Bayesian Tensor Regression: A Scalable Bayesian Framework in Neuroscience Applications

Department of Applied Mathematics and Statistics, University of California Santa Cruz
Rajarshi Guhaniyogi, Ph.D

April 9, 2016

High Dimensional Regression

- In typical high dimensional regression problems we have response $y_i \in \mathbb{R}$ $(i = 1, .., n)$ associated with a high dimensional predictor vector $\boldsymbol{x}_i \in \mathbb{R}^p$.

Central Theme: Statistical + Computational Efficiency

High Dimensional Regression

- In typical high dimensional regression problems we have response $y_i \in \mathbb{R}$ $(i = 1, .., n)$ associated with a high dimensional predictor vector $\boldsymbol{x}_i \in \mathbb{R}^p$.
- p is very big and n is moderate–"large p, small n" problem.

High Dimensional Regression

- In typical high dimensional regression problems we have response $y_i \in \mathbb{R}$ $(i = 1, .., n)$ associated with a high dimensional predictor vector $\boldsymbol{x}_i \in \mathbb{R}^p$.
- p is very big and n is moderate–"large p, small n" problem.
- Occurs routinely in many Biomedical applications.

High Dimensional Regression

- In typical high dimensional regression problems we have response $y_i \in \mathbb{R}$ $(i = 1, .., n)$ associated with a high dimensional predictor vector $\boldsymbol{x}_i \in \mathbb{R}^p$.
- p is very big and n is moderate–"large p, small n" problem.
- Occurs routinely in many Biomedical applications.
- Dimensionality reduction is critical.

High Dimensional Regression

- In typical high dimensional regression problems we have response $y_i \in \mathbb{R}$ $(i = 1, .., n)$ associated with a high dimensional predictor vector $\boldsymbol{x}_i \in \mathbb{R}^p$.
- p is very big and n is moderate–"large p, small n" problem.
- Occurs routinely in many Biomedical applications.
- Dimensionality reduction is critical.

Central Theme: Statistical + Computational Efficiency

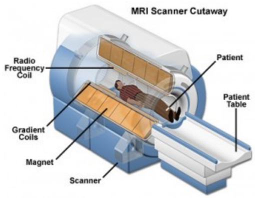
High Dimensional Regression

- In typical high dimensional regression problems we have response $y_i \in \mathbb{R}$ $(i = 1, .., n)$ associated with a high dimensional predictor vector $\boldsymbol{x}_i \in \mathbb{R}^p$.
- p is very big and n is moderate–"large p, small n" problem.
- Occurs routinely in many Biomedical applications.
- Dimensionality reduction is critical.

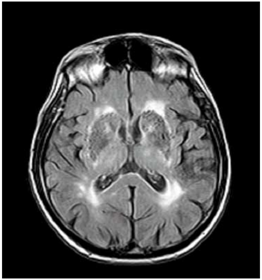
Object Oriented Regression

- Answering complex inferential questions can lead to massive dimensional regression.

Detecting Voxels in Diseased Brain



(a) MRI machine



(b) cross section MRI scan

Tensor predictor: Resting state fMRI for 550 people (some patients, some normal).
scalar predictors: volume of the brain, sex, smoking during pregnancy.
Response: Binary indicator whether diseased or not.

Penalized Optimization:Unsatisfactory Predictive Performance

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

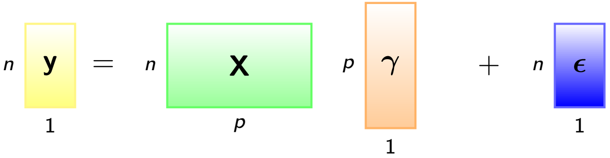
Penalized Optimization:Unsatisfactory Predictive Performance

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

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

$arg\,\min_{\gamma} \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\gamma)^2 + \zeta \sum_{j=1}^{p} \psi(\gamma_j) \rightarrow$ **Penalized Opt.**

