Bayesian Factor Models in High Dimensions

Natesh S. Pillai
Department of Statistics,
Harvard University

November 10, 2017

Vienna University of Economics and Business

Collaborators

- Anirban Bhattacharya (Texas A&M)
- Debdeep Pati (Texas A& M)
- David B. Dunson (Duke)
- Gautam Sabnis (U of Michigan), Barbara Englehardt (Princeton)

and thanks to James Scott (UTA).

Outline of the Talk

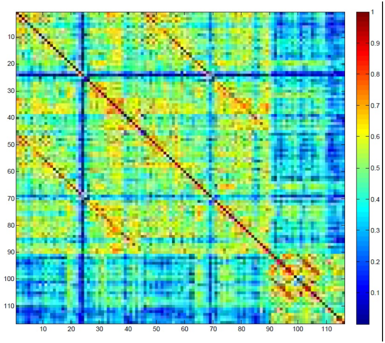
- Bayesian Estimation in "Large-p, small-n".
- Statistical Efficiency *vs.* Computational Efficiency, a key issue.
- In this talk: a concrete formulation of this problem.
- Shrinkage priors.
- Hierarchical Factor models for computational speed up.

Conversation with Peter E. Huybers

- Motivation: Time variability in covariance patterns for climate data: stationarity?
- Instrumental measurements, only for the past $n = 150$ years.
- Measurements on $p = 2000$ latitude-longitude points.
- Estimate $O(p^2)$ parameters.
- Need judicious modeling.

Autism spectra-matrix

- Brain spectra covariance matrix for autism infected adults at the National Taiwan University Hospital.



- Understand these patterns

Motivation

- An important class of models: Latent factor methods (West, 2003; Lucas et al., 2006; Carvalho et al., 2008).
- Set $y_i = (y_{i1}, \ldots, y_{ip})^{\mathrm{T}}, i = 1, \ldots, n$
- $y_i \sim \mathrm{N}_p(0, \Sigma)$
- Goal: Estimate Σ.
- Note $p \gg n$.

Gaussian factor models

- Unstructured Σ has $O(p^2)$ free elements
- Assume a factor model

$$\Sigma = \Lambda\Lambda' + \sigma^2 \mathrm{I}_p$$

  via parsimonious factorization
- $k = O(1)$, the number of factors.
- Λ is the factor loadings.
- Λ is $p \times k$ and thus model complexity $O(p)$ - huge dimensionality reduction, but still challenging.

Sparse factor modeling

- *Sparse factor modeling* (West, 2003); also (Lucas et al., 2006; Carvalho et al., 2008) and many others
- Allow zeros in loadings.
- Assume each column of Λ has only s non-zero elements.
- Here s denotes the sparsity.

High-dimensional covariance estimation

- ‘Frequentist’ solution – MLE doesn’t work.

High-dimensional covariance estimation

- ‘Frequentist’ solution – MLE doesn’t work.
- Start with sample covariance matrix:

$$\Sigma^{\mathrm{sample}} = \frac{1}{n} \sum_{i=1}^{n} y_i y_i^T \ .$$

High-dimensional covariance estimation

- 'Frequentist' solution – MLE doesn't work.
- Start with sample covariance matrix:

$$\Sigma^{\text{sample}} = \frac{1}{n}\sum_{i=1}^{n} y_i y_i^T \ .$$

- Great interest in regularized estimation (Bickel & Levina, 2008a, b; Wu and Pourahmadi, 2010, Cai and Zhou, 2011 ...)

High-dimensional covariance estimation

- 'Frequentist' solution – MLE doesn't work.
- Start with sample covariance matrix:

$$\Sigma^{\text{sample}} = \frac{1}{n} \sum_{i=1}^{n} y_i y_i^T \ .$$

- Great interest in regularized estimation (Bickel & Levina, 2008a, b; Wu and Pourahmadi, 2010, Cai and Zhou, 2011 ...)
- Efficient Estimators based on Thresholding:

$$\hat{\Sigma}_{ij} = \Sigma_{ij}^{\text{sample}} \, 1_{|\Sigma_{ij}^{\text{sample}}| > t_n} \ .$$

High-dimensional covariance estimation

- ‘Frequentist’ solution – MLE doesn’t work.
- Start with sample covariance matrix:

$$\Sigma^{\text{sample}} = \frac{1}{n}\sum_{i=1}^{n} y_i y_i^T \ .$$

- Great interest in regularized estimation (Bickel & Levina, 2008a, b; Wu and Pourahmadi, 2010, Cai and Zhou, 2011 ...)
- Efficient Estimators based on Thresholding:

$$\hat{\Sigma}_{ij} = \Sigma_{ij}^{\text{sample}} 1_{|\Sigma_{ij}^{\text{sample}}| > t_n} \ .$$

- Unstable; Confidence intervals..?

Bayes-Frequentist Synergy

- Question: Given most regularization estimators are posterior modes of a Bayesian model, can one run a Markov Chain Monte Carlo algorithm to sample from the posterior distribution, and compute the uncertainty intervals?

Bayes-Frequentist Synergy

- Question: Given most regularization estimators are posterior modes of a Bayesian model, can one run a Markov Chain Monte Carlo algorithm to sample from the posterior distribution, and compute the uncertainty intervals?
- Successfully exploited in "classical statistics": *i.e.*, *fixed*-p, large-n situation.

