A Divide and Conquer Approach to High Dimensional Bayesian Factor Models

Gautam Sabnis

Boston University

Outline

- Overview
- Covariance Matrix Estimation
  - Estimators for the covariance matrix
- Factor Models
  - Sparse factor models
  - Bayesian sparse fatent models
- Choice of Priors
- Posterior Computation: MCMC sampler
- A Divide and Conquer Framework
  - Divide
  - Fit
  - Conquer
- Numerical Experiments
- Application to Thyroid Gene Expression Data

Overview

- High-dimensional data is ubiquitous in modern applications
- Example: Omics study where $p =$ biological variables, $n =$ number of samples
- High-dimensional data: $p \gg n$
- Curse of dimensionality: classical statistical methods break down in such settings
- Dimensionality reduction is crucial

Covariance Matrix Estimation

- Goal: estimate a covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$
- given i.i.d. samples $y_i \sim N(0, \Sigma)$, for $i = \ldots, n$

Why covariance estimation?

- Principal Component Analysis (PCA)
- Linear/Quadratic Discriminant analysis (LDA/QDA)
- Generalized Method of Moments (GMM)
- Generalized Least Squares (GLS)
- Canonical Correlation Analysis (CCA)

Estimators for Covariance Matrix

Classical approach

Estimate Σ via sample covariance matrix

$$\hat{\Sigma}_{\mathrm{n}} = \frac{1}{n} \sum_{i=1}^{n} y_i y_i^T$$

The performance of $\hat{\Sigma}_{\mathrm{n}}$ degrades as $c = n/p \to 0$.

Structured Estimation

Great interest in regularized estimation
(Ledoit & Wolf, 2003b; Bickel & Levina 2008 a,b; Cai and Zhou, 2011 ...)

Low intrinsic dimensionality
(Johnstone, 2001; Bhattacharya & Dunson, 2011; Fan et al., 2013 ...)

Factor Models

- Highly successful approach for dimensionality reduction
- Relate high-dimensional $\mathbf{y}$ to $\eta \in \mathbb{R}^k$ through

$$\mathbf{y_i} = \Lambda \eta_\mathbf{i} + \epsilon_\mathbf{i}, \quad \epsilon_\mathbf{i} \sim N(0, \Omega_p)$$

- $\Lambda \in \mathbb{R}^{p \times k}$ is a tall-skinny (k << p) factor loadings matrix
- $\mathbb{E}(\eta) = 0$, $\mathrm{Var}(\eta) = \mathrm{I}_k$
- Assumptions: 1) Ω is a diagonal matrix, 2) ϵ is independent of η.

With assumptions (1) and (2),

$$\Sigma = \Lambda\Lambda^T + \Omega$$

which leads to dimension reduction

$$\mathcal{O}(p^2) \longrightarrow \mathcal{O}(pk + p)$$

Sparse Factor Models

- A *sparse factor model* (West 2003) is a factor model in which Λ is sparse.

$$\Lambda_{p\times k} = \begin{pmatrix} \lambda_{11} & 0 & \lambda_{13} & \dots & 0 \\ \lambda_{21} & \lambda_{22} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_{jh} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \lambda_{p3} & \dots & 0 \end{pmatrix}$$

- For example,

> $y =$ gene expression, $\eta =$ pathway expression

Bayesian Sparse Factor Models

Setup

- Observe $y_i \sim N(0, \Sigma),\ i = 1, \ldots, n$
- $\eta \sim N(0, \mathrm{I}_k)$
- $\Lambda \sim$ Prior distribution
- $\Omega \sim$ Prior distribution

Posterior distribution

$$\underbrace{\Pi(\Lambda, \eta, \Omega \mid y)}_{posterior} \propto \prod_{i=1}^{n} \underbrace{\mathrm{N}(y_i \mid \Lambda\eta_i, \Omega)}_{Likelihood} \underbrace{\pi(\Lambda)\pi(\eta_i)\pi(\Omega)}_{priors}$$

Choice of Priors

- Multiplicative Gamma Process Shrinkage Prior (MGPS) (Bhattacharya & Dunson 2011)

$$\lambda_{jh} \mid \phi_{jh}, \tau_h \sim N(0, (\phi_{jh}\tau_h)^{-1}), \quad \phi_{jh} \sim \Gamma(\nu/2, \nu/2), \quad \tau_h = \prod_{l=1}^{h} \delta_l$$

$$\delta_1 \sim \Gamma(a_1, 1), \quad \delta_l \sim \Gamma(a_2, 1),\ l \geq 2, \quad \omega_j^{-2} \sim \Gamma(a_\omega, b_\omega)$$

where

- τ_h: global shrinkage parameter
- ϕ_{jh}: local shrinkage parameter
- $\Omega = \text{diag}(\omega_1^2, \ldots, \omega_p^2)$

