Bayesian High-Dimensional Regression with Tensors and Distributed Computation with Space-Time Data

Rajarshi Guhaniyogi, Ph.D

Department of Statistics, Texas A & M University

Partially Supported by

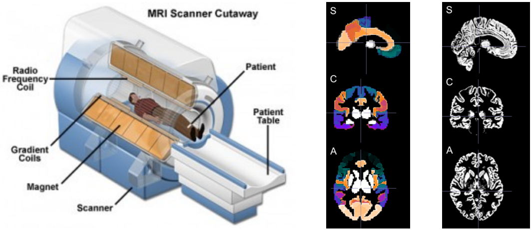
March 13, 2022

Outline

1. Bayesian tensor regression
2. Bayesian tensor response regression
3. Bayesian symmetric tensor response regression
4. Distributed computation with space-time data

Why Tensor Regression? Application in Primary Progressive Aphasia

Primary Progressive Aphasia is manifested in terms of language loss and indicates early stage of Alzheimers.

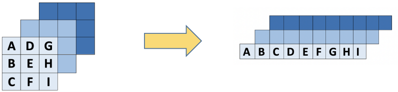


(a) MRI scan (b) Atlas (c) GM

Tensor predictor: Structural MRI for 142 patients of language loss.
scalar predictors: gender, age.
Response: Language score representing degree of language loss.

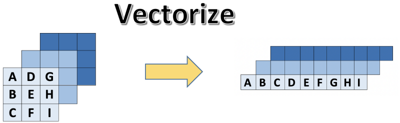
Penalized Optimization:Unsatisfactory Predictive Performance

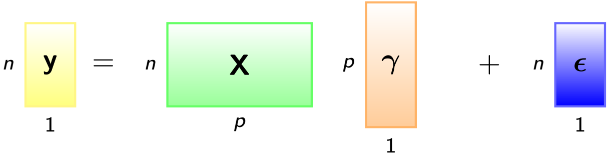
Vectorize



$$\underset{n \times 1}{\mathbf{y}} = \underset{n \times p}{\mathbf{X}} \; \underset{p \times 1}{\gamma} + \underset{n \times 1}{\epsilon}$$

Penalized Optimization:Unsatisfactory Predictive Performance



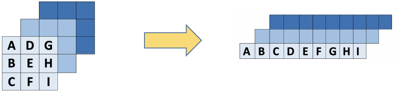


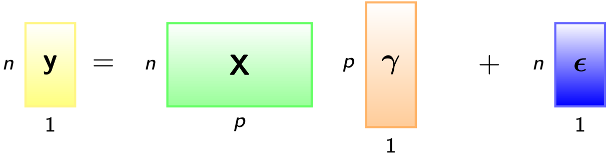
$\psi(\cdot)$ = convex penalty function, ζ = tuning parameter

$\arg\min_{\gamma} ||\boldsymbol{y} - \boldsymbol{X}\gamma||^2 + \zeta \sum_{j=1}^{p} \psi(\gamma_j) \rightarrow$ **Penalized Opt.**

Penalized Optimization:Unsatisfactory Predictive Performance

Vectorize



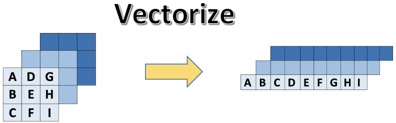


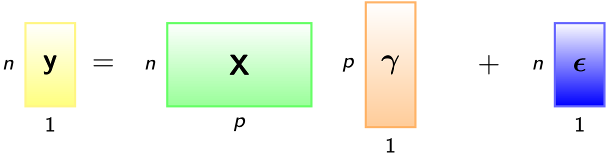
$\psi(\cdot)$ = convex penalty function, ζ = tuning parameter

$\arg\min_{\gamma} ||\boldsymbol{y} - \boldsymbol{X}\gamma||^2 + \zeta \sum_{j=1}^{p} \psi(\gamma_j) \rightarrow$ **Penalized Opt.**

- LASSO (Tibshirani, 1996), Elastic Net (Zhou et al., 2005), tons of other variants.

Penalized Optimization:Unsatisfactory Predictive Performance





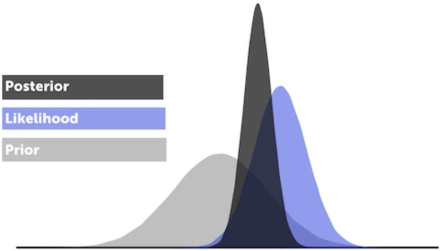
$\psi(\cdot)$ = convex penalty function, ζ = tuning parameter

$\arg\min_{\gamma} ||\boldsymbol{y} - \boldsymbol{X}\gamma||^2 + \zeta \sum_{j=1}^{p} \psi(\gamma_j) \rightarrow$ **Penalized Opt.**

- LASSO (Tibshirani, 1996), Elastic Net (Zhou et al., 2005), tons of other variants.
- Unsatisfactory predictive uncertainty.

Bayesian Inference

- Start with a prior distribution of γ.
- "Combine" data likelihood and prior distribution to obtain posterior distribution of γ.



- **point estimation** $\rightarrow$ mean of the posterior, **uncertainty** $\rightarrow$ 95% credible interval from the posterior.
- Markov Chain Monte Carlo (MCMC) and its variants exist to empirically estimate the posterior distribution of γ.

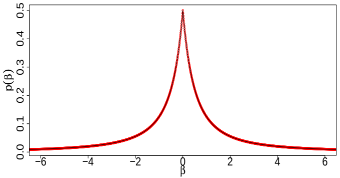
Bayesian High Dim. Reg.:Unsuitable in High Dimension

- Bayesians choose sparsity-favoring priors on $\gamma \in \mathbb{R}^P$ which will set components of γ to be 0.

- Bayesians choose sparsity-favoring priors on $\gamma \in \mathbb{R}^P$ which will set components of γ to be 0.

Spike & Slab Prior (Computationally Inefficient)

$$\gamma_j \sim \pi\delta_0 + (1-\pi)g, \;\; g \text{ is a cont. density.}$$



Bayesian Shrinkage Prior (Statistically Inefficient)

$$\gamma_j \sim N(0, \zeta_j\tau), \; \zeta_j \sim f_1, \; \tau \sim f_2.$$

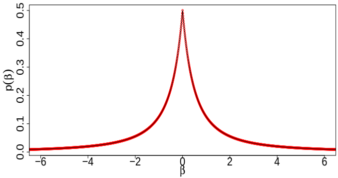
Marginally, γ_j has a heavy tailed density

Bayesian High Dim. Reg.:Unsuitable in High Dimension

- Bayesians choose sparsity-favoring priors on $\gamma \in \mathbb{R}^P$ which will set components of γ to be 0.

