220.289 (22.772)	473.386 (258.081)	0.977 (0.205)
Scenario 2	0.193 (0.003)	589.850 (89.233)	327.251 (35.156)	0.976 (0.002)

we construct the averaging matrix  $F_t$ , each of order  $12800 \times 764$ , such that  $F_t\psi_t$  provides the vector of average of state values at the original  $0.5^\circ \times 0.5^\circ$  spatial resolution. This scenario is particularly useful if we want to obtain spatial interpolation at a finer resolution. Furthermore, finer resolution allows us to achieve approximate inference from the limiting continuum geostatistical model; see, e.g., Besag and Mondal (2005), and Dutta and Mondal (2015, 2016) for examples of such inferences in spatial statistics. At spatial resolution  $0.125^\circ \times 0.125^\circ$ , computations are particularly challenging as we have a  $153600 \times 153600$  non-sparse block triangular matrix  $A$ , where each block of  $A$  is of order  $12800 \times 12800$ .

Table S1 summarizes the REML estimates for standardized monthly

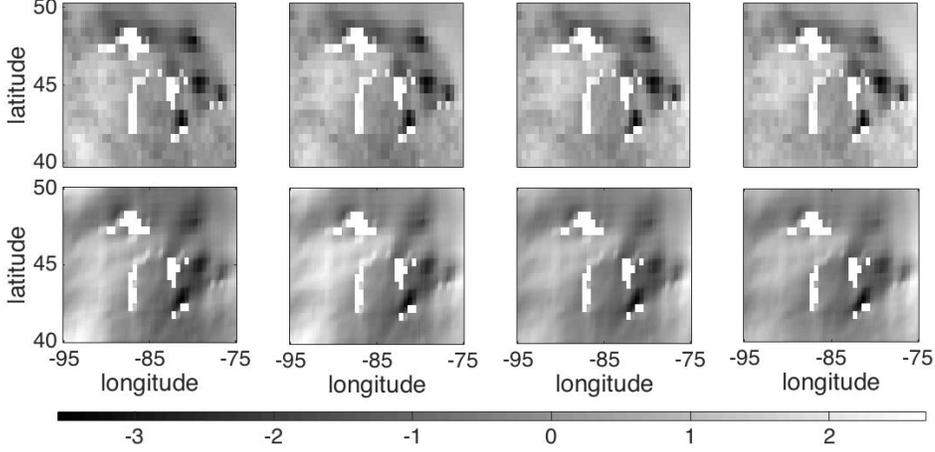


Figure S2: The top panel shows image plots of  $y_9, \dots, y_{12}$ . The bottom panel displays  $\hat{\psi}_9, \dots, \hat{\psi}_{12}$  at  $0.125^\circ \times 0.125^\circ$  spatial resolution. The horizontal bar gives the gray scale.

soil moisture values. In Scenario 2, the estimate of  $\lambda_0$  decreases, and the estimate of  $\beta = \lambda_1 / (4\lambda_1 + 2\lambda_2)$  increases. This explains that  $\psi_t$ s are spatially and temporally more dependent at spatial resolution  $0.125^\circ \times 0.125^\circ$ . Finally, Figure S2 displays the actual observations, and the prediction for latent variables for the last four months. We see that the model performs reasonably well in predicting soil moisture at spatial resolution  $0.125^\circ \times 0.125^\circ$ . In particular, in Scenario 2, about 76% percentage of total variation is explained by spatial-temporal state-space model. The algorithms detailed in Section 4 made these computations possible without dimension reduction or ensembles of stochastic simulations.

## S4 Further discussion

Within the framework of the paper, we can consider further possible extensions of the spate-space dynamical model in (2.4) and (2.5). For example, we can consider higher neighborhood-order conditional autoregressions, or fractional spatial random fields. In the former, we replace the precision matrix  $C$  in equation (2.5) with  $C = J(W)$  where  $J$  is a suitable positive polynomial; see Mondal (2017) for details. In the latter,  $C$  in equation (2.5) is replaced with  $C^\kappa$ ,  $\kappa > 0$ , which corresponds to the fractional Laplacian differenced random fields and approximate spatial Matérn processes. For both the higher neighborhood-order conditional autoregressions and the fractional Laplacian differenced random fields,  $M$  still provides the eigenvectors of the precision matrix of the state variables. We can thus calculate fast matrix-vector products without storing any matrices and pursue computations as presented in Section 4. If needed, we can also include covariate information and consider mixed effect models. Furthermore, various complex spatial-temporal autoregressions arise from small time-step discretization of a wide variety of SDPEs. Sections 5 and 6 demonstrate how we can implement these elaborate models. The computations proposed here are better than those presented in Rue et al. (2009) and Lindgren et al. (2011). The method presented here will have further applications in data

assimilation and computer simulations involving discrete linearizations of complex non-linear stochastic particle differential equations. These applications typically involve a small or moderate value of  $s$  and a large value of  $n$  and are ideal for our matrix-free computations.

If in certain applications, both  $s$  and  $n$  are very large, one can also adopt various strategies including parallel and distributive computing. The discrete cosine transform, the matrix-vector product, the Lanczos algorithm, the stochastic trace approximations are in fact all parallelizable; see e.g., Frigo and Johnson (2005) and Kim and Chronopoulos (1991) for details.