Penalized Optimization:Unsatisfactory Predictive Performance



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

$arg \min_{\gamma} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i' \gamma)^2 + \zeta \sum_{j=1}^{p} \psi(\gamma_j) \rightarrow$ **Penalized Opt.**

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

Penalized Optimization:Unsatisfactory Predictive Performance

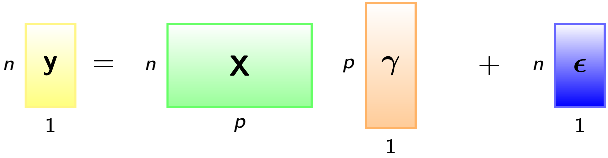
$$\underset{n \times 1}{\mathbf{y}} = \underset{n \times p}{\mathbf{X}} \; \underset{p \times 1}{\boldsymbol{\gamma}} + \underset{n \times 1}{\boldsymbol{\epsilon}}$$

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

$arg \min\limits_{\gamma} \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\gamma)^2 + \zeta \sum_{j=1}^{p} \psi(\gamma_j) \rightarrow$ **Penalized Opt.**

- LASSO (Tibshirani, 1996), Elastic Net (Zhou et al., 2005), tons of other variants.
- Efficient convex optimization algorithms (Hastie, 2003; Friedman, 2010) to produce point prediction for high dimensional regression.

Penalized Optimization:Unsatisfactory Predictive Performance



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

$$\arg\min_{\gamma} \sum_{i=1}^{n} (y_i - \mathbf{x}_i' \gamma)^2 + \zeta \sum_{j=1}^{p} \psi(\gamma_j) \rightarrow \textbf{Penalized Opt.}$$

- LASSO (Tibshirani, 1996), Elastic Net (Zhou et al., 2005), tons of other variants.
- Efficient convex optimization algorithms (Hastie, 2003; Friedman, 2010) to produce point prediction for high dimensional regression.
- Unsatisfactory predictive uncertainty.

Bayesian High Dim. Reg.:Unsuitable in High Dimension

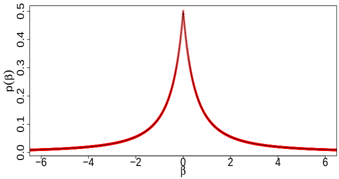
- Bayesians choose sparsity-favoring priors on γ concentrating around an S-sparse vector $\gamma_0 \in \mathscr{R}^p$.

Bayesian High Dim. Reg.:Unsuitable in High Dimension

- Bayesians choose sparsity-favoring priors on γ concentrating around an S-sparse vector $\gamma_0 \in \mathscr{R}^p$.

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)

$$\gamma_j \sim g,$$

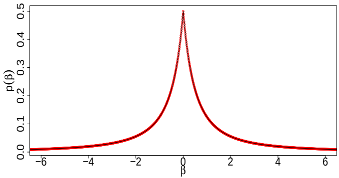
g heavy tailed density.

Bayesian High Dim. Reg.:Unsuitable in High Dimension

- Bayesians choose sparsity-favoring priors on γ concentrating around an S-sparse vector $\gamma_0 \in \mathscr{R}^p$.

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)

$$\gamma_j \sim g,$$

g 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 γ concentrating around an S-sparse vector $\gamma_0 \in \mathscr{R}^p$.

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

- $p = p_1 \times p_2 \times \cdots p_D$, 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 brain images carry.

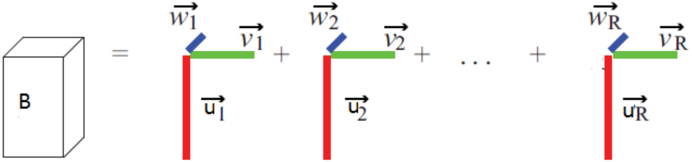
- Important shrinkage priors, Bayesian Lasso (Park et al., 2008; Hans, 2009), Horseshoe (Carvalho et al., 2009), Generalized Double Pareto (Armagan et al., 2013).