Spike & Slab Prior (Computationally Inefficient)

$$\gamma_j \sim \pi\delta_0 + (1-\pi)g, \;\; g \text{ is a cont. density.}$$



Bayesian Shrinkage Prior (Statistically Inefficient)

$$\gamma_j \sim N(0, \zeta_j\tau), \; \zeta_j \sim f_1, \; \tau \sim f_2.$$

Marginally, γ_j has a heavy tailed density

- Important shrinkage priors, Bayesian Lasso (Park et al., 2008; Hans, 2009), Horseshoe (Carvalho et al., 2009), Generalized Double Pareto (Armagan et al., 2013).

Bayesian High Dim. Reg.:Unsuitable in High Dimension

- Bayesians choose sparsity-favoring priors on $\gamma \in \mathbb{R}^P$ which will set components of γ to be 0.

Spike & Slab Prior (Computationally Inefficient)

$$\gamma_j \sim \pi\delta_0 + (1-\pi)g, \;\; g \text{ is a cont. density.}$$

Serious Drawbacks of Penalization and Shrinkage

- $p = p_1 \times p_2 \times p_3$, each $p_i = 64$ typically, implies massive dimensional regression with close to half a million predictors $\Rightarrow$ Infeasibility
- Misses out on wealth of information that the tensor valued images carry.

p(β)

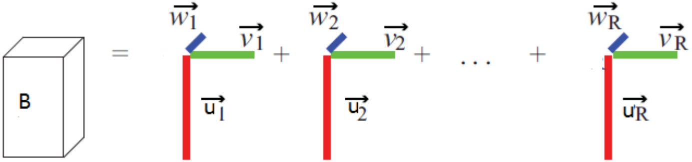
- Important shrinkage priors, Bayesian Lasso (Park et al., 2008; Hans, 2009), Horseshoe (Carvalho et al., 2009), Generalized Double Pareto (Armagan et al., 2013).

Tensor Regression Model with PARAFAC Decomposition

Data Model

$$y = \langle \boldsymbol{X}, \boldsymbol{B} \rangle + \boldsymbol{z}'\gamma + \epsilon, \epsilon \sim \mathrm{N}(0, \sigma^2)$$

rank-R PARAFAC decomposition of B for dimension reduction



For $D > 3$, need a better notation $\Rightarrow \boldsymbol{B} = \sum_{r=1}^{R} \beta_1^{(r)} \circ \cdots \circ \beta_D^{(r)}$
$\beta_j^{(r)} \in \mathbb{R}^{p_j}$, $\circ$ denotes *outer product* between vectors.

Tensor Regression Model with PARAFAC Decomposition

Data Model

$$y = \langle \boldsymbol{X}, \boldsymbol{B} \rangle + \boldsymbol{z}'\gamma + \epsilon, \epsilon \sim \mathrm{N}(0, \sigma^2)$$

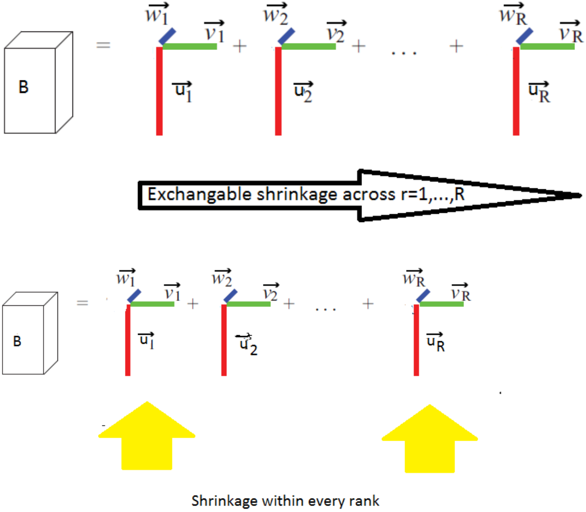
rank-R PARAFAC decomposition of B for dimension reduction

Advantages

- Number of parameters needed to model is $R \sum_{j=1}^{D} p_j$ as opposed to $\prod_{j=1}^{D} p_j \Rightarrow$ Dimension Reduction.
- Exploits neighborhood structure of $\boldsymbol{X} \Rightarrow$ potentially better inference.

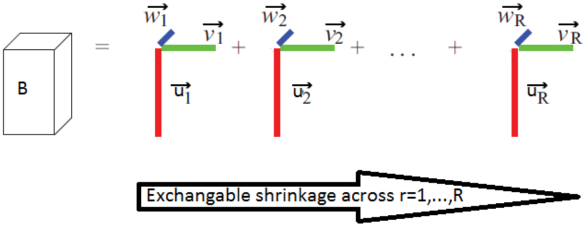
For $D > 3$, need a better notation $\Rightarrow \boldsymbol{B} = \sum_{r=1}^{R} \beta_1^{(r)} \circ \cdots \circ \beta_D^{(r)}$
$\beta_j^{(r)} \in \mathbb{R}^{p_j}$, $\circ$ denotes *outer product* between vectors.

Need for a Multiway Shrinkage Prior on B

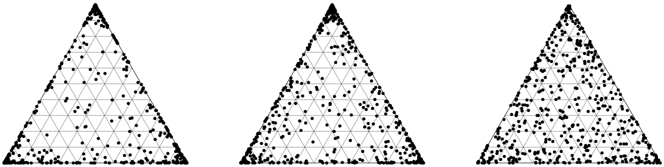


- Protects from overfitting due to a higher rank (R) than needed.
- Estimate tensor margins with an approximate sparsity.

Multiway Shrinkage Prior for B (G. et al. 2017, JMLR)

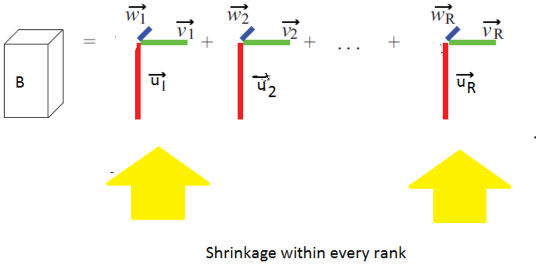


$\beta_j^{(r)} \sim N(\mathbf{0}, diag(w_{jr,1}, .., w_{jr,p_j})\tau\phi_r)$, ϕ_r's rank specific parameters.
Shrinkage across ranks: $(\phi_1, ..., \phi_R) \sim Dirichlet(\alpha, ..., \alpha)$, $\alpha > 0$.



