Multivariate spatial modeling and BLMC Factor model

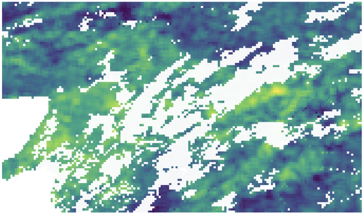
Lu Zhang

Biostats Division
Population and Public Health Sciences Department
University of Southern California
lzhang63@usc.edu

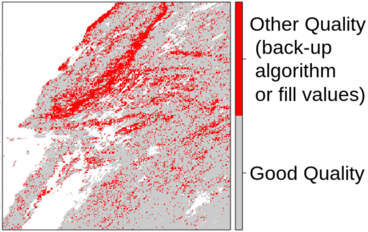
Apr 03, 2024

"The only way this planet is going to deal with its Global challenges to feed people, supply them with medical care, supply them with energy, electricity, and to make sure they're not burnt to a crisp because of global warming is the effective use of data" —- Kenneth Cukier: Big data is better data (Ted talk)

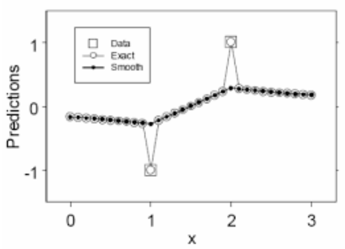
Different type of "messy" data



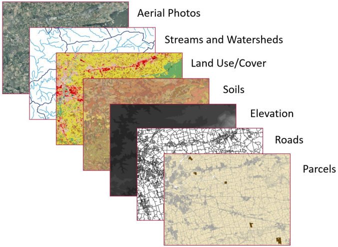
(a) blocked by clouds



(b) poor quality

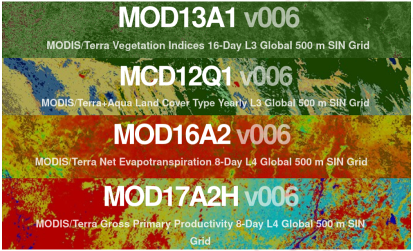


(c) measurement error



(d) combine multiple datasets

Motivating Data (source)



- Vegetation Indices data
- Global Terrestrial Evapotranspiration Product,
- Gross Primary Productivity data
- Land cover type data

Motivating Data (key features)

- **Large scale spatial data:** $N = 1,020,000$ observed locations
- **Multiple responses from different sources:** $q = 10$ responses
- **Misalignment (missing observations):** Not all responses are recorded on every observed location (Heatmap) Has missing responses on over 65% of observed locations

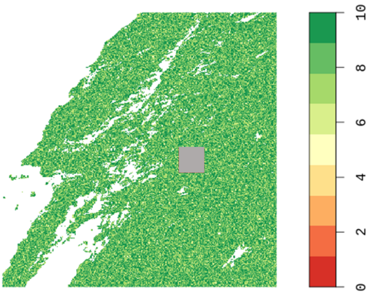


Figure: Heat-map of counts of observed responses, the greener the color, the higher the value. Grey square: a blocked region with no responses.

Motivating Data (illustration)

- **NDVI** (Normalized Difference Vegetation Index) is an index to identify vegetation amount and measure plant health vitality.
- **Range**:(-1, 1); The higher the VI, the higher amounts of vegetation. Shift to (0, 2) and transform it by log for better model fitting
- *Understanding the global distribution of vegetation types as well as their biophysical and structural properties and spatial variations.*

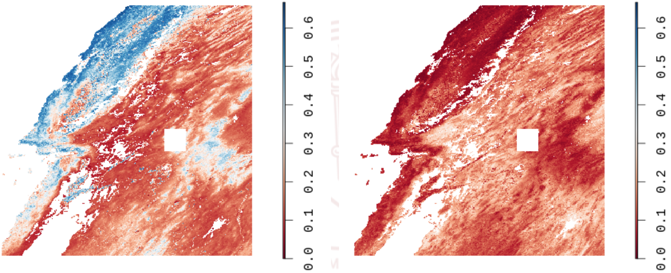


Figure: Colored transformed NDVI and red reflectance images of zone h08v05 western United States; Same color scale; The warmer the color, the lower the value.