Bayes-Frequentist Synergy

- Question: Given most regularization estimators are posterior modes of a Bayesian model, can one run a Markov Chain Monte Carlo algorithm to sample from the posterior distribution, and compute the uncertainty intervals?
- Successfully exploited in "classical statistics": *i.e.*, *fixed*-p, large-n situation.
- Here we assume, $p_n = O(e^{n^\alpha})$ with $\alpha < 1/3$ (ultra high-dimensions).

- Set $k = 1$, thus
$$\Sigma = \sigma^2 \mathrm{I}_p + \Lambda\Lambda'$$

Bayes-Frequentist Synergy

- Set $k = 1$, thus
$$\Sigma = \sigma^2 \mathrm{I}_p + \Lambda\Lambda'$$
- Λ is a $p \times 1$ vector, with only s many non-zeroes.

Bayes-Frequentist Synergy

- Set $k = 1$, thus
$$\Sigma = \sigma^2 \mathrm{I}_p + \Lambda\Lambda'$$
- Λ is a $p \times 1$ vector, with only s many non-zeroes.
- Questions:
  1. What is the minimax rate for estimating Σ?
  2. What prior on the vector Λ leads to a posterior which concentrates at the minimax rate?

Bayes-Frequentist Synergy

- Set $k = 1$, thus
$$\Sigma = \sigma^2 \mathrm{I}_p + \Lambda\Lambda'$$
- Λ is a $p \times 1$ vector, with only s many non-zeroes.
- Questions:
  1. What is the minimax rate for estimating Σ?
  2. What prior on the vector Λ leads to a posterior which concentrates at the minimax rate?
- Answer to the above two questions: a first step towards Bayes-Frequentist agreement in this "large-p, small-n" problem.

Assumptions:

- Recall, $k = 1$, thus

$$\Sigma = \sigma^2 \mathrm{I}_p + \Lambda\Lambda'$$

- Λ is a $p \times 1$ vector, with only s many non-zeroes.
- $p_n = O(e^{n^\alpha})$ with $\alpha < 1/3$
- Key facet:

$$\sigma^2 < \|\Lambda\Lambda'\|_2 = \|\Lambda\|^2 = O(\log p_n)$$

- Thus Σ is not a "small" perturbation of identity (different from other common assumptions...)

Minimax rate:

- Theorem (Minimax Lower Bound)
  (Pati, Bhattacharya, P., Dunson, 2014)

$$\inf_{\hat{\Sigma}} \sup_{\Sigma} \|\hat{\Sigma} - \Sigma\|_2 \geq \sqrt{\frac{(\log p_n)^3 s}{n}}$$

  - Proof uses a variant of Le Cam's method/ Fano's Lemma.
  - Questions:
    1. What is the minimax rate for estimating Σ? $= \sqrt{\frac{(\log p_n)^3 s}{n}}$
    2. What prior on the vector Λ leads to a posterior which concentrates at the minimax rate?

Posterior rate

- What prior on the vector Λ leads to a posterior which concentrates at the minimax rate?

Posterior rate

- What prior on the vector Λ leads to a posterior which concentrates at the minimax rate?
- Let Σ_0 be the true data generating parameter.

Posterior rate

- What prior on the vector Λ leads to a posterior which concentrates at the minimax rate?
- Let Σ_0 be the true data generating parameter.
- We seek ϵ_n such that (Ghosh and Ramamoorthi)

$$\lim_{n\to\infty} \mathbb{P}_{\Sigma_0}(\|\Sigma - \Sigma_0\|_2 \geq \epsilon_n|\mathrm{Data}) = 0.$$

Posterior rate

- What prior on the vector Λ leads to a posterior which concentrates at the minimax rate?
- Let Σ_0 be the true data generating parameter.
- We seek ϵ_n such that (Ghosh and Ramamoorthi)

$$\lim_{n\to\infty} \mathbb{P}_{\Sigma_0}(\|\Sigma - \Sigma_0\|_2 \geq \epsilon_n | \mathrm{Data}) = 0.$$

- Where to look for possible priors?

Posterior rate

- What prior on the vector Λ leads to a posterior which concentrates at the minimax rate?
- Let Σ_0 be the true data generating parameter.
- We seek ϵ_n such that (Ghosh and Ramamoorthi)

$$\lim_{n \to \infty} \mathbb{P}_{\Sigma_0}(\|\Sigma - \Sigma_0\|_2 \geq \epsilon_n | \mathrm{Data}) = 0.$$

- Where to look for possible priors?
- First choice: point mass priors These can be thought of as the Bayesian analogue of thresholding estimates.

Posterior rate: Point Mass Priors

- Set
$$\Lambda_j \sim (1-\pi)\delta_0 + \pi g(\cdot)$$
where $g(\cdot)$ has exponential or heavier tails.

Posterior rate: Point Mass Priors

- Set
$$\Lambda_j \sim (1-\pi)\delta_0 + \pi g(\cdot)$$
where $g(\cdot)$ has exponential or heavier tails.
- If s is known, then $\pi = s/p$ is a natural choice.

Posterior rate: Point Mass Priors

- Set
$$\Lambda_j \sim (1-\pi)\delta_0 + \pi g(\cdot)$$
where $g(\cdot)$ has exponential or heavier tails.
- If s is known, then $\pi = s/p$ is a natural choice.
- If s is unknown, set a hyper-prior (Scott & Berger 2010, Castillo & van der Vaart, 2012)
$$\pi \sim \text{Beta}(1, p+1).$$

Posterior rate: Point Mass Priors

- Set
$$\Lambda_j \sim (1-\pi)\delta_0 + \pi g(\cdot)$$
where $g(\cdot)$ has exponential or heavier tails.
- If s is known, then $\pi = s/p$ is a natural choice.
- If s is unknown, set a hyper-prior (Scott & Berger 2010, Castillo & van der Vaart, 2012)
$$\pi \sim \text{Beta}(1, p+1).$$
- Has connections to automatic multiplicity adjustments, and also optimal in other contexts.