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 $\Rightarrow$ $\boldsymbol{B} = \sum_{r=1}^{R} \beta_1^{(r)} \circ \cdots \circ \beta_D^{(r)}$

$\beta_j^{(r)} \in \mathscr{R}^{p_j}$, $\circ$ 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.
- Keeps spatial structure of $\boldsymbol{X}$ intact $\Rightarrow$ potentially better inference.

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

Frequentist Tensor Regression (Zhou et al. 2013, 2014):

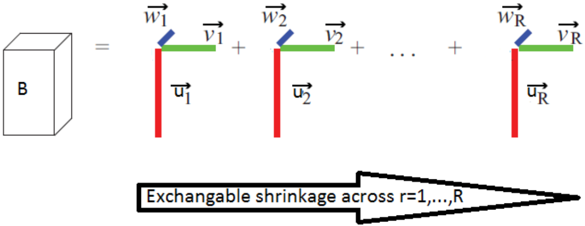
$\psi(\cdot)$ = convex penalty function, ζ = tuning parameter

$$arg \min_{\gamma, \beta_j^{(r)}} \sum_{i=1}^{n} (y_i - \langle \boldsymbol{X}_i, \boldsymbol{B} \rangle - \boldsymbol{z}_i' \gamma)^2 + \zeta \left[\psi(\gamma) + \sum_{r=1}^{R} \sum_{j=1}^{D} \psi\left(\beta_j^{(r)}\right) \right]$$

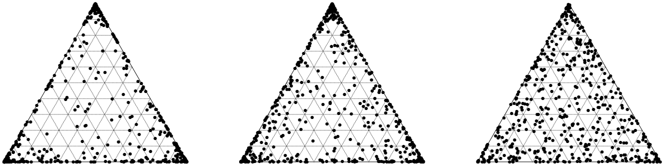
Issues with Frequentist Tensor Regression (FTR)

1. Choice of R is adhoc.
2. Result depends heavily on the tuning parameter ζ. Choice of the tuning parameter is also uncertain.
3. Prediction and inference can be improved.

Multiway Shrinkage Prior for B (Guhaniyogi et al. 2015)

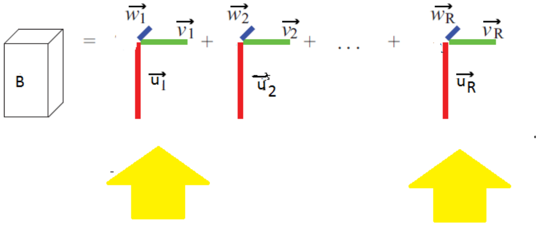


$\beta_j^{(r)} \sim N(\mathbf{0}, \boldsymbol{W}_{jr}\tau\phi_r)$, ϕ_r's rank specific parameters. Shrinkage across ranks: $(\phi_1, ..., \phi_R) \sim Dirichlet(\alpha_1, ..., \alpha_R)$.



(c) $\alpha_i = 0.2$ (d) $\alpha_i = 0.3$ (e) $\alpha_i = 0.5$

Multiway Dirichlet Generalized Double Pareto Prior (M-DGDP)



Shrinkage within every rank

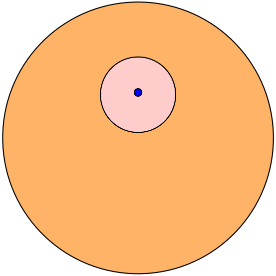
$$w_{jr,k} \sim \mathrm{Exp}(\lambda_{jr}^2/2), \quad \lambda_{jr} \sim \mathrm{Ga}(a_\lambda, b_\lambda), \tau \sim IG(a_\tau, b_\tau)$$

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

$$\boxed{\beta_{j,k}^{(r)} | \lambda_{jr}, \phi_r, \tau \overset{i.i.d}{\sim} \mathrm{DE}(\lambda_{jr}/\sqrt{\phi_r \tau}),\ 1 \le k \le p_j,}$$