(a) $\alpha = 0.2$ (b) $\alpha = 0.3$ (c) $\alpha = 0.5$

Multiway Dirichlet Generalized Double Pareto Prior (M-DGDP)

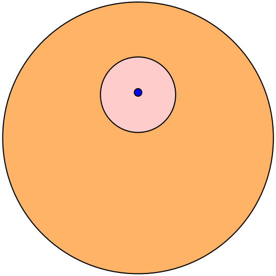


$$w_{jr,k} \sim \mathrm{Exp}(\lambda_{jr}^2/2), \quad \lambda_{jr} \sim \mathrm{Ga}(a_\lambda, b_\lambda), \tau \sim IG(a_\tau, b_\tau)$$

Integrating out $\boldsymbol{W}_{jr}$

$\beta_{j,k}^{(r)} \mid \phi_r, \tau$ marginally follows GDP shrinkage prior.

General Theoretical Setup: G. et al., 2017, JMLR



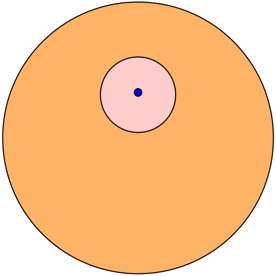
True Model

$\left(f(y|\boldsymbol{B}_n^0) = \mathrm{N}\left(\langle \boldsymbol{X}, \boldsymbol{B}_n^0\rangle, \sigma^2\right)\right)$

Class of tensor reg. models fitted to the data

KL metric ball of radius ϵ around the truth

General Theoretical Setup: G. et al., 2017, JMLR



True Model
$(f(y|\boldsymbol{B}_n^0) = \mathrm{N}\left(\langle \boldsymbol{X}, \boldsymbol{B}_n^0\rangle, \sigma^2\right))$

Class of tensor reg. models fitted to the data

KL metric ball of radius ϵ around the truth

$$\boxed{\mathscr{B}_n = \left\{\boldsymbol{B}_n : \frac{1}{n}\sum_{i=1}^{n} \mathrm{KL}(f(y_i|\boldsymbol{B}_n^0), f(y_i|\boldsymbol{B}_n)) < \epsilon\right\} \Rightarrow \textit{Neighborhood}}$$

Posterior Consistency

$$\Pi_n\left(\mathscr{B}_n^c\right) \to 0 \quad \text{under } \boldsymbol{B}_n^0 \quad \text{a.s. as } n \to \infty. \tag{1}$$

Π_n posterior distribution given $y_1, \ldots, y_n$.

Posterior Consistency Results, G. et al. 2017, JMLR

Theorem

The posterior is consistent under the following assumptions.

Posterior Consistency Results, G. et al. 2017, JMLR

Theorem

The posterior is consistent under the following assumptions.

1. $\boldsymbol{B}_n^0 = \sum_{r=1}^{R^0} \beta_{1,n}^{0(r)} \circ \cdots \circ \beta_{D,n}^{0(r)}$ follows rank-R_0 decomposition, $R > R_0$. (Structure on the true coefficients)

Theorem

The posterior is consistent under the following assumptions.

1. $\boldsymbol{B}_n^0 = \sum_{r=1}^{R^0} \boldsymbol{\beta}_{1,n}^{0(r)} \circ \cdots \circ \boldsymbol{\beta}_{D,n}^{0(r)}$ follows rank-R_0 decomposition, $R > R_0$. (Structure on the true coefficients)
2. $\sup_{l=1,\dots,p_{j,n}} |\boldsymbol{\beta}_{j,n,l}^{0(r)}| < \infty$, for all $j = 1, \dots, D$; $r = 1, \dots, R_0$. (Structure on the true coefficients)

Posterior Consistency Results, G. et al. 2017, JMLR

Theorem

The posterior is consistent under the following assumptions.

1. $\boldsymbol{B}_n^0 = \sum_{r=1}^{R^0} \beta_{1,n}^{0(r)} \circ \cdots \circ \beta_{D,n}^{0(r)}$ follows rank-R_0 decomposition, $R > R_0$. (Structure on the true coefficients)
2. $\sup_{l=1,\ldots,p_{j,n}} |\beta_{j,n,l}^{0(r)}| < \infty$, for all $j = 1, \ldots, D;\ r = 1, \ldots, R_0$. (Structure on the true coefficients)
3. $\sum_{j=1}^{D} p_{j,n} \log(p_{j,n}) = o(n)$. (Dimensions of tensor margins)

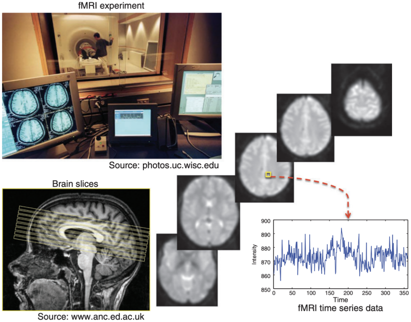
Posterior Consistency Results, G. et al. 2017, JMLR

Theorem

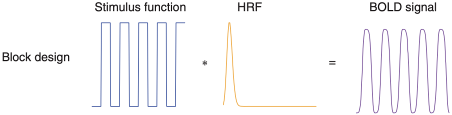
The posterior is consistent under the following assumptions.

1. $\boldsymbol{B}_n^0 = \sum_{r=1}^{R^0} \beta_{1,n}^{0(r)} \circ \cdots \circ \beta_{D,n}^{0(r)}$ follows rank-R_0 decomposition, $R > R_0$. (Structure on the true coefficients)
2. $\sup_{l=1,\ldots,p_{j,n}} |\beta_{j,n,l}^{0(r)}| < \infty$, for all $j = 1, \ldots, D;\ r = 1, \ldots, R_0$. (Structure on the true coefficients)
3. $\sum_{j=1}^{D} p_{j,n} \log(p_{j,n}) = o(n)$. (Dimensions of tensor margins)

More involved convergence results can be found in G. (2017), JMVA.



- $\boldsymbol{Y}_t$ is $64 \times 64 \times 64$ dimensional tensor response at time $t = 1, ..., T$.



- x_t is a scalar activation related predictor at time $t = 1, ..., T$.

- Identify brain voxels activated by an external stimulus.