Frequentist validation of Bayesian procedures

- Bernstein von Mises Theorem (1949 Doob) - posterior is independent of prior if sample size is large
- True only for finite dimensions
- Inconsistency of nonparametric Bayes (1986 Freedman and Diaconis)
- Apparently simple minded priors can go wrong

Minimax rate: Pati, Bhattacharya, P., Dunson, 2014

- Theorem (Posterior Convergence Rate)

  For point mass priors, with $\epsilon_n = \sqrt{\frac{(\log p_n)^3 s}{n}}$*, we have*

$$\lim_{n\to\infty} \mathbb{P}_{\Sigma_0}(\|\Sigma - \Sigma_0\|_2 \geq \epsilon_n |\mathrm{Data}) = 0.$$

Minimax rate: Pati, Bhattacharya, P., Dunson, 2014

- Theorem (Posterior Convergence Rate)

  For point mass priors, with $\epsilon_n = \sqrt{\frac{(\log p_n)^3 s}{n}}$, *we have*

  $$\lim_{n\to\infty} \mathbb{P}_{\Sigma_0}(\|\Sigma - \Sigma_0\|_2 \geq \epsilon_n | \mathrm{Data}) = 0.$$

  - Proof involves novel ideas, and uses results from non-asymptotic random matrix theory (Vershynin 2010, Tropp 2012).

Minimax rate: Pati, Bhattacharya, P., Dunson, 2014

- Theorem (Posterior Convergence Rate)

  For point mass priors, with $\epsilon_n = \sqrt{\frac{(\log p_n)^3 s}{n}}$, *we have*

  $$\lim_{n \to \infty} \mathbb{P}_{\Sigma_0}(\|\Sigma - \Sigma_0\|_2 \geq \epsilon_n | \mathrm{Data}) = 0.$$

  - Proof involves novel ideas, and uses results from non-asymptotic random matrix theory (Vershynin 2010, Tropp 2012).
  - The sample covariance will not be efficient in detecting the points.

Minimax rate: Pati, Bhattacharya, P., Dunson, 2014

- Theorem (Posterior Convergence Rate)

  For point mass priors, with $\epsilon_n = \sqrt{\frac{(\log p_n)^3 s}{n}}$, *we have*

  $$\lim_{n\to\infty} \mathbb{P}_{\Sigma_0}(\|\Sigma - \Sigma_0\|_2 \geq \epsilon_n | \mathrm{Data}) = 0.$$

  - Proof involves novel ideas, and uses results from non-asymptotic random matrix theory (Vershynin 2010, Tropp 2012).
  - The sample covariance will not be efficient in detecting the points.
  - Take the low dimensional projections of the data, and then compute the sample covariance matrix.

Minimax rate: Pati, Bhattacharya, P., Dunson, 2013

- Theorem (Posterior Convergence Rate)

  For point mass priors, with $\epsilon_n = \sqrt{\frac{(\log p_n)^3 s}{n}}$*, we have*

$$\lim_{n \to \infty} \mathbb{P}_{\Sigma_0}(\|\Sigma - \Sigma_0\|_2 \geq \epsilon_n | \mathrm{Data}) = 0.$$

Minimax rate: Pati, Bhattacharya, P., Dunson, 2013

- Theorem (Posterior Convergence Rate)

  For point mass priors, with $\epsilon_n = \sqrt{\frac{(\log p_n)^3 s}{n}}$, *we have*

  $$\lim_{n\to\infty} \mathbb{P}_{\Sigma_0}(\|\Sigma - \Sigma_0\|_2 \geq \epsilon_n | \mathrm{Data}) = 0.$$

  - Questions:
    1. What is the minimax rate for estimating Σ? $= \sqrt{\frac{(\log p_n)^3 s}{n}}$
    2. What prior on the vector Λ leads to a posterior which concentrates at the minimax rate? = Point mass priors achieve this!

Have we solved the problem?

- ...Not yet!

Have we solved the problem?

- ...Not yet!
- The MCMC algorithm for sampling the posterior has to explore a model space of dimension $O(2^p)$.

Have we solved the problem?

- ...Not yet!
- The MCMC algorithm for sampling the posterior has to explore a model space of dimension $O(2^p)$.
- Curse of dimensionality catches up fast; not even feasible for moderate p.

Have we solved the problem?

- ...Not yet!
- The MCMC algorithm for sampling the posterior has to explore a model space of dimension $O(2^p)$.
- Curse of dimensionality catches up fast; not even feasible for moderate p.
- Effective sample size is small; Point mass priors are statistically efficient, but computationally NOT efficient!

Have we solved the problem?

- ...Not yet!
- The MCMC algorithm for sampling the posterior has to explore a model space of dimension $O(2^p)$.
- Curse of dimensionality catches up fast; not even feasible for moderate p.
- Effective sample size is small; Point mass priors are statistically efficient, but computationally NOT efficient!
- OK, what now?

Funny you should ask...

- Continuous Shrinkage Priors!

Funny you should ask...

- Continuous Shrinkage Priors!
- Appealing computationally & philosophically to relax assumption of exact zeros.

Funny you should ask...