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

General Theoretical Setup: Guhaniyogi et al., 2015



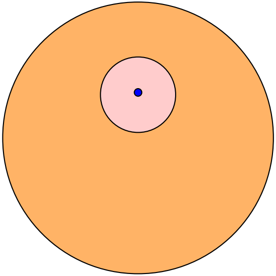
True Model

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

Class of tensor reg. models fitted to the data

KL metric ball of radius ϵ around the truth

General Theoretical Setup: Guhaniyogi et al., 2015



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, Guhaniyogi et al. 2015

Theorem

The posterior is consistent under the following assumptions.

Posterior Consistency Results, Guhaniyogi et al. 2015

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. (Structure on the true coefficients)

Posterior Consistency Results, Guhaniyogi et al. 2015

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. (Structure on the true coefficients)
2. $\sup_{l=1,\ldots,p_{j,n}} |\beta_{j,n,l}^{0(r)}| < \infty$, for all $j = 1, \ldots, D;\ r = 1, \ldots, R$. (Structure on the true coefficients)

Posterior Consistency Results, Guhaniyogi et al. 2015

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. (Structure on the true coefficients)
2. $\sup_{l=1,\ldots,p_{j,n}} |\beta_{j,n,l}^{0(r)}| < \infty$, for all $j = 1,\ldots,D;\ r = 1,\ldots,R$. (Structure on the true coefficients)
3. $\sum_{j=1}^{D} p_{j,n} \log(p_{j,n}) = o(n)$. (Dimension of margins)

Posterior Consistency Results, Guhaniyogi et al. 2015

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. (Structure on the true coefficients)
2. $\sup_{l=1,\dots,p_{j,n}} |\beta_{j,n,l}^{0(r)}| < \infty$, for all $j = 1, \dots, D;\ r = 1, \dots, R$. (Structure on the true coefficients)
3. $\sum_{j=1}^{D} p_{j,n} \log(p_{j,n}) = o(n)$. (Dimension of margins)
4. $M_n = \frac{1}{n}\sqrt{\sum_{i=1}^{n} ||\boldsymbol{X}_i||_2^2}$, $H_1 n^{\rho_1} < M_n < H_2 n^{\rho_2}$, $H_1, H_2, \rho_1, \rho_2 > 0$.

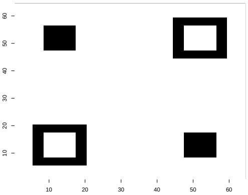
Simulation Setup

Data Generation

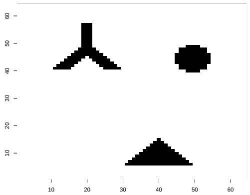
$$y_i = \langle \boldsymbol{X}_i, \boldsymbol{B}^0 \rangle + \epsilon_i, \epsilon_i \sim N(0, \sigma_0^2), \;\; i = 1, ..., n$$

(i) $n = 1000$
(ii) $\sigma_0^2 = 1$
(iii) $\boldsymbol{B}^0$ is 64×64
(iv) $x_{i_1,i_2} \sim N(0,1) \; \forall \;\; i_1 = 1:64, \; i_2 = 1:64.$

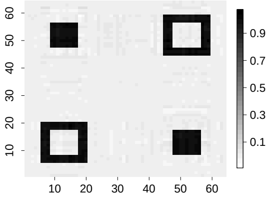
Competitors
Frequentist Tensor Regression (FTR)
Vectorized Lasso (Lasso)



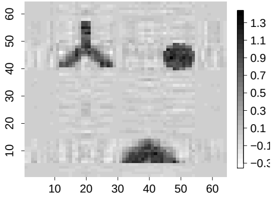
(f) Boxes (11.0%)



(g) Shapes (6.8%)

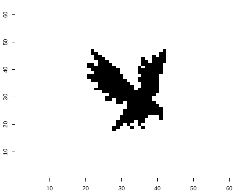


(h) Boxes Recov.