Tensor Response Regression Model

Data Model

$$\boldsymbol{Y}_t = \boldsymbol{B}_1 x_{1t} + \cdots + \boldsymbol{B}_m x_{mt} + \boldsymbol{E}_t$$

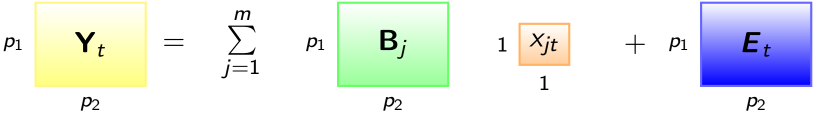
- $\boldsymbol{Y}_t$ is a $p_1 \times \cdots \times p_D$ dimensional tensor response, $x_{1,t}, \ldots, x_{m,t}$ are m predictors.
- $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_m$ are $p_1 \times \cdots \times p_D$ dim. tensor coefficients.
- $\mathrm{vec}(\boldsymbol{E}_t) \sim$ Stationary AR(1) process with the lag parameter ϕ.

Tensor Response Regression Model

Data Model

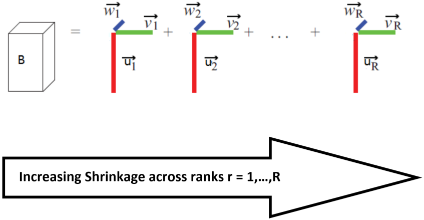
$$\boldsymbol{Y}_t = \boldsymbol{B}_1 x_{1t} + \cdots + \boldsymbol{B}_m x_{mt} + \boldsymbol{E}_t$$

- $\boldsymbol{Y}_t$ is a $p_1 \times \cdots \times p_D$ dimensional tensor response, $x_{1,t}, \ldots, x_{m,t}$ are m predictors.
- $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_m$ are $p_1 \times \cdots \times p_D$ dim. tensor coefficients.
- $\text{vec}(\boldsymbol{E}_t) \sim$ Stationary AR(1) process with the lag parameter ϕ.



- (k, l)-th entry of $\boldsymbol{B}_j$ determines the effect of j-th predictor on the (k, l)-th cell of the response tensor.

Multiway Stick Breaking Prior for $\boldsymbol{B}_j$ (Spencer et al., Psychometrika (2020); G. & Spencer, Bayesian Analysis (2021))



- $\boldsymbol{B}_j = \sum_{r=1}^{R} \boldsymbol{\beta}_{j,1}^{(r)} \circ \cdots \circ \boldsymbol{\beta}_{j,D}^{(r)}$

- Shrinkage within every rank through generalized double pareto shrinkage prior.
- Shrinkage prior involves rank specific parameters $\phi_{j,r}$, $r = 1, ..., R$.
- They assume a stick breaking construction $\phi_{j,1} = \xi_{j,1}, \phi_{j,2} = \xi_{j,2}(1 - \xi_{j,1}), ..., \phi_{j,R} = \prod_{r=1}^{R-1}(1 - \xi_{j,r})$.

Theoretical Study: Bayesian Tensor Response Regression

$\boldsymbol{B}$ (tensor of dimensions $m \times p_1 \times \cdots \times p_D$): stacking tensor coefficients $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_m$ together.

$$\boxed{\mathscr{A}_T = \{\boldsymbol{B} : ||\boldsymbol{B} - \boldsymbol{B}_0||_2 < \epsilon\}}, \ \boldsymbol{B}_0 = \text{true value of } \boldsymbol{B}.$$

$\Pi_T(\cdot)$ is the posterior distribution of $\boldsymbol{B}$ with T observations.

Notion of Posterior Consistency

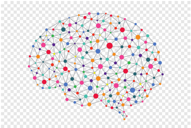
$\Pi_T(\mathscr{A}_T^c) \to 0$, a.s., when $T \to \infty$.

Posterior consistency holds under the following conditions:

1. $\boldsymbol{B}_{0,j}$ assumes rank $R_{0,j}$ PARAFAC decomposition with $R > \max_j R_{0,j}$.
2. $m \sum_{d=1}^{D} p_d \log(p_d) = o(T)$, $s \log(m \prod_{d=1}^{D} p_d) = o(T)$, where s is the number of nonzero entries in $\boldsymbol{B}_0$.
3. Covariate matrix has bounded singular values.

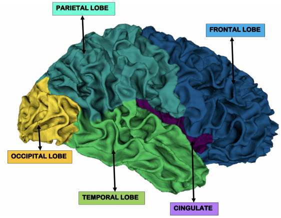
Motivation: Brain Connectomes with Phenotypes

- **Data:** Brain connectome network ($\boldsymbol{Y}_i$), creative achievement (x_i) for subjects.
- (k, l)-th entry of $\boldsymbol{Y}_i$ represents "association" between kth and lth brain regions.



Nodes

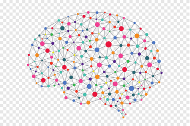
Nodes	1	2	3	4	5
1	0	a	b	c	d
2		0	e	f	g
3			0	h	i
4				0	j
5					0



- 68 Regions of interest (ROI), 34 in each hemisphere.
- 12 Lobes, 6 in each hemisphere.

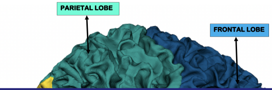
Motivation: Brain Connectomes with Phenotypes

- **Data:** Brain connectome network ($\boldsymbol{Y}_i$), creative achievement (x_i) for subjects.
- (k, l)-th entry of $\boldsymbol{Y}_i$ represents "association" between kth and lth brain regions.



Nodes

Nodes	1	2	3	4	5
1	0	a	b	c	d
2		0	e	f	g
3			0	h	i
4				0	j
5					0



- 68 Regions of interest (ROI), 34 in each

Inferential Goal

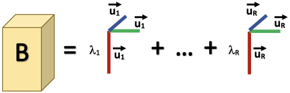
Develop regression of $\boldsymbol{Y}_i$ on x_i, identify important network nodes related to creativity.

Bayesian Symmetric Tensor on Vector Regression (Guha & G., Technometrics, 2021)

Data Model

$$\boldsymbol{Y}_i = \boldsymbol{B}_1 x_{1i} + \cdots + \boldsymbol{B}_m x_{mi} + \boldsymbol{E}_i$$

- $\boldsymbol{Y}_i$ is a $p \times \cdots \times p$ dimensional *symmetric* tensor response, $x_{1i}, \ldots, x_{mi}$ are m predictors.
- $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_m$ are $p \times \cdots \times p$ dim. symmetric tensor coefficients.
- $\boldsymbol{B}_j$ follows a symmetric rank-R PARAFAC decomposition $\boldsymbol{B}_j = \sum_{r=1}^{R} \lambda_{j,r} \boldsymbol{\beta}_j^{(r)} \circ \cdots \circ \boldsymbol{\beta}_j^{(r)}$.