Motivating Data (illustration)

- **Red Reflectance** Waveband in red (655 nm)
- Healthy green plants absorb red light
- **Unit**:Reflectance: ratio of the amount of light leaving a target to the amount of light striking the target. (transformed by log)

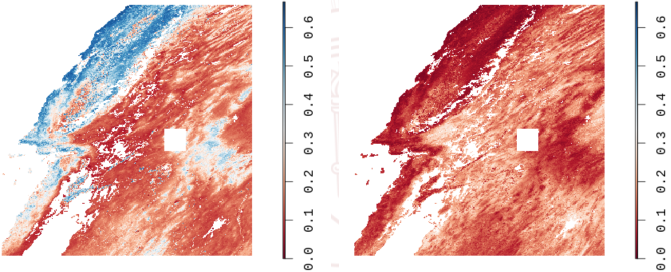


Figure: Colored transformed NDVI and red reflectance of zone h08v05 western United States; Same color scale; The warmer the color, the lower the value.

Motivating Data

- **Other 8 responses**: Enhanced Vegetation Index (EVI), Gross Primary Productivity (GPP), Net Photosynthesis (PSN), blue reflectance, evapotranspiration (ET), latent heat flux (LE), potential ET (PET), potential LE (PLE)
- **Explanatory variables**: Land type data, indicator of no vegetation or urban area based on the 2016 land cover data (+ intercept)

Analysis goals:

- Impute the missing responses
- Predict in the blocked region
- Measure the impact of the land type on different measurements of greenness

Multivariate Spatial Regression Model

- All Responses: $\mathbf{y}(\mathbf{s}) = (y_1(\mathbf{s}), \ldots, y_q(\mathbf{s}))^\top$
- Explanatory variables: $\mathbf{x}(\mathbf{s}) = (x_1(\mathbf{s}), \ldots, x_p(\mathbf{s}))^\top$
- **Multivariate Spatial Regression Model**

$$\mathbf{y}(\mathbf{s}) = \boldsymbol{\beta}^\top \mathbf{x}(\mathbf{s}) + \boldsymbol{\omega}(\mathbf{s}) + \boldsymbol{\epsilon}(\mathbf{s}) \ , \ \mathbf{s} \in \mathcal{D} \quad (1)$$

$\boldsymbol{\beta}$: $p \times q$ regression coefficient matrix $\mathcal{D}$: Study domain

$\boldsymbol{\epsilon}(\mathbf{s}) = (\epsilon_1(\mathbf{s}), \ldots, \epsilon_q(\mathbf{s}))^\top$: noise $\boldsymbol{\epsilon}(\mathbf{s}) \overset{iid}{\sim} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$

$\boldsymbol{\omega}(\mathbf{s})$: latent processes, capture the underlying spatial pattern that cannot be explained by explanatory variables for all responses.

Multivariate Spatial Regression Model

- $\boldsymbol{\omega}(\mathbf{s}) \sim \mathrm{GP}(\mathbf{0}_q, \mathbf{C}(\cdot,\cdot))$ Gaussian process with cross-covariance function $\mathbf{C}(\cdot,\cdot)$

- For $(\mathbf{s}, \mathbf{s}') \in \mathcal{D} \times \mathcal{D}$, $\mathbf{C}(\mathbf{s}, \mathbf{s}')$ is $q \times q$ matrix with

$$\{\mathbf{C}(\mathbf{s}, \mathbf{s}')\}_{ij} = \mathrm{cov}(\boldsymbol{\omega}_i(\mathbf{s}), \boldsymbol{\omega}_j(\mathbf{s}')) \quad 1 \leq i, j \leq q \; .$$

  How $\boldsymbol{\omega}(\mathbf{s})$ affect $\boldsymbol{\omega}(\mathbf{s}')$; Well-defined if for any spatial locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ and any integer n, $\{\mathbf{C}(\mathbf{s}_i, \mathbf{s}_j)\}_{i=1,j=1}^n$ is positive definite

- Denote $\omega^\top = (\omega_1(\mathbf{s}_1), \ldots, \omega_q(\mathbf{s}_1), \omega_1(\mathbf{s}_2), \ldots \omega_q(\mathbf{s}_2), \ldots, \omega_1(\mathbf{s}_n), \ldots \omega_q(\mathbf{s}_n))$