- Continuous Shrinkage Priors!
- Appealing computationally & philosophically to relax assumption of exact zeros.
- Zillions of them (Park and Casella, 2008; Carvalho, Polson and Scott, 2010; Armagan, Dunson and Lee, 2011; Hans, 2011,..)

Funny you should ask...

- Continuous Shrinkage Priors!
- Appealing computationally & philosophically to relax assumption of exact zeros.
- Zillions of them (Park and Casella, 2008; Carvalho, Polson and Scott, 2010; Armagan, Dunson and Lee, 2011; Hans, 2011,..)
- Polson & Scott (2010) unifies them as

$$\Lambda_j \sim N(0, \psi_j \tau), \quad \psi_j \overset{i.i.d}{\sim} g, \quad \tau \sim f$$

Funny you should ask...

- Continuous Shrinkage Priors!
- Appealing computationally & philosophically to relax assumption of exact zeros.
- Zillions of them (Park and Casella, 2008; Carvalho, Polson and Scott, 2010; Armagan, Dunson and Lee, 2011; Hans, 2011,..)
- Polson & Scott (2010) unifies them as

$$\Lambda_j \sim N(0, \psi_j \tau), \quad \psi_j \overset{i.i.d}{\sim} g, \quad \tau \sim f$$

- Many penalized least squares estimators correspond to mode of a Bayesian posterior (e.g., $L_1 \equiv$ Laplace prior)

Global-local scale mixtures of Gaussians (Polson & Scott, 2010)

- Essentially all shrinkage priors can be represented as

$$\Lambda_j \overset{ind}{\sim} N(0, \psi_j \tau), \quad \psi_j \overset{i.i.d}{\sim} g, \quad \tau \sim f$$

Global-local scale mixtures of Gaussians (Polson & Scott, 2010)

- Essentially all shrinkage priors can be represented as

$$\Lambda_j \overset{ind}{\sim} N(0, \psi_j \tau), \quad \psi_j \overset{i.i.d}{\sim} g, \quad \tau \sim f$$

- τ - global shrinkage toward zero, ψ_j's - avoid over-shrinking signals locally

Global-local scale mixtures of Gaussians (Polson & Scott, 2010)

- Essentially all shrinkage priors can be represented as

$$\Lambda_j \overset{ind}{\sim} N(0, \psi_j \tau), \quad \psi_j \overset{i.i.d}{\sim} g, \quad \tau \sim f$$

- τ - global shrinkage toward zero, ψ_j's - avoid over-shrinking signals locally
- g exponential = (Bayesian Lasso, Park & Casella, 2008; Hans, 2009)

Global-local scale mixtures of Gaussians (Polson & Scott, 2010)

- Essentially all shrinkage priors can be represented as

$$\Lambda_j \overset{ind}{\sim} N(0, \psi_j \tau), \quad \psi_j \overset{i.i.d}{\sim} g, \quad \tau \sim f$$

- τ - global shrinkage toward zero, ψ_j's - avoid over-shrinking signals locally
- g exponential = (Bayesian Lasso, Park & Casella, 2008; Hans, 2009)
- g inverse-gamma = (RVM, Tipping, 2001)

Global-local scale mixtures of Gaussians (Polson & Scott, 2010)

- Essentially all shrinkage priors can be represented as

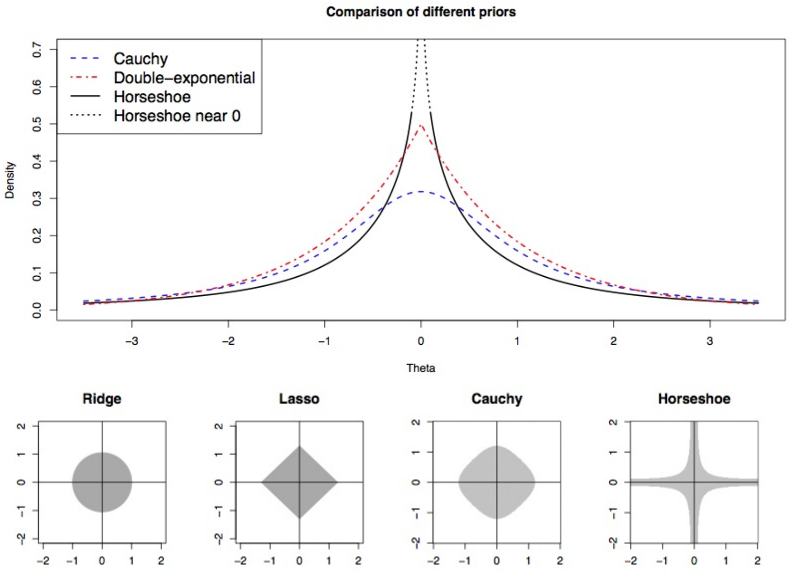
$$\Lambda_j \stackrel{ind}{\sim} N(0, \psi_j \tau), \quad \psi_j \stackrel{i.i.d}{\sim} g, \quad \tau \sim f$$

- τ - global shrinkage toward zero, ψ_j's - avoid over-shrinking signals locally
- g exponential = (Bayesian Lasso, Park & Casella, 2008; Hans, 2009)
- g inverse-gamma = (RVM, Tipping, 2001)
- g half-Cauchy = (Carvalho et al., 2009)

Shrinkage priors

- Appealing computationally & philosophically to relax assumption of exact zeros
- Rich literature on continuous shrinkage priors - student-t (T 01), normal/Jeffreys (BM 04), Laplace (Bayes Lasso) (PC 08, H 09), horseshoe (CPS 09), normal-gamma (GB 10, 12), generalized double Pareto (ADL 12), bridge (PSW 12) etc
- Many penalized least squares estimators correspond to mode of a Bayesian posterior (e.g., $L_1 \equiv$ Laplace prior)