- $\boldsymbol{\beta}_j^{(r)} = (\beta_{j,1}^{(r)}, \ldots, \beta_{j,p}^{(r)})' \in \mathbb{R}^p$, $\lambda_{j,r} \in \{0, 1\}$.

Influential Node Identification

- $\boldsymbol{u}_{j,k} = (\beta_{j,k}^{(1)}, ..., \beta_{j,k}^{(R)})' = \boldsymbol{0}$ implies k-th node is unrelated to the jth predictor.
- Variable selection prior to identify important nodes,

$$\boldsymbol{u}_{j,k} \sim \begin{cases} N(\boldsymbol{0}, \boldsymbol{M}), & \text{if } \xi_{j,k} = 1 \\ \delta_{\boldsymbol{0}}, & \text{if } \xi_{j,k} = 0 \end{cases}, \quad \xi_{j,k} \sim Ber(\Delta),$$

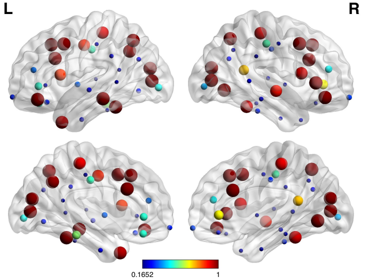
where $\delta_{\boldsymbol{0}}$ is the Dirac-delta function at $\boldsymbol{0}$, $\boldsymbol{M}$ is a covariance matrix of order $R \times R$.

Near Optimal Estimation of Predictive Density (Guha and G., 2021)

The predictive density of the proposed model can be estimated at a rate close to $n^{-1/2}$ upto a $\log(n)$ factor.

Inference on Significant Nodes

- Compute posterior probability of $\{\boldsymbol{u}_{j,k} = \mathbf{0}\}$ empirically from MCMC samples.
- kth network node is related to the jth predictor if this probability is less than 0.5.



13 frontal, **6** temporal ROIs are significant among **34** significant ROIs. (More than half of the identified ROIs).

- We offer posterior probabilities of each node being related to a predictor, which quantifies the statistical uncertainty.
- Posterior prob. close to 0 or 1 means less uncertainty with the decision.

Functional Regression: Computational Infeasibility

Space-Time Varying Coefficient Model

$$y(\boldsymbol{s}_i, t_i) = \boldsymbol{x}(\boldsymbol{s}_i, t_i)' \boldsymbol{\beta} + \boldsymbol{z}(\boldsymbol{s}_i, t_i)' \gamma(\boldsymbol{s}_i, t_i) + \epsilon(\boldsymbol{s}_i, t_i)$$

- $p \times 1$ Fixed Effect
- $m \times 1$ Space-Time Varying Coefficients
- Non-Spatial Error following i.i.d. $N(0, \tau^2)$

Some Observations

- Only m of the p predictors have varying coefficients, $m \leq p$.
- $\boldsymbol{z}(\boldsymbol{s}_i, t_i) = 1 \Rightarrow$ spatio-temporal geo-statistical model.
- $\boldsymbol{x}(\boldsymbol{s}_i, t_i) = \boldsymbol{z}(\boldsymbol{s}_i, t_i) \Rightarrow$ all predictor coefficients are space-time varying.

Spatio-Temporal Gaussian Process

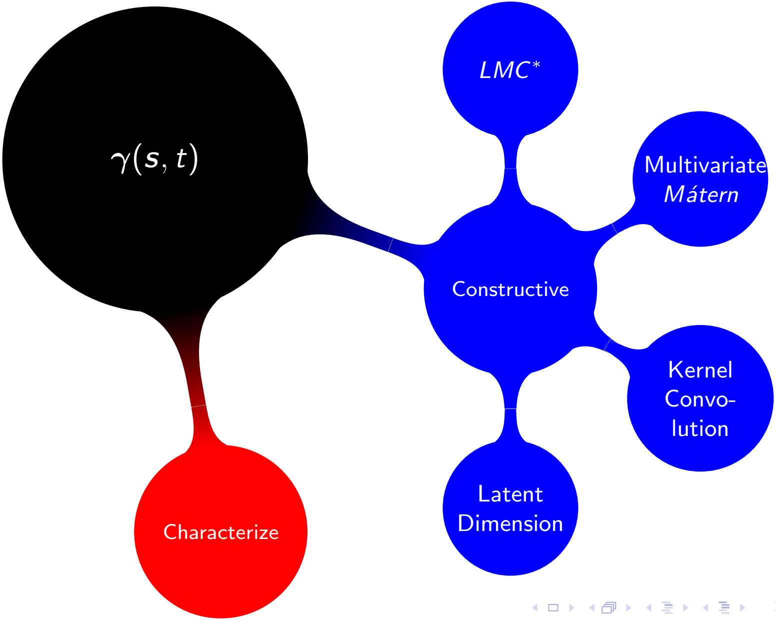
$\{\gamma(\boldsymbol{s}, t) : (\boldsymbol{s}, t) \in \mathscr{D} \times \mathscr{T}\} \sim GP(\boldsymbol{0}, \boldsymbol{C}_{\boldsymbol{\theta}}(\cdot, \cdot))$ → $(\gamma(\boldsymbol{s}_1, t_1), \ldots, \gamma(\boldsymbol{s}_n, t_n))' \sim N(\boldsymbol{0}, \boldsymbol{C}_{\boldsymbol{\theta}})$

for any finite set of location-time tuples $(\boldsymbol{s}_1, t_1), \ldots, (\boldsymbol{s}_n, t_n)$.

Cross Covariance Kernel Matrix

- $\boldsymbol{C}_{\boldsymbol{\theta}}(\cdot, \cdot)$ is the $m \times m$ cross-covariance kernel matrix.
- $\boldsymbol{C}_{\boldsymbol{\theta}}$ is the $nm \times nm$ covariance matrix with the (i, j)-th block given by the $m \times m$ matrix $\boldsymbol{C}_{\boldsymbol{\theta}}((\boldsymbol{s}_i, t_i), (\boldsymbol{s}_j, t_j))$.