- $\Sigma_\omega = \{\mathbf{C}(\mathbf{s}_i, \mathbf{s}_j)\}_{i=1,j=1}^n$ the covariance matrix of ω, $nq \times nq$ matrix

Challenges and Motivation

- Big-data problems, observed locations in millions
  - Σ_ω : $nq \times nq$ matrix. Calculation needs Cholesky decomposition of this $nq \times nq$ matrix. For $q = 1$, $n = 1,000,000$, this decomp. in R would require over 3700 GB RAM, and hundreds of hours for an i7 processor.
  - Challenging for obtaining Bayesian inference when using iterative method like MCMC algorithm
- Pathological geometry features of the posterior
- Studies on scalable spatial modeling focus on single process modeling. Limited discussions on scalable multivariate spatial data modeling, especially for data with missing responses

Extend scalable modeling strategies for a single process to Bayesian multivariate process modeling.

Multivariate spatial modeling: Separable models

Let ρ be a valid correlation function for a univariate spatial process, Let T be a $q \times q$ positive definite matrix and let

$$\mathbf{C}(\mathbf{s}, \mathbf{s}') = \rho(\mathbf{s}, \mathbf{s}') \cdot T \; . \tag{2}$$

In (2), $T \equiv (T_{ij})$ is interpreted as the covariance matrix associated with $\omega(\mathbf{s})$, and ρ attenuates association as s and s' become farther apart.

$$\Sigma_\omega = H \otimes T \; ,$$

where $(H)_{ij} = \rho(\mathbf{s}_i, \mathbf{s}_j)$ and $\otimes$ denotes the Kronecker product. Σ_ω is evidently positive definite since H and T are.

Multivariate spatial modeling: Separable models

- pros:
  - Convenient to work with

$$|\Sigma_\omega| = |H|^q |T|^n; \Sigma_\omega^T = H^{-1} \otimes T^{-1}$$

  - Accelerate posterior sampling through Gibbs sampler under certain prior settings
- cons:
  - $\mathbf{C}(\mathbf{s}, \mathbf{s}')$ is symmetric , i.e., $\text{cov}(\omega_l(\mathbf{s}_i), \omega_{l'}(\mathbf{s}_{i'}))) = \text{cov}(\omega_{l'}(\mathbf{s}_i), \omega_l(\mathbf{s}_{i'})))$ for all i, i', l, l'.
  - If ρ is stationary, the generalized correlation (coherence in time series literature)

$$\frac{\text{cov}(\omega_l(s), \omega_{l'}(s+h))}{\sqrt{\text{cov}(\omega_l(s), \omega_l(s+h))\text{cov}(\omega_{l'}(s), \omega_{l'}(s+h))}} = \frac{T_{ll'}}{\sqrt{T_{ll} T_{ll'}}}$$

  is regardless of $\mathbf{s}$ and h
  - ρ is isotropic and strictly decreasing, then the spatial range is identical for each component of $\omega(\mathbf{s})$

Kronecker Product

Definition

The Kronecker product, denoted by $\otimes$, is an operation on two matrices of arbitrary size resulting in a block matrix.

- If A is an $m \times n$ matrix and B is a $p \times q$ matrix, then the Kronecker product $A \otimes B$ is an $mp \times nq$ matrix.
- Operationally, $A \otimes B$ is obtained by multiplying each element a_{ij} of A by the matrix B.

Example

Given $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ and $B = \begin{pmatrix} e & f \\ g & h \end{pmatrix}$, the Kronecker product $A \otimes B$: is

$$\begin{pmatrix} ae & af & be & bf \\ ag & ah & bg & bh \\ ce & cf & de & df \\ cg & ch & dg & dh \end{pmatrix}$$

Multivariate spatial modeling: Coregionalization models

Linear Model of Coregionalization (LMC)