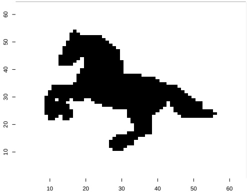


(i) Shapes Recov.

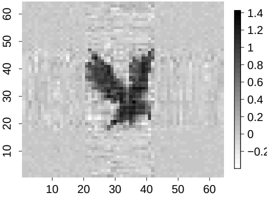
Results: True Coefficients "Ready-made" Images



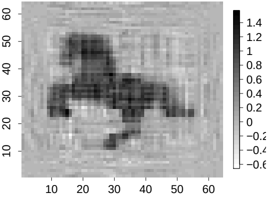
(j) Bird



(k) Horse



(l) Bird Recov.



(m) Horse Recov.

Comparison with Competitors: Lower Mean Squared Error (MSE) with Excellent Coverage

Case	BTR	FTR	Lasso	vox
Eagle	$\mathbf{0.226}_{0.02}$	$0.354_{0.03}$	$0.665_{0.03}$	> 0
Horse	$\mathbf{0.278}_{0.01}$	$0.391_{0.03}$	$0.888_{0.01}$	> 0
Eagle	$\mathbf{0.085}_{0.00}$	$0.163_{0.03}$	$0.097_{0.00}$	$= 0$
Horse	$\mathbf{0.137}_{0.00}$	$0.215_{0.02}$	$0.155_{0.02}$	$= 0$

Comparison with Competitors: Lower Mean Squared Error (MSE) with Excellent Coverage

Case	BTR	FTR	Lasso	vox
Eagle	$\mathbf{0.226}_{0.02}$	$0.354_{0.03}$	$0.665_{0.03}$	> 0
Horse	$\mathbf{0.278}_{0.01}$	$0.391_{0.03}$	$0.888_{0.01}$	> 0
Eagle	$\mathbf{0.085}_{0.00}$	$0.163_{0.03}$	$0.097_{0.00}$	$= 0$
Horse	$\mathbf{0.137}_{0.00}$	$0.215_{0.02}$	$0.155_{0.02}$	$= 0$

Coverage for M-DGDP is 0.94 and 0.92 respectively.

Simulated Response with Real Vector and Tensor Covariates

- $30 \times 30 \times 30$ MRI images (predictor tensor) for 550 individuals.
- Response is simulated as $y \sim N(\langle \boldsymbol{X}, \boldsymbol{B}^0 \rangle + \boldsymbol{z}'\gamma, 1)$ for every individual.
- Three different rank-2 tensor coefficients are simulated with varying sparsity.

Case	BTR	FTR	Lasso
Case 1	$\mathbf{0.13}_{0.01}$	$0.15_{0.01}$	$0.15_{0.01}$
Case 2	$\mathbf{0.20}_{0.01}$	$0.23_{0.01}$	$0.24_{0.01}$
Case 3	$\mathbf{0.17}_{0.01}$	$0.19_{0.01}$	$0.19_{0.01}$

Brain Connectome Data Application

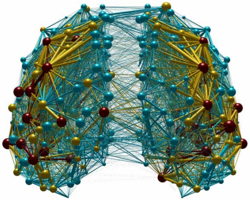
- Data are extracted from diffusion tensor imaging (DTI) for 109 individuals.

Brain Connectome Data Application

- Data are extracted from diffusion tensor imaging (DTI) for 109 individuals.
- For each individual, brain connections are encoded by a 70×70 weighted adjacency matrix.

Brain Connectome Data Application

- Data are extracted from diffusion tensor imaging (DTI) for 109 individuals.
- For each individual, brain connections are encoded by a 70×70 weighted adjacency matrix.
- The (i, j)-th entry of the matrix is the estimated number of fiber tracts connecting the i-th and j-th brain region.



Goal

Developing a predictive model of composite creativity index (CCI) based on neuronal connectivity.

Predictive Inference: Brain Connectome Data

- **Response:** Composite Creativity Index (CCI).