It is worthwhile to point out that the proposed computations do not break down easily for long-range dependencies which may well be the case for the circulant embedding approximation of covariance matrices. We are working with precision matrices directly and not the covariances. The eigenvalues of a precision matrix typically correspond to the reciprocal of the spectral density whereas the eigenvalues of a covariance matrix correspond to the spectral density. Thus, precision matrices including sparse precision matrices are well suited for long range dependencies. In Dutta and Mondal (2016), we did extensive computations with intrinsic and long range dependence Matérn and power variogram models. The results in Tables 1-5 of Dutta and Mondal (2016) clearly show that the computations did not break

down. Similarly in Section 6.2, we considered long-range dependencies, but again the computations did not break down.

In this paper, we focused on spatial-temporal model where the time dynamics follow an autoregressive structure of order 1 (i.e.,  $\psi_{t+1}$  given  $\psi_1, \dots, \psi_t$  depends only on  $\psi_t$  through the dependence matrix  $\exp(-\lambda_0 C/2)$ ). Furthermore, equations (S2.1)–(S2.2) in the Supplement explicitly show how both the temporal autoregressive structure and the spatial dependence structure enter into the spectral factorization of the spatial-temporal models. This was possible because the two-dimensional discrete cosine transform of  $\psi_t$  breaks the spatial dependence structure and convert  $\psi_t$  into independent components. This can be extended further to construct spatial-temporal models where the temporal dynamics have an autoregressive structure of order  $p$  or have a fractional dependence structure. The construction of general spatial-temporal lattice systems will be pursued in a future work.

For applications in Sections 6 and the Supplement, there might be benefit in assuming that the precision parameters are random variables and obey some probability distributions. Such ideas can be incorporated in a Bayesian hierarchical model and one can then pursue a matrix-free sampling based-method for inference. The advantage of this approach is that we

can even allow precision parameters to vary with time. Some penalization may also be possible. However, obtaining frequentist penalized estimation of precision parameters presents challenges, especially when these will be treated as random effects and are high-dimensional; see e.g., the discussion of Firth (2006).

Next, spatial-temporal non-Gaussian state-space models and Bayesian computations are also of interest. Examples include binomial or Poisson models and for application see Brix and Diggle (2001). They often arise in generalized linear model when data  $y_t$  is a response to a linear predictor  $\chi_t$  that can be represented as

$$\chi_t = T_t\delta + F_t\psi_t + \epsilon_t. \tag{S4.1}$$

In the above  $\delta$  is covariate effect,  $T_t$  provides covariate information at time  $t$ ,  $\psi_t$ ,  $F_t$  and  $\epsilon_t$  are the same as those defined in Section 2. The model (S4.1) generalizes the Gaussian state-space model in (2.3)–(2.5). REML analysis does not extend to the non-Gaussian model in (S4.1) and statistical inference typically requires (Bayesian) Markov chain Monte Carlo (MCMC) and other simulation-based computations. The best linear unbiased prediction calculations and the method of conditional simulations presented in this paper is however relevant in this context. These computations can be used to develop various matrix-free scalable block Gibbs-Metropolis-Hastings and

Hamiltonian MCMC computations.

Finally, there are some challenges to routinely implement the described algorithms in R. This is because some key functions such as computing sparse incomplete Cholesky decomposition and efficient implementation of the two dimensional discrete cosine transformation are still missing in its current library of sparse matrices. Nevertheless, we implemented the algorithms in Matlab by modifying the prototype codes provided in the supplements of the papers by Dutta and Mondal (2015, 2016). However, there is still some way to go towards engineering implementation. Currently, work is in progress to develop an R package by converting these prototype codes into a set of near optimized codes in C and R so that the methods can be applied in a robust way to various data sets.

## Bibliography

- Besag, J. and D. Mondal (2005). First-order intrinsic autoregressions and the de Wijs process. *Biometrika* 92(4), 909–920.
- Brix, A. and P. J. Diggle (2001). Spatiotemporal prediction for log-Gaussian Cox processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63(4), 823–841.
- Dutta, S. and D. Mondal (2015). An h-likelihood method for spatial mixed

- linear models based on intrinsic auto-regressions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 77(3), 699–726.
- Dutta, S. and D. Mondal (2016). REML estimation with intrinsic Matérn dependence in the spatial linear mixed model. *Electronic Journal of Statistics* 10(2), 2856–2893.
- Fan, Y. and H. van den Dool (2004). Climate prediction center global monthly soil moisture data set at 0.5 resolution for 1948 to present. *Journal of Geophysical Research: Atmospheres* 109, D10102.
- Firth, D. (2006). Contribution to the discussion of “Double hierarchical generalized linear models” by Y. Lee and J. A. Nelder. *Journal of the Royal Statistical Society Series C (Applied Statistics)* 55(2), 168–170.
- Frigo, M. and S. G. Johnson (2005). The design and implementation of FFTW3. *Proceedings of the IEEE* 93(2), 216–231.
- Kim, S. K. and A. T. Chronopoulos (1991). A class of lanczos-like algorithms implemented on parallel computers. *Parallel Computing* 17, 763–778.
- Lindgren, F., H. Rue, and J. Lindström (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73(4), 423–498.

Mondal, D. (2017). On edge correction of conditional and intrinsic autoregressions. To appear in *Biometrika*.

Rue, H., S. Martino, and N. Chopin (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.