- $$\boldsymbol{\omega}(\mathbf{s}) = A\mathbf{w}(\mathbf{s}) \ ,$$
  where $\mathbf{w}(\mathbf{s})^\top = (\mathbf{w}_1(\mathbf{s}), \ldots, \mathbf{w}_q(\mathbf{s}))$, A is a $q \times q$ matrix.
- Let the $\mathbf{w}_j(s)$ process have mean μ_j, variance 1, and correlation function $\rho_j(h)$. Then $\mathbb{E}(\boldsymbol{\omega}(s)) = A\boldsymbol{\mu}$ where $\boldsymbol{\mu}^\top = (\mu_1, \ldots, \mu_q)$ and
  $$\Sigma_{\boldsymbol{\omega}(\mathbf{s}),\boldsymbol{\omega}(\mathbf{s}')} = \mathbf{C}(\mathbf{s}, \mathbf{s}') = \sum_{k=1}^{q} \rho_k(\mathbf{s}, \mathbf{s}') T_j \ ,$$
  where $T_j = a_j a_j^\top$ with a_j the jth column of A
- Define $T = \sum_j T_j$. The one-to-one relationship between T and lower triangular A is standard.

Coregionalization models

- Each $\omega_j(\mathbf{s})$ has its own range.
- Equivalence of likelihoods.
  - In the context of $\boldsymbol{\omega}(\mathbf{s}) = A\mathbf{w}(\mathbf{s})$ where the $\mathbf{w}_j(\mathbf{s})$ are mean zero Gaussian processes, by taking A to be lower triangular the equivalence and associated reparametrization are easy to see

$$p(\boldsymbol{\omega}_1(\mathbf{s}))p(\boldsymbol{\omega}_2(\mathbf{s}) \mid \boldsymbol{\omega}_1(\mathbf{s})) \cdots p(\boldsymbol{\omega}_q(\mathbf{s}) \mid \boldsymbol{\omega}_1(\mathbf{s}), \ldots, \boldsymbol{\omega}_{q-1}(\mathbf{s}))$$

  - Advantages to working with the conditional form of the model are certainly computational and possibly mechanistic or interpretive.
- Other extensions, the column number of A need not be equal to q

Build $\omega(s)$ from univariate models

More general LMC model

$$\mathbf{C}(\mathbf{s}, \mathbf{s}') = \sum_{k=1}^{K} \underbrace{\rho_k(\mathbf{s}, \mathbf{s}')}_{\text{scale}} \underbrace{\boldsymbol{\lambda}_k \boldsymbol{\lambda}_k^\top}_{q \times q \text{ matrix}} , \; \boldsymbol{\omega}(\mathbf{s}) = \sum_{k=1}^{K} \underbrace{f_k(\mathbf{s})}_{scale} \underbrace{\boldsymbol{\lambda}_k}_{q \times 1} = \boldsymbol{\Lambda}^\top \mathbf{f}(\mathbf{s})$$

- $\boldsymbol{\Lambda}^\top = [\boldsymbol{\lambda}_1 : \cdots : \boldsymbol{\lambda}_K]$: $q \times K$ loading matrix. $\boldsymbol{\lambda}_k \in \mathbb{R}^q$
- $\mathbf{f}(\mathbf{s}) = (f_1(\mathbf{s}), \ldots, f_K(\mathbf{s}))^\top$, $f_k(\mathbf{s})$ are independent, each follows a random field with covariance function $\rho_k(\cdot, \cdot)$.

Spatial Regression Model with $\omega(\mathbf{s})$ modeled by LMC

$$Model: \begin{array}{l} \mathbf{y}(\mathbf{s}) = \boldsymbol{\beta}^\top \mathbf{x}(\mathbf{s}) + \boldsymbol{\Lambda}^\top \mathbf{f}(\mathbf{s}) + \boldsymbol{\epsilon}(\mathbf{s}) \;,\; \mathbf{s} \in \mathcal{D} \;, \\ \boldsymbol{\epsilon}(\mathbf{s}) \overset{iid}{\sim} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}) \;,\; f_k(\mathbf{s}) \sim \mathrm{GP}(0, \rho_k(\cdot, \cdot\,; \psi_k)) \end{array}$$

Notations

- Define $\boldsymbol{\gamma} = [\boldsymbol{\beta}^\top, \boldsymbol{\Lambda}^\top]^\top$: $(p + K) \times q$ matrix

Priors for $\boldsymbol{\gamma}, \boldsymbol{\Sigma}, \{\psi_k\}_{k=1}^K$?