Modeling Cross Covariance:Popular Approaches



Full Likelihood from Gaussian Process (GP) Model

$\boldsymbol{y} = (y(\boldsymbol{s}_1, t_1), .., y(\boldsymbol{s}_n, t_n))'$, ($n \times 1$ vector)
$\boldsymbol{X} = [\boldsymbol{x}(\boldsymbol{s}_1, t_1) : \cdots : \boldsymbol{x}(\boldsymbol{s}_n, t_n)]'$, ($n \times p$ matrix)
$\boldsymbol{Z} = Block - diag(\boldsymbol{z}(\boldsymbol{s}_1, t_1)', \cdots, \boldsymbol{z}(\boldsymbol{s}_n, t_n)')$, ($n \times nm$ matrix)

- Model: $\boldsymbol{y} \sim N(\boldsymbol{X}\beta, \boldsymbol{Z}\boldsymbol{C}_{\boldsymbol{\theta}}\boldsymbol{Z}' + \tau^2\boldsymbol{I})$.
- Estimating parameters $\beta, \boldsymbol{\theta}, \tau^2$ from the likelihood

$$-\frac{\log(det(\boldsymbol{Z}\boldsymbol{C}_{\boldsymbol{\theta}}\boldsymbol{Z}' + \tau^2\boldsymbol{I}))}{2} - \frac{(\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{Z}\boldsymbol{C}_{\boldsymbol{\theta}}\boldsymbol{Z}' + \tau^2\boldsymbol{I})^{-1}(\boldsymbol{y} - \boldsymbol{X}\beta)}{2}$$

Challenges

- Store $\boldsymbol{Z}\boldsymbol{C}_{\boldsymbol{\theta}}\boldsymbol{Z}' + \tau^2\boldsymbol{I}$
- Compute $\texttt{Chol}(\boldsymbol{Z}\boldsymbol{C}_{\boldsymbol{\theta}}\boldsymbol{Z}' + \tau^2\boldsymbol{I}) = \boldsymbol{L}\boldsymbol{L}'$.

n^3 floating point operations per MCMC iteration $\rightarrow$Big-n problem

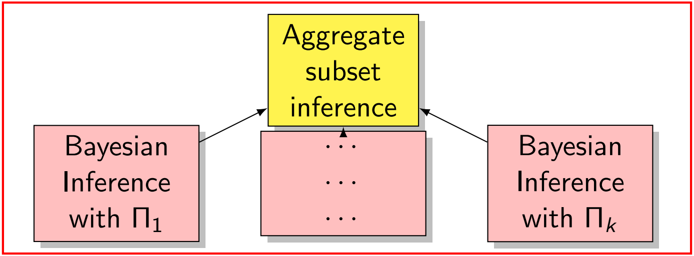
Literature on Spatial/Spatio-Temporal Big Data

- Low rank model (Wahba, 1990; Higdon, 2001; Kamman & Wand, 2003; Paciorek, 2007; Lemos and Sanso, 2006; Banerjee et al., 2008; Cressie & Johannesson, 2008; Finley et al., 2009; Gramacy and Lee, 2008; **Guhaniyogi et al., 2011 & 2013**; Sang et al. 2012; Katzfuss, 2016).
- Multiscale approaches (Nychka, 2002; Johannesson et al., 2007; Tzeng and Huang, 2015; Nychka et al., 2015; Katzfuss, 2016; Katzfuss & Guinness, 2021, **Guhaniyogi & Sanso, 2017**).
- Spectral approximations and composite likelihoods (Fuentes, 2007; Eidvisk, 2016).
- Sparsity: Covariance tapering (Kaufman et al., 2008; Du et al., 2009; Sang et al., 2012; **Guhaniyogi, 2017**),`INLA` (Rue et al., 2009; Lindgren et al., 2011), `lagp` (Gramacy and Apley, 2015), nearest neighbor processes (Stein et al., 2004; Stroud et al., 2014; Datta et al., 2016).

Divide-and-Conquer Inference with Big Data

- Split the data $\mathscr{S} = \{(\boldsymbol{s}_1, t_1), ..., (\boldsymbol{s}_n, t_n)\}, \mathscr{Y} = \{y(\boldsymbol{s}_1, t_1), ..., y(\boldsymbol{s}_n, t_n)\}, \mathscr{X} = \{\boldsymbol{x}(\boldsymbol{s}_1, t_1), ..., \boldsymbol{x}(\boldsymbol{s}_n, t_n)\}$, $\mathscr{Z} = \{\boldsymbol{z}(\boldsymbol{s}_1, t_1), ..., \boldsymbol{z}(\boldsymbol{s}_n, t_n)\}$ into k exhaustive subsets $\mathscr{S}_j, \mathscr{Y}_j, \mathscr{X}_j, \mathscr{Z}_j, j = 1, .., k$.

- The jth subset posterior Π_j computed from the j-th subset containing M_j data points drawn randomly from the entire domain, $M_1 + \cdots + M_k \geq n$.



Construction of Subset Posteriors

Subset Posterior: "Weak Learner" of Full Posterior

$$\Pi_j(\boldsymbol{\beta}, \boldsymbol{\theta}, \tau^2 | \mathscr{Y}_j, \mathscr{X}_j, \mathscr{Z}_j) \propto [\, p(\mathscr{Y}_j | \mathscr{X}_j, \mathscr{Z}_j, \boldsymbol{\beta}, \boldsymbol{\theta}, \tau^2)\,]^{n/M_j}\; p(\boldsymbol{\beta}, \boldsymbol{\theta}, \tau^2)$$

- Likelihood: $N(\boldsymbol{y}_j | \boldsymbol{X}_j \beta, \boldsymbol{Z}_j \boldsymbol{C}_{\boldsymbol{\theta},j} \boldsymbol{Z}_j' + \tau^2 \boldsymbol{I}_{M_j})$
- Prior Distribution

- These are "Stochastic Approximations" of the full posterior $\Pi(\boldsymbol{\beta}, \boldsymbol{\theta}, \tau^2 | \mathscr{Y}, \mathscr{X}, \mathscr{Z})$.
- For easier implementation you may get rid of the power $n/M_j \Rightarrow$ satisfactory point estimation, wider confidence intervals.

How to combine Π_j's optimally?

Combine Subset Posteriors Marginally: Li et al. (2017), Biometrika; G. et al. (2022), Stat. Sci.

- Compute Wasserstein mean $\bar{\Pi}$ of $\Pi_1, \ldots, \Pi_k$.

Each Π_j multivariate normal $\implies$ $\bar{\Pi}$ multivariate normal

- $h(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}$, a $1D$ parametric function of $(\boldsymbol{\theta}, \boldsymbol{\beta})$.