Global-local priors



Common choices of the kernel $\mathcal{K}$ & associated penalty functions

Global-local priors

- Scale mixtures of Gaussians appealing computationally - block update possible
- However, understanding of such priors limited
- How to evaluate and compare such shrinkage priors relative to point-mass mixture priors ?
- Marginal properties not enough

Posterior rate: Pati, Bhattacharya, P., Dunson, 2014

- Theorem (Posterior Rate)

  For most global-local shrinkage priors defined as above, with $\epsilon_n = \sqrt{\frac{(\log p_n)^3 s}{n}}$*, we have*

$$\lim_{n \to \infty} \mathbb{P}_{\Sigma_0}(\|\Sigma - \Sigma_0\|_2 \geq \epsilon_n | \mathrm{Data}) \neq 0.$$

Posterior rate: Pati, Bhattacharya, P., Dunson, 2014

- Theorem (Posterior Rate)

  For most global-local shrinkage priors defined as above, with $\epsilon_n = \sqrt{\frac{(\log p_n)^3 s}{n}}$, *we have*

$$\lim_{n \to \infty} \mathbb{P}_{\Sigma_0}(\|\Sigma - \Sigma_0\|_2 \geq \epsilon_n | \mathrm{Data}) \neq 0.$$

  - Questions:
    1. What is the minimax rate for estimating Σ?$= \sqrt{\frac{(\log p_n)^3 s}{n}}$
    2. What prior on the vector Λ leads to a posterior which concentrates at the minimax rate? $=$ Point mass priors achieve this! Most global-local priors do NOT!

Statistical Inefficiency of Global-local priors

- What goes wrong? Two things:
  1. *A priori* independence of coordinates: inefficient shrinkage, a la Stein.
  2. Concentration of Measure.

Statistical Inefficiency of Global-local priors

- What goes wrong? Two things:
  1. *A priori* independence of coordinates: inefficient shrinkage, a la Stein.
  2. Concentration of Measure.
- As before:

$$\Lambda_j \overset{ind}{\sim} N(0, \psi_j \tau), \quad \psi_j \overset{i.i.d}{\sim} g, \quad \tau \sim f$$

Statistical Inefficiency of Global-local priors

- What goes wrong? Two things:
  1. *A priori* independence of coordinates: inefficient shrinkage, a la Stein.
  2. Concentration of Measure.
- As before:

$$\Lambda_j \stackrel{ind}{\sim} N(0, \psi_j \tau), \quad \psi_j \stackrel{i.i.d}{\sim} g, \quad \tau \sim f$$

- Local scales ψ are apriori independent; thus no *a priori* borrowing of information across coordinates, needed for efficient shrinkage estimators!

Statistical Inefficiency of Global-local priors

- What goes wrong? Two perspectives:
  1. A priori independence of coordinates: inefficient shrinkage, a la Stein.
  2. Concentration of Measure.

Statistical Inefficiency of Global-local priors

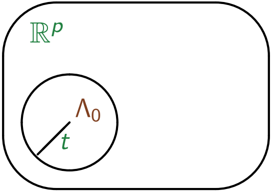
- What goes wrong? Two perspectives:
  1. A priori independence of coordinates: inefficient shrinkage, a la Stein.
  2. Concentration of Measure.
- In constructing prior distributions, we have to make sure that the prior gives sufficient mass around the "true parameter".

Statistical Inefficiency of Global-local priors

- What goes wrong? Two perspectives:
  1. A priori independence of coordinates: inefficient shrinkage, a la Stein.
  2. Concentration of Measure.
- In constructing prior distributions, we have to make sure that the prior gives sufficient mass around the "true parameter".
- Joint concentration $\mathbb{P}(\|\Lambda - \Lambda_0\|_2 \leq t)$ crucial for sparse Λ_0

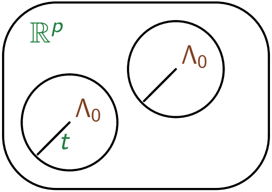
Statistical Inefficiency of Global-local priors

- Need joint concentration $\mathbb{P}(\|\Lambda - \Lambda_0\|_2 \leq t)$ crucial for sparse θ_0



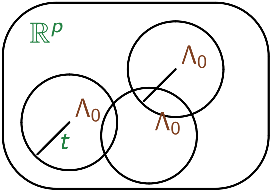
Statistical Inefficiency of Global-local priors

- Need joint concentration $\mathbb{P}(\|\Lambda - \Lambda_0\|_2 \leq t)$ crucial for sparse θ_0



Statistical Inefficiency of Global-local priors

- Need joint concentration $\mathbb{P}(\|\Lambda - \Lambda_0\|_2 \leq t)$ crucial for sparse θ_0



Prior concentration - some initial examples

- Recall The truth $\Lambda_0 \in \mathbb{R}^p$: with at most s non-zero elements.