Matrix-Normal Inverse-Wishart (MNIW)

Consider priors for $\{\gamma, \boldsymbol{\Sigma}\}$

$$\gamma \mid \boldsymbol{\Sigma} \sim \mathrm{MN}(\boldsymbol{\mu}_\gamma, \mathbf{V}_\gamma, \boldsymbol{\Sigma}) \;,\; \boldsymbol{\Sigma} \sim \mathrm{IW}(\boldsymbol{\Psi}, \nu)$$

$\boldsymbol{\mu}_\gamma$: $(p+K) \times q$ matrix
$\mathbf{V}_\gamma$: $(p+K) \times (p+K)$ positive definite matrix.

$$\{\gamma, \boldsymbol{\Sigma}\} \sim \mathrm{MNIW}(\boldsymbol{\mu}_\gamma, \mathbf{V}_\gamma, \boldsymbol{\Psi}, \nu).$$

- Matrix-Normal distribution: If $\mathbf{Z}_{n\times p} \sim \mathrm{MN}_{n,p}(\mathbf{M}, \mathbf{U}, \mathbf{V})$ then $\mathrm{vec}(\mathbf{Z}) = \{\mathbf{z}_1^\top, \ldots, \mathbf{z}_p^\top\}^\top$ follows a Gaussian distribution

$$\mathrm{vec}(\mathbf{Z}) \sim \mathrm{N}_{np}(\mathrm{vec}(\mathbf{M}), \mathbf{V} \otimes \mathbf{U}).$$

Bayesian LMC factor model (BLMC)

$$\textit{Model}: \begin{aligned} &\mathbf{y}(\mathbf{s}) = \boldsymbol{\beta}^\top \mathbf{x}(\mathbf{s}) + \boldsymbol{\Lambda}^\top \mathbf{f}(\mathbf{s}) + \boldsymbol{\epsilon}(\mathbf{s}) \;,\; \mathbf{s} \in \mathcal{D} \;, \\ &\boldsymbol{\epsilon}(\mathbf{s}) \overset{iid}{\sim} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}) \;,\; f_k(\mathbf{s}) \sim \mathrm{GP}(0, \rho_k(\cdot, \cdot\,; \psi_k)) \end{aligned}$$

$$\textit{Priors}: \{\boldsymbol{\gamma}, \boldsymbol{\Sigma}\} \sim \mathrm{MNIW}(\boldsymbol{\mu}_\gamma, \mathbf{V}_\gamma, \boldsymbol{\Psi}, \nu) \;,\; p(\psi_k)$$

The design of the prior yields the conditional posterior distributions of all parameters except for $\{\psi_k\}_{k=1}^K$ in closed form.

Bayesian LMC factor model with diagonal $\boldsymbol{\Sigma}$

When restricting $\boldsymbol{\Sigma}$ to be diagonal ($\boldsymbol{\Sigma} = \mathrm{diag}(\{\sigma_i^2\}_{i=1}^q)$)

$$\textit{Model}: \begin{array}{l} \mathbf{y}(\mathbf{s}) = \boldsymbol{\beta}^\top \mathbf{x}(\mathbf{s}) + \boldsymbol{\Lambda}^\top \mathbf{f}(\mathbf{s}) + \boldsymbol{\epsilon}(\mathbf{s}) \;,\; \mathbf{s} \in \mathcal{D} \;, \\ \boldsymbol{\epsilon}(\mathbf{s}) \overset{iid}{\sim} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}) \;,\; f_k(\mathbf{s}) \sim \mathrm{GP}(0, \rho_k(\cdot, \cdot\,; \psi_k)) \end{array}$$

$$\textit{Priors}: \sigma_i^2 \sim \mathrm{IG}(a, b_i) \;,\; \boldsymbol{\gamma} \mid \boldsymbol{\Sigma} \sim \mathrm{MN}(\boldsymbol{\mu}_\gamma, \mathbf{V}_\gamma, \boldsymbol{\Sigma}) \;,\; p(\psi_k)$$