- $\Pi_j^{-1}(u)$: uth quantile of the jth subset posterior distribution of $h(\boldsymbol{\theta}, \boldsymbol{\beta})$, $u \in (0, 1)$.

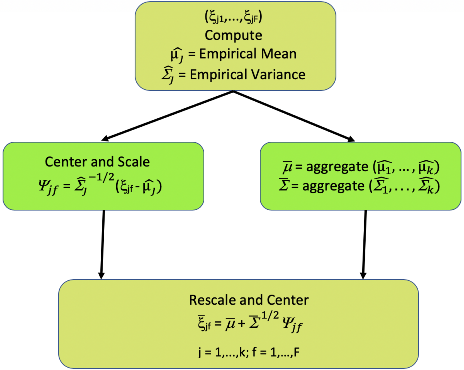
- If $\bar{\Pi}^{-1}(u)$ is the uth quantile of the Wasserstein mean, then

PIE Combination: $\bar{\Pi}^{-1}(u) = \frac{1}{k} \sum_{j=1}^{k} \Pi_j^{-1}(u), \ \forall \ u \in (0, 1)$

- Combines marginals of posterior distribution separately.

General Subset Posterior Aggregation Approach

- ξ_{jf}: f-th post burn-in iterate of model parameters from the j-th subset.



Important Special Cases

Aggregated Monte Carlo (AMC) (G. et al., 2022, JMLR)

$$\bar{\boldsymbol{\mu}} = \frac{1}{k} \sum_{j=1}^{k} \widehat{\boldsymbol{\mu}}_j, \;\; \bar{\boldsymbol{\Sigma}} = \mathrm{AM}\{\widehat{\boldsymbol{\Sigma}}_1, \ldots, \widehat{\boldsymbol{\Sigma}}_k\}$$

Wasserstein Posterior (WASP) (G. et al., 2022, JMLR)

$$\bar{\boldsymbol{\mu}} = \frac{1}{k} \sum_{j=1}^{k} \widehat{\boldsymbol{\mu}}_j, \;\; \bar{\boldsymbol{\Sigma}} = \mathrm{GM}\{\widehat{\boldsymbol{\Sigma}}_1, \ldots, \widehat{\boldsymbol{\Sigma}}_k\}$$

- WASP uses geometric mean for aggregating $\widehat{\boldsymbol{\Sigma}}_j$'s.
- Geometric mean requires computing an iterative algorithm.

Novelty vis-a-vis Existing Divide & Conquer Techniques

- Aggregation of Subset Posteriors through Median (Minsker et al., 2017)
- Computation of Meta Posterior (Guhaniyogi and Banerjee, 2017)
- Consensus Monte Carlo (CMC) (Scott et al., 2016), Semiparametric Density Product (SDP) (Neiswenger et al., 2014).
- ADVI (Kucukelbir et al., 2017).
- Both theory and practice for uncertainty quantification of parameters are unavailable for spatial/spatio-temporal process models.

Convergence Rate Results, G. et al. 2022, JMLR

For notational simplicity, we denote $\boldsymbol{u} = (\boldsymbol{s}, t)$, and assume $M_1 = \cdots = M_k = M = n/k$.

True Model ($\mathscr{M}_0$) and Fitted Model ($\mathscr{M}$)

$\mathscr{M}$: $y(\boldsymbol{u}) = z(\boldsymbol{u})'\boldsymbol{\gamma}(\boldsymbol{u}) + \epsilon(\boldsymbol{u})$, $\boldsymbol{\gamma}(\boldsymbol{u}) = (\gamma_1(\boldsymbol{u}), ..., \gamma_m(\boldsymbol{u}))'$

$\mathscr{M}_0$: $y(\boldsymbol{u}) = z(\boldsymbol{u})'\boldsymbol{\gamma}_0(\boldsymbol{u}) + \epsilon(\boldsymbol{u})$, $\boldsymbol{\gamma}_0(\boldsymbol{u}) = (\gamma_{0,1}(\boldsymbol{u}), ..., \gamma_{0,m}(\boldsymbol{u}))'$

1. The cross-covariance function for $\boldsymbol{\gamma}(\boldsymbol{u})$ has bounded eigenfunctions and polynomially decaying eigenvalues.

Convergence Rate Results, G. et al. 2022, JMLR

For notational simplicity, we denote $\boldsymbol{u} = (\boldsymbol{s}, t)$, and assume $M_1 = \cdots = M_k = M = n/k$.

True Model ($\mathscr{M}_0$) and Fitted Model ($\mathscr{M}$)

$\mathscr{M}: y(\boldsymbol{u}) = z(\boldsymbol{u})'\gamma(\boldsymbol{u}) + \epsilon(\boldsymbol{u}),\ \gamma(\boldsymbol{u}) = (\gamma_1(\boldsymbol{u}), ..., \gamma_m(\boldsymbol{u}))'$

$\mathscr{M}_0: y(\boldsymbol{u}) = z(\boldsymbol{u})'\gamma_0(\boldsymbol{u}) + \epsilon(\boldsymbol{u}),\ \gamma_0(\boldsymbol{u}) = (\gamma_{0,1}(\boldsymbol{u}), ..., \gamma_{0,m}(\boldsymbol{u}))'$

1. The cross-covariance function for $\gamma(\boldsymbol{u})$ has bounded eigenfunctions and polynomially decaying eigenvalues.
2. The function $\gamma_{0,g}(\boldsymbol{u})$ has v degrees of smoothness.
3. subsets are disjoint and subset size M must be greater than a certain fraction of n depending on v.

Convergence Rate Results, G. et al. 2022, JMLR

For notational simplicity, we denote $\boldsymbol{u} = (\boldsymbol{s}, t)$, and assume $M_1 = \cdots = M_k = M = n/k$.