Posterior Computation

MCMC Sampler

If the current state is $(\Lambda_t, \eta_t, \Omega_t)$ at the t^{th} iteration, then update the parameter values by cycling through the following steps,

- Sample $\eta_{(t+1)} \sim N_k(\mu_\eta, \Sigma_\eta \mid \Lambda_t, \Omega_t)$
- Sample $\Lambda_{j_{(t+1)}} \sim N_k(\mu_\Lambda, \Sigma_\Lambda \mid \eta_{(t+1)}, \Omega_t)$
- Sample $\omega^{-2}_{(t+1)} \sim \Gamma(a_\omega, b_\omega \mid \Lambda_{(t+1)}, \eta_{(t+1)})$
- Sample $\Phi_{j_{(t+1)}} \sim \Gamma(a_\phi, b_\phi \mid \Lambda_{(t+1)}, \delta_t)$
- Sample $\delta_{(t+1)} \sim \Gamma(a_\delta, b_\delta \mid \Lambda_{(t+1)}, \Phi_{(t+1)})$
- Sample the hyperparameters a_1 and a_2 using a Metropolis-Hastings step within the sampler.

Computational & Storage Constraints

The MCMC sampler, at each iteration,

- performs several matrix operations such as matrix multiplication, matrix inversion, Cholesky decomposition
- stores Λ for updating η and vice versa

Matrix Operation	Example	Complexity
Multiplication	$\Lambda^T\Lambda$	$\mathcal{O}(pk^2)$
Inversion	$(\Lambda^T\Lambda)^{-1}$	$\mathcal{O}(k^3)$
Cholesky Decomposition	$\Lambda^T\Lambda$	$\mathcal{O}(k^3)$

> One iteration of the Gibbs sampler for estimating Σ requires $O(k^3 + npk + nk^2 + pk^2)$ floating point operations for matrix computations, and a storage complexity of $O(pk + k^2)$.

Divide-and-Conquer

- Basic idea
  - divide the high-dimensional data into low dimensional subproblems
  - solve the subproblems in parallel using existing MCMC samplers
  - combine the estimates to produce a global estimate of the covariance matrix
- 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)

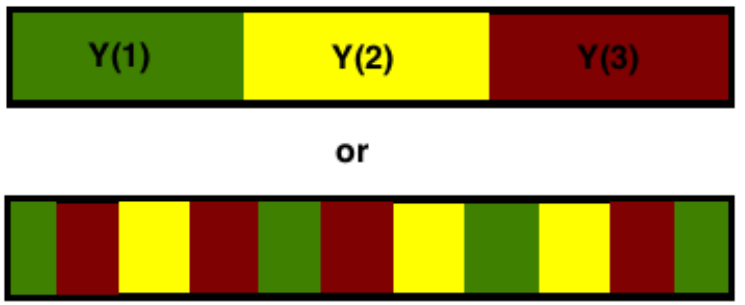
> We split dimension p across the subproblems.

Divide step

Randomly partition $y_i \in \mathbb{R}^p$

$$y_i \longrightarrow (y_i^{(1)}, \ldots, y_i^{(g)}), \quad i = 1, \ldots, n$$

so that $y_i^{(m)} \in \mathbb{R}^{p_g}$ where $p_g = p/g$ and $m = 1, \ldots, g$.



Fit Step

On machine m,

$$y_i^{(m)} = \Lambda^{(m)}\ \eta_i^{(m)} + \epsilon_i^{(m)}, \quad i = 1, \ldots, n,$$

where $\Lambda^{(m)} \in \mathbb{R}^{p_g \times k_g}$, $\eta_i^{(m)} \in \mathbb{R}^{k_g}$ and $\epsilon_i^{(m)} \in \mathbb{R}^{p_g}$.

$$\underbrace{\Pi(\Lambda^{(m)}, \eta^{(m)}, \Omega^{(m)} \mid y)}_{posterior} \propto \prod_{i=1}^{n} \underbrace{\mathrm{N}(y_i^{(m)} \mid \Lambda^{(m)} \eta_i^{(m)}, \Omega^{(m)})}_{Likelihood} \underbrace{\pi(\Lambda^{(m)}) \pi(\eta_i^{(m)}) \pi(\Omega^{(m)})}_{priors}$$