Flexible in modeling

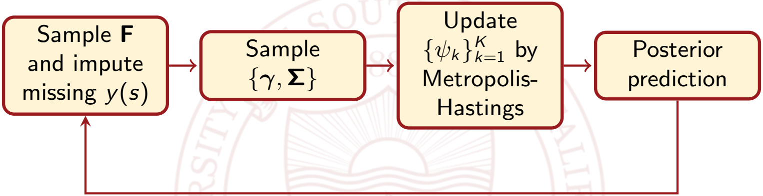
q is small

- $K \geq q$:
- $\boldsymbol{\Sigma}$ p.d: Assume potential inner correlation in measurement error

q is large

- $K < q$:
- $\boldsymbol{\Sigma}$ Diagonal: Assume independence measurement error

Block update MCMC algorithm



Notations

- $\mathbf{F} = [\mathbf{f}(\mathbf{s}_1) : \cdots : \mathbf{f}(\mathbf{s}_n)]^\top$: $n \times K$ matrix

Scalable, i.e. the computational burden and storage requirement is linear to n, once f_ks are modeled by scalable spatial modeling methods. (Majority of popular univariate spatial modeling strategies.)

Features of BLMC factor model and MCMC algorithm

- **Parameter expansion**: $\boldsymbol{\Lambda}$ and $\mathbf{F}$ are not jointly identified. However, it brings flexibility in model and efficiency in posterior sampling.
  - Conditional conjugacy contains geometry info to generate posterior samples. Allows all parameters in step 1 or 2 to be sampled simultaneously through a linear transformation of independent random variables.
  - Reduce the dependence between the updates in MCMC to improve the convergence rate and mixing of MCMC chains.
- **Posterior inference on high-dimensional space**
  - latent processes; missing response; prediction; parameter (identifiable)
- **Scalable for fitting massive spatial data with missing responses**

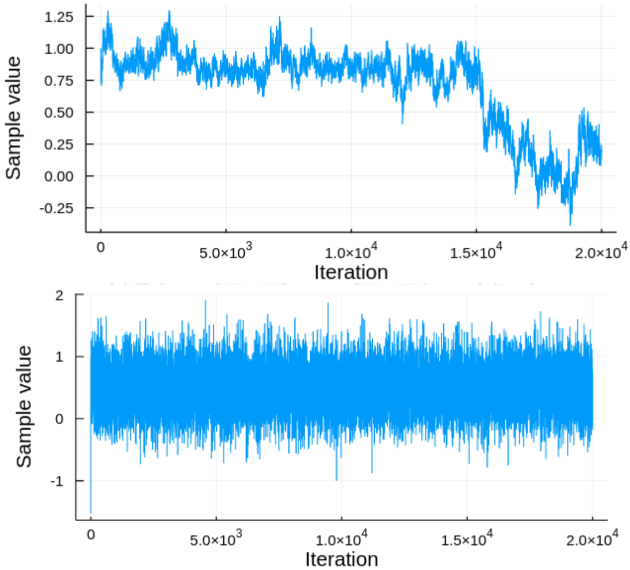


Figure: MCMC chain of $\mathbf{\Lambda}_{11}$ (1.0) and $w_1(\mathbf{s}_1) + \beta_{11}$ (0.453)

$$\mathrm{vec}(\mathbf{F}) \mid \cdots \sim \mathrm{N}((\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top\tilde{\mathbf{Y}}, (\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1})$$

$$(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})\,\mathrm{vec}(\mathbf{F}) = \tilde{\mathbf{X}}^\top(\tilde{\mathbf{Y}} + \boldsymbol{\eta})\,, \quad \boldsymbol{\eta} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_{nq})$$

$$\boldsymbol{\rho}_k = \{\rho_k(s_i, s_j\,; \psi_k)\}_{i,j=1}^n$$

A large family of popular scalale spatial models result in either sparse $\boldsymbol{\rho}_k$ or sparse $\boldsymbol{\rho}_k^{-1}$ that have sparse Cholesky decompositions

- When covariance matrix $\boldsymbol{\rho}_k$'s have special pattern, the direct calculation of $(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top(\tilde{\mathbf{Y}} + \boldsymbol{\eta})$ is often scalable.
- When precision matrics $\boldsymbol{\rho}_k^{-1}$ are **sparse**, i.e., the number of non-zero elements is linear to n.