Prior concentration - some initial examples

- Recall The truth $\Lambda_0 \in \mathbb{R}^p$: with at most s non-zero elements.
- Focus: Need non-asymptotic concentration bounds for

$$\mathbb{P}(\|\Lambda - \Lambda_0\|_2 < \sqrt{p})$$

Prior concentration - some initial examples

- Recall The truth $\Lambda_0 \in \mathbb{R}^p$: with at most s non-zero elements.
- Focus: Need non-asymptotic concentration bounds for

$$\mathbb{P}(\|\Lambda - \Lambda_0\|_2 < \sqrt{p})$$

- If Λ_j's are i.i.d. $\mathrm{N}(0, 1)$, then

$$\mathbb{P}(\|\Lambda - \Lambda_0\|_2 < \sqrt{p}) \leq e^{-Cp}$$

Prior concentration - some initial examples

- Recall The truth $\Lambda_0 \in \mathbb{R}^p$: with at most s non-zero elements.
- Focus: Need <u>non-asymptotic</u> concentration bounds for

$$\mathbb{P}(\|\Lambda - \Lambda_0\|_2 < \sqrt{p})$$

- If Λ_j's are i.i.d. $\mathrm{N}(0,1)$, then

$$\mathbb{P}(\|\Lambda - \Lambda_0\|_2 < \sqrt{p}) \leq e^{-Cp}$$

- On the other hand, for suitable point mass priors (g Laplace)

$$\mathbb{P}(\|\Lambda - \Lambda_0\|_2 < \sqrt{p}) \geq e^{-Cs \log p}$$

Prior concentration

- KEY RESULT: Most continuous shrinkage priors give poor concentration

Prior concentration

- KEY RESULT: Most continuous shrinkage priors give poor concentration
- Bayesian LASSO:

$$\mathbb{P}(\|\Lambda - \Lambda_0\|_2 < \sqrt{p}) \leq e^{-C\sqrt{p}}$$

Prior concentration

- KEY RESULT: Most continuous shrinkage priors give poor concentration
- Bayesian LASSO:

$$\mathbb{P}(\|\Lambda - \Lambda_0\|_2 < \sqrt{p}) \leq e^{-C\sqrt{p}}$$

- Thus the concentration improves only a little.

Dirichlet Laplace prior & properties

- Next key idea: Can we induce dependence across local scales?
- We propose a simple dependent modification leading to optimal concentration & efficient computation

$$\Lambda_j \sim \mathrm{Double\,Exp}(\psi_j \tau)$$

- IDEA: Constrain ψ to the simplex - this allows for dependence
- We let $\psi \sim \text{Diri}(\alpha, \ldots, \alpha)$ - $\alpha < 1$ favors small # dominant values with remaining ≈ 0.

Dirichlet Laplace prior & properties

- $\psi \sim \text{Diri}(\alpha, \ldots, \alpha)$ with $\alpha < 1$ favors small # dominant values with remaining ≈ 0
- Induced marginal of $\Lambda_j \propto |\Lambda_j|^{\alpha/2-1} K_{1-\alpha}(\sqrt{2|\Lambda_j|})$, where $K_\nu(\cdot)$ modified Bessel function of second kind
- Spike at zero controlled by α - use $U(0,1)$ prior
- Tune α to incorporate prior knowledge about sparsity
- Similar to horseshoe prior marginally
- $\tau \sim \mathrm{Ga}(p\alpha, 1/2)$.

Recall: for suitable point mass priors (g Laplace)

$$\mathbb{P}(\|\Lambda - \Lambda_0\|_2 < \sqrt{p}) \geq e^{-Cs \log p}$$

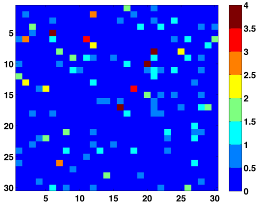
A Key Result.

Theorem

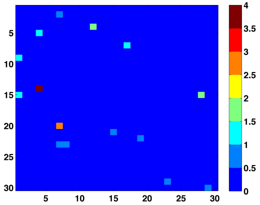
Dirichlet-Laplace prior distributions have similar concentration as the point mass prior distributions.

Prior draws

- $\alpha = 1/2$



- $\alpha = 1/10$



Novel sampling scheme for $(\phi \mid \Lambda)$

- Normal scale mixture rep: $\Lambda_j \sim \mathrm{N}(0, \psi_j \phi_j^2 \tau^2)$, $\psi_j \sim \mathrm{Exp}(1/2)$
- $\phi = (\phi_1, \ldots, \phi_p)^{\mathrm{T}} \sim \mathrm{Dir}(\alpha, \alpha, \ldots, \alpha)$
- generalized inverse Gaussian: $Y \sim giG(\lambda, \rho, \chi)$

$$f_Y(y) \propto y^{\lambda-1} e^{-\frac{1}{2}(\rho y + \chi/y)}$$

- Theorem

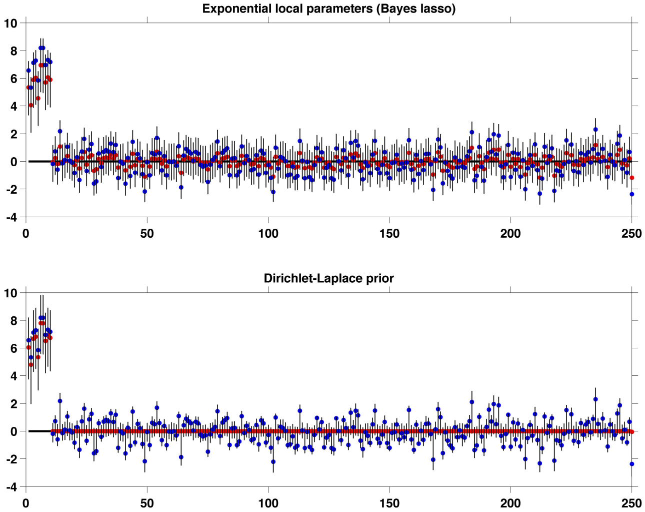
  The joint posterior of $\boxed{\phi \mid \Lambda \stackrel{d}{=} (T_1/T, \ldots, T_p/T)}$, *where* $T_j \sim giG(\alpha - 1, 1, 2|\theta_j|)$ *independently.*