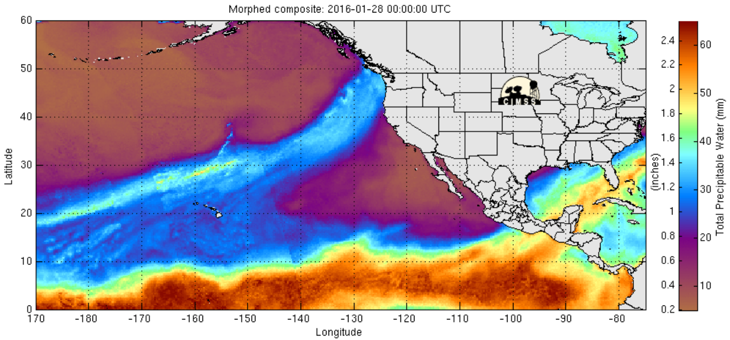
True Model ($\mathscr{M}_0$) and Fitted Model ($\mathscr{M}$)

$\mathscr{M}$: $y(\boldsymbol{u}) = z(\boldsymbol{u})'\boldsymbol{\gamma}(\boldsymbol{u}) + \epsilon(\boldsymbol{u})$, $\boldsymbol{\gamma}(\boldsymbol{u}) = (\gamma_1(\boldsymbol{u}), \ldots, \gamma_m(\boldsymbol{u}))'$
$\mathscr{M}_0$: $y(\boldsymbol{u}) = z(\boldsymbol{u})'\gamma_0(\boldsymbol{u}) + \epsilon(\boldsymbol{u})$, $\gamma_0(\boldsymbol{u}) = (\gamma_{0,1}(\boldsymbol{u}), \ldots, \gamma_{0,m}(\boldsymbol{u}))'$

1. The cross-covariance function for $\boldsymbol{\gamma}(\boldsymbol{u})$ has bounded eigenfunctions and polynomially decaying eigenvalues.
2. The function $\gamma_{0,g}(\boldsymbol{u})$ has v degrees of smoothness.
3. subsets are disjoint and subset size M must be greater than a certain fraction of n depending on v.
4. Subset posterior aggregation schemes follow a general rule satisfied by PIE, AMC and WASP.

Then $\gamma_{0,g}$ is estimated close to the optimal rate of $n^{-2v/(2v+3)}$.

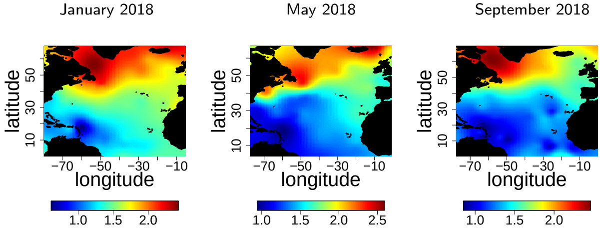
Sea Surface Temperature (SST) and Sea Surface Salinity (SSS)



- Data from Hadley Center of the MET office in UK.
- Data on SST and SSS between $0^0 - 70^0$ N. latitude and $0^0 - 80^0$ W. longitude.
- We consider $\sim 110K$ space-time observations on SST and SSS over the 12 months in 2018.

Fit: $\mathrm{SST}(\boldsymbol{s}, t) = \gamma_0(\boldsymbol{s}, t) + \gamma_1(\boldsymbol{s}, t)\ \mathrm{SSS}(\boldsymbol{s}, t) + \epsilon(\boldsymbol{s}, t)$

Estimated Map of $\gamma_1(\boldsymbol{s}, t)$ in the North Atlantic



1. Divide data into 400 subsets.
2. An overall positive association between SSS and SST from equator to the pole.
3. In lower latitude, due to the pronounced salt accumulation as a result of excess heating and oceanic currents, SSS surges.
4. SSS decreases in comparison with SST during winter, except for the Brazilian coast due to the strong North Brazil Current.

Predictive Inferential Accuracy

Predictive Inference on 600 hold out observations

	Coverage	MSPE	95% PI Length	Efficiency$=\frac{\log_2(\text{Eff. Sample Size})}{\text{Comp. Time}}$
AMC	0.98	2.92	6.60	9.99
PIE	0.97	2.93	5.93	-
CMC	0.90	74.95	24.71	1.06
WASP	0.98	2.92	5.66	9.99

Some Important Findings

1. All divide-and-conquer schemes with the theoretical backing perform similarly.
2. Popular ML aggregation scheme CMC offers suboptimal inference.

References

- Guhaniyogi, R. and Qamar, S. and Dunson, D.B. (2017). Bayesian tensor regression. *The Journal of Machine Learning Research*, 18(1), 2733–2763.
- Guhaniyogi, R. (2017). Convergence rate of Bayesian supervised tensor modeling with multiway shrinkage priors. *Journal of Multivariate Analysis*, 160, 157–168.
- Guhaniyogi, R. and Spencer, D. (2021). Bayesian tensor response regression with an application to brain activation studies. *Bayesian Analysis*, 16(4), 1221–1249.
- Spencer, D. and Guhaniyogi, R. and Prado, R. (2020). Joint Bayesian estimation of voxel activation and inter-regional connectivity in fMRI experiments. *Psychometrika*, 85(4), 845–869.
- Guha, S. and Guhaniyogi, R. (2021). Bayesian generalized sparse symmetric tensor-on-vector regression. *Technometrics*, 63(2), 160–170.
- Guhaniyogi, R. and Banerjee, S. (2018). Meta-kriging: Scalable Bayesian modeling and inference for massive spatial datasets. *Technometrics*, 60(4), 430–444.
- Guhaniyogi, R. and Li, C. and Savitsky, T. D. and Srivastava, S. (2022+). A divide-and-conquer Bayesian approach to large-scale kriging. *arXiv preprint arXiv:1712.09767*.
- Guhaniyogi, R. and Li, C. and Savitsky, T. D. and Srivastava, S. (2022). Distributed Bayesian varying coefficient modeling using a Gaussian process prior. Accepted, *Journal of Machine Learning Research (arXiv preprint arXiv:2006.00783)*.

References

- Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I. and McCulloch, R. E. (2016). Bayes and big data: The consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, 11(2), 78–88.
- Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A. and Blei, D. M. (2017). Automatic differentiation variational inference. *Journal of Machine Learning Research*.
- Neiswanger, W. and Xing, E. (2017). Post-inference prior swapping. *International Conference on Machine Learning*, 2594–2602.
- Minsker, S., Srivastava, S., Lin, L. and Dunson, D. B. (2017). Robust and scalable Bayes via a median of subset posterior measures. *Journal of Machine Learning Research*, 18(1), 4488–4527.
- Li, C., Srivastava, S. and Dunson, D. B. (2017). Simple, scalable and accurate posterior interval estimation. *Biometrika*, 104(3), 665–680.
- Armagan, A. and Dunson, D. B. and Lee, J. (2013). Generalized double Pareto shrinkage. *Statistica Sinica*, 23(1), 119–143.
- Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review*, 51(3), 455–500.
- Guhaniyogi, R. (2020). Bayesian methods for tensor regression. *Wiley StatsRef: Statistical References Online*, 1–18.