$$\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} = \mathrm{diag}\{\boldsymbol{\rho}_k^{-1}\} + \text{sparse matrix}$$

We use iterative methods for solving linear system to facilitate the sampling of **F**

Conjugate gradient (CG)

CG is an iterative method for solving a linear system.

$$Ax = b, \; A \text{ non-singular, } n \times n \text{ matrix, solution } x^* = A^{-1}b$$

- Belongs to Krylov subspace methods.
- No matrix-matrix operations, Krylov subspace methods access matrices only through matrix-vector multiplies and work with the resulting vectors.
- More efficient for solving large sparse system
  - avoid storing matrix factors or even A
  - utilize the sparsity of A to reduce the storage and computational burden.
  - be significantly faster when A has certain features.

- **Lu Zhang** (2022)[1] wiley statref "Applications of Conjugate Gradient in Bayesian computation"
  - focus on sparse regression and spatial analysis
  - A self-contained introduction of conjugate gradient is provided to facilitate potential applications in a broader range of problems.

[1] **Lu, Zhang**. "Applications of Conjugate Gradient in Bayesian Computation". In: *Wiley StatsRef: Statistics Reference Online* (2022), pp. 1–7.

Illustration with Vege Indices data (I)[2]

- $q = 10$ Responses:
  - NDVI (Normalized Difference Vegetation Index): an index to identify vegetation amount and measure plant health vitality
  - Red Reflectance (waveband in red)
  - 8 other responses highly related to the distribution of vegetation.
- $p = 2$ Explanatory variables: intercept and indicator of no vegetation or urban area
- $n = 1,020,000$ observed locations
- Not all responses are recorded on every observed location (Heatmap)
  Has missing responses on over 65% of observed locations locations

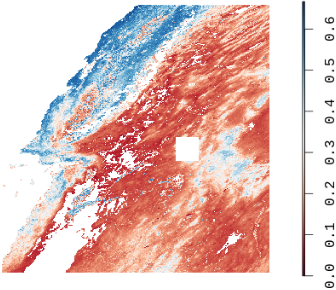


Figure: Colored transformed NDVI image of zone western United States

[2] **Lu, Zhang** and Sudipto Banerjee. "Spatial Factor Modeling: A Bayesian Matrix-Normal Approach for Misaligned Data". In: *Biometrics* (2021).

Illustration with Vege Indices data (II)

- Fit Bayesian LMC factor model with diagonal $\mathbf{\Sigma}$
- $K = 2$ Model $f_k(\mathbf{s})$ with Nearest Neighbor Gaussian Process
- 10,000 iterations 60.7 hours for a desktop with a single 8 Intel Core i7-7700K CPU @ 4.20GHz processor and 32 Gbytes of random-access memory running Ubuntu 18.04.2 LTS. locations

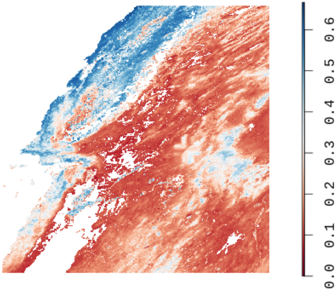


Figure: Interpolated maps of latent process + intercept for transformed NDVI

Conjugate (multivariate) spatial regression model[3]

Features: no MCMC, directly sample from posterior

Performance

18.88 mins to generate 500 independent posterior samples for a larger NDVI data with $N = 3,115,934$, $q = 2$ when fixing a small set of hyperparameters.

- Pragmatic way of obtaining quick inference in multivariate spatial data analysis, including the latent process
- More restricted model assumptions
  - Cannot fit data with misalignment
  - For Matérn model with measurement errors, we fix ϕ, $\delta^2 = \tau^2/\sigma^2$ and ν if unknown *No interval estimates of the fixed hyper-parameters*

[3] **Lu, Zhang**, Sudipto Banerjee, and Andrew O Finley. "High-dimensional multivariate geostatistics: A Bayesian matrix-normal approach". In: *Environmetrics* (2021), e2675.

Summary

- Multivariate spatial modeling
- Modeling approaches based separability and coregionalization
- Other related topics:
  - Multivariate spatial modeling approaches based upon moving averages and convolution
  - Multivariate models for areal data