  - Makes block update possible - highly efficient Gibbs sampler

Minimax rate: Pati, Bhattacharya, P., Dunson, 2013

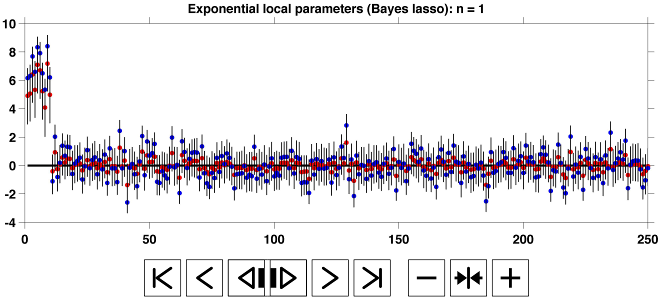
- Questions:
  1. What is the minimax rate for estimating Σ? $= \sqrt{\frac{(\log p_n)^3 s}{n}}$
  2. What prior on the vector Λ leads to a posterior which concentrates at the minimax rate? $=$ Point mass priors achieve this! The Dirichlet-Laplace Prior above also achieves this!

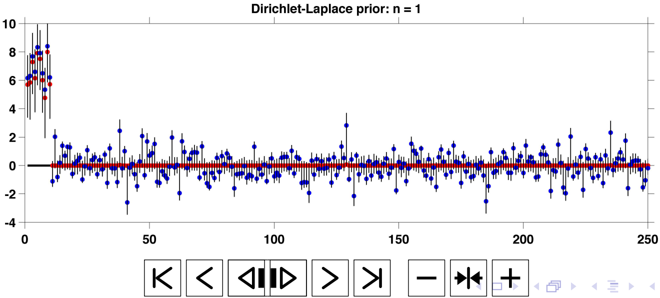
Improved prior concentration reflected in the posterior



Draw $y \sim \mathrm{N}_{250}(\theta_0, \mathrm{I}_{250})$ with $\theta_0[1:10] = 7$, $\theta_0[11:250] = 0$. Blue dots: entries of y, red dots: posterior median of θ, bars: point wise 95% credible intervals

Increase sample size





Now, have we solved the problem?

- Not completely!
- There are different MCMC algorithms for posterior sampling.
- The only commonly used measure so far is the "effective sample" size.
- Hard to get exact bounds theoretically for most examples!

The Divide-and-Conquer Framework

- Basic Idea -
  - divide the high-dimensional data into low dimensional subproblems
  - solve the subproblems in parallel using existing MCMC techniques
  - combine the estimates to produce a global estimate of the covariance matrix
- Other divide-and-conquer approaches in the literature focus on tackling "large n" problems where the data are assumed to be independent and identically distributed (Mackey et al. 2011, Zhang et al. 2013, Minsker et al. 2014, Cheng & Shang 2015)

Divide step

Randomly partition $\mathbf{y_i} \in \mathbb{R}^p$ into g p_g-dimensional subvectors, $\{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(g)}\}$ where $\mathbf{y}_i^{(m)} \in \mathbb{R}^{p_g}, \quad m = 1, \ldots, g$ and $p_g = p/g$

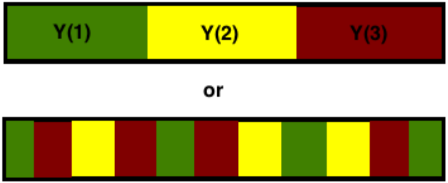


Figure : Y_i is partitioned into 3 groups, namely, $\mathbf{Y}(1)$, $\mathbf{Y}(2)$ and $\mathbf{Y}(3)$

Fit step -

- Fit factor models to the group m for $m = 1, \ldots, g$ as

$$\mathbf{y_i^{(m)}} = \Lambda^{(m)} \eta_{\mathbf{i}}^{\mathbf{(m)}} + \epsilon_{\mathbf{i}}^{\mathbf{(m)}}, \quad \epsilon_{\mathbf{i}}^{\mathbf{(m)}} \sim \mathrm{N}(0, \Omega^{(m)}).$$

  and obtain posterior distribution of $\Sigma^{(m)} \in \mathbb{R}^{p_g \times p_g}$ based on a shrinkage prior on $(\Lambda^{(m)}, \Omega^{(m)})$ conditional on the latent factors $\eta_i^{(m)}$.

Conquer Step

How to combine estimates from different groups to form a global estimator for the covariance matrix?