Predictive Inference: Brain Connectome Data

- **Response:** Composite Creativity Index (CCI).
- **Vector Predictor:** 10 clinical covariates e.g. openness, agreeableness, conscientiousness.

Predictive Inference: Brain Connectome Data

- **Response:** Composite Creativity Index (CCI).
- **Vector Predictor:** 10 clinical covariates e.g. openness, agreeableness, conscientiousness.
- **Tensor Predictor:** 70×70 weighted adjacency matrix.
- Predictive inference of lasso and BTR with 10 folds of the data.

Method	avg(RMSE)	sd(RMSE)	avg(cov.)	sd(cov.)	avg(cor.)	sd(cor.)
Lasso	9.18	1.64	0.63	0.20	0.31	0.11
BTR	9.03	2.18	0.91	0.10	0.32	0.13

What have we achieved so far?

- Penalized optimization unsatisfactory for predictive uncertainties; Bayesian shrinkage priors statistically inefficient, computationally not scalable to tensor predictors with large number of voxels, destroy tensor structure in the predictors.

What have we achieved so far?

- Penalized optimization unsatisfactory for predictive uncertainties; Bayesian shrinkage priors statistically inefficient, computationally not scalable to tensor predictors with large number of voxels, destroy tensor structure in the predictors.
- Frequentist Tensor Regression is less robust with choice of the tuning parameter, selects R in an adhoc way.

Brief Overview of Tensor Regression

What have we achieved so far?

- Penalized optimization unsatisfactory for predictive uncertainties; Bayesian shrinkage priors statistically inefficient, computationally not scalable to tensor predictors with large number of voxels, destroy tensor structure in the predictors.
- Frequentist Tensor Regression is less robust with choice of the tuning parameter, selects R in an adhoc way.

Tensor Regression with M-DGDP prior

- A novel multiway shrinkage prior– R selection is automated,
- significantly better performance, excellent parametric and predictive coverage.

Brief Overview of Tensor Regression

What have we achieved so far?

- Penalized optimization unsatisfactory for predictive uncertainties; Bayesian shrinkage priors statistically inefficient, computationally not scalable to tensor predictors with large number of voxels, destroy tensor structure in the predictors.
- Frequentist Tensor Regression is less robust with choice of the tuning parameter, selects R in an adhoc way.

Tensor Regression with M-DGDP prior

- A novel multiway shrinkage prior– R selection is automated,
- significantly better performance, excellent parametric and predictive coverage.
- Supported by theoretical convergence results.

Brief Overview of Tensor Regression

What have we achieved so far?

- Penalized optimization unsatisfactory for predictive uncertainties; Bayesian shrinkage priors statistically inefficient, computationally not scalable to tensor predictors with large number of voxels, destroy tensor structure in the predictors.
- Frequentist Tensor Regression is less robust with choice of the tuning parameter, selects R in an adhoc way.

Tensor Regression with M-DGDP prior

Useful in Other Tensor Regression Framework?

Nontrivial extension of BTR useful in providing a scalable framework for the brain activation study. Stay tuned....

predictive coverage.

- Supported by theoretical convergence results.

References

- Armagan, A., Dunson, D.B., and Lee, J. (2013), "Generalized Double Pareto Shrinkage," *Statistica Sinica*, 23, 119-143.
- Guhaniyogi, R., Qamar, S., and Dunson, D. B. (2015), "Bayesian Tensor Regression," *arXiv:1509.06490*.
- Zhou, H. (2013), "Tensor Regression with Applications in Neuroimaging Data Analysis," *Journal of the American Statistical Association*, **108**, 540-552.
- Zhou, H. and Li, Lexin (2014), "Regularized Matrix Regression," *Journal of the Royal Statistical Society, Series B*, **76**, 463-483.
- Carvalho, C.M., Polson, N.G., and Scott, J.G. (2009), "Handling Sparsity via The Horseshoe," *JMLR: W & CP*, 5, 73-80.