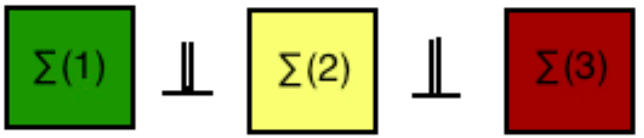
Run the Gibbs sampler to obtain posterior estimate of $\Sigma^{(m)} \in \mathbb{R}^{p_g \times p_g}$ via

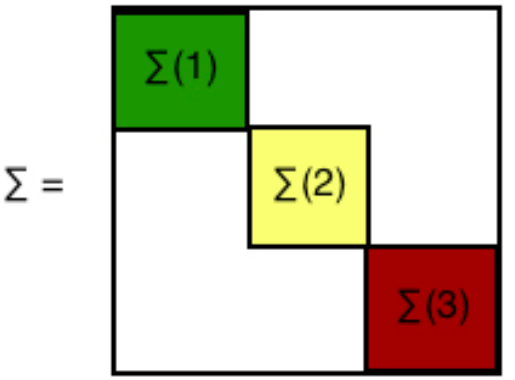
$$\hat{\Sigma}^{(m)} = \hat{\Lambda}^{(m)} \hat{\Lambda}^{(m)T} + \hat{\Omega}^{(m)}$$

Computational Gains

Model	Complexity
Original Fit step	$O(k^3 + npk + nk^2 + pk^2)$
Fit step in DNC	$O((k/g)^3 + n(pk/g^2) + n(k/g)^2 + (pk^2/g^3))$

How to combine estimates? How to Conquer?





Inducing Dependence

Consider the hierarchical model

$$\eta_i^{(m)} = \sqrt{\rho} X_i + \sqrt{1-\rho} Z_i^{(m)}$$

where

- $X_i \sim N(0, I_{k_g})$: shared component
- $Z_i^{(m)} \sim N(0, I_{k_g})$: machine-specific component (not shared)
- $0 < \rho < 1$

> If $\mathrm{Cov}\left(Z_i^{(m)}, Z_i^{(m')}\right) = 0$, then $\mathrm{Cov}\left(y_i^{(m)}, y_i^{(m')}\right) = \rho \Lambda^{(m)} \Lambda^{(m')}$.

Conquer Step

The covariance matrix is given by

$$\Sigma = \Lambda_B \Lambda_B^T + (1-\rho)\left(\text{diag}\{\Lambda^{(1)}\Lambda^{(1)T}, \ldots, \Lambda^{(g)}\Lambda^{(g)T}\} - \Lambda_B \Lambda_B^T\right) + \check{\Omega}$$

where $\Lambda_B^T = (\Lambda^{(1)}, \ldots, \Lambda^{(g)}) \in \mathbb{R}^{\frac{k}{g} \times p}$ and $\check{\Omega} = \text{diag}(\Omega^{(1)}, \ldots, \Omega^{(m)})$.

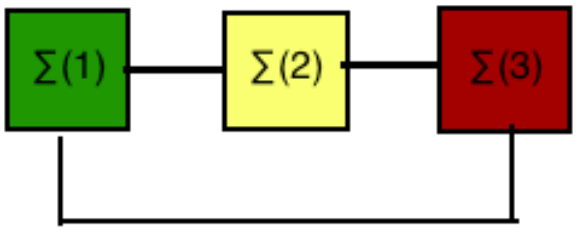
Notes

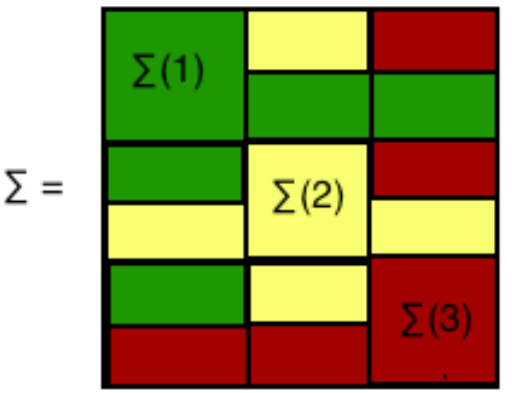
- To sample from $\eta^{(m)} \sim N_k(\mu_\eta, \Sigma_\eta \mid \Lambda^{(m)}, \Omega^{(m)})$, draw samples from
  - $X \sim N_k(\mu_\eta, \Sigma_\eta \mid \Lambda^{(m)}, \Omega^{(m)}, Z^{(m)})$
  - $Z^{(m)} \sim N_k(\mu_\eta, \Sigma_\eta \mid \Lambda^{(m)}, \Omega^{(m)}, X)$

Conquer Step: Illustration

For $g = 3$, an estimate of Σ is given by

$$\Sigma = \begin{pmatrix} \Lambda^{(1)}\Lambda^{(1)T} + \Omega^{(