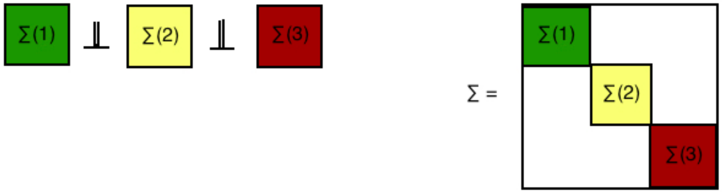


Figure : The task is divided across 3 groups/machines and the estimates obtained from each subproblem are assumed to be independent

Inducing Dependence Via Factor Augmentation

- Consider the hierarchical model,

$$\eta_{\mathbf{i}}^{(\mathbf{m})} \mid \mathbf{X_i}, \mathbf{Z_i^{(m)}} = \sqrt{\rho}\,\mathbf{X_i} + \sqrt{1-\rho}\,\mathbf{Z_i^{(m)}},\ i = 1, \ldots, n,\ m = 1, \ldots, g$$

  where
  - $\mathbf{X_i} \sim N_{k_g}(0, I)$, is the component that is shared across all the latent sub-factors
  - $\mathbf{Z_i^{(m)}} \sim N_{k_g}(0, I)$ is the component that is idiosyncratic to the specific sub-factor
  - ρ is the correlation induced between the latent sub-factors. $\rho \sim U(0, 1)$ is a convenient choice

Inducing Dependence Via Factor Augmentation

The hierarchical structure has two distinct advantages:

- it induces a correlation structure among sub-estimates $\hat{\Sigma}^{(m)}$ in the conquer step (Lemma 1)
- it does not increase the computational complexity of the algorithm
- $Cov\{\mathbf{Y_i^{(m)}}, \mathbf{Y_i^{(m')}}\} = \rho\Lambda^{(m)}\Lambda^{(m')}$.

Conquer step I

The estimate for the original covariance matrix Σ is obtained using $\Sigma_E = DED^{\mathrm{T}} + \Omega$, where $D = \mathrm{diag}\{\Lambda^{(1)}, \cdots, \Lambda^{(g)}\}$, $\Omega = \mathrm{diag}\{\Omega^{(1)}, \cdots, \Omega^{(g)}\}$, $E = I_{k_g} \otimes C$ for a $g \times g$ positive definite matrix C such that $C_{mm'} = 1$ if $m = m'$ and $C_{mm'} = \rho$ if $m \neq m'$

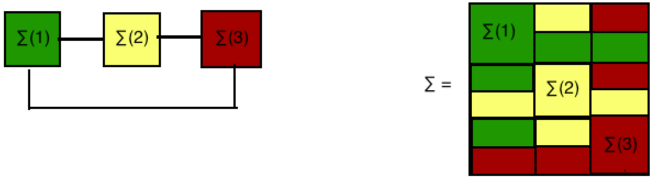


Figure : The task is divided across 3 groups/machines and the estimates obtained are pooled using the hierarchical framework

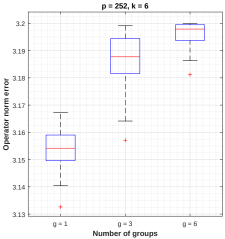
Conquer step II

For $g = 2$ groups, an estimate of the covariance matrix Σ is given by

$$\Sigma_E = \begin{bmatrix} \hat{\Lambda}^{(1)}\hat{\Lambda}^{(1)\mathrm{T}} + \hat{\Omega}^{(1)} & \rho\hat{\Lambda}^{(1)}\hat{\Lambda}^{(2)\mathrm{T}} \\ \rho\hat{\Lambda}^{(1)}\hat{\Lambda}^{(2)\mathrm{T}} & \hat{\Lambda}^{(2)}\hat{\Lambda}^{(2)\mathrm{T}} + \hat{\Omega}^{(2)} \end{bmatrix}$$

We have some theory to show to what extent is $\Sigma_E = DED^{\mathrm{T}} + \Omega$ is a good approximation to $\Sigma = \Lambda\Lambda^{\mathrm{T}} + \Omega$ where $\Lambda \in \mathbb{R}^{p\times k}$?

Sensitivity to Random Splitting I



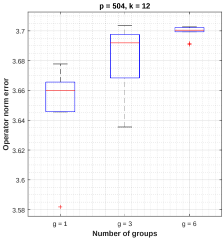


Figure : Results across 10 replicates for $n = 100$

Table : Comparison of Divide and Conquer ($g = 20$) with POET and Carvalho et al. across 1 simulation replicate for $p = 20,000$ and $k = 200$.

	Carvalho	DnC	POET
meanop	630.02	84.63	Fail
meanfro	3470.2	423.38	Fail
time	48890	9858	Fail

Paradigm Shift for Large Data Sets

- The computational complexity of the some of the commonly used MCMC algorithms are exponential in the data set.
- Need new theoretical framework for evaluating the efficiency MCMC algorithms with fixed computational complexity.
- Evaluate the efficiency of MCMC algorithms keeping the CPU time fixed- Widely open area!
- Can we MCMC algorithms which scale polynomially with the sample size (n) and/or the dimension of the parameter space (p)?

Summary

- Concrete formulation of the statistical efficiency vs. computational efficiency.
- Under mild conditions, efficient posterior convergence is possible even if $p \gg n$.
- Prior concentration very important - should give enough probability near sparse subspaces.
- Appropriate point mass mixture priors can achieve this - prior probability of subset size important
- Most continuous shrinkage priors do not achieve this.
- Also developed a continuous shrinkage prior which does indeed meets both the theoretical and computational efficiency criteria.
- Divide and conquer factor model seems to be a promising area, and worth exploring!

References

- Dirichlet-Laplace priors for optimal shrinkage (Bhattacharya, A., Pati. D., Dunson, D.B.), JASA 2014.
- Posterior Contraction Rates in Sparse Bayesian Models for Massive Covariance Matrices, (Bhattacharya, A., Pati. D., Dunson, D.B.), Annals of Statistics 2014.
- A Divide and Conquer Strategy for High Dimensional Bayesian Factor Models, (Gautam Sabnis, Debdeep Pati and Barbara Engelhardt), arXiv, 2017.

Acknowledgements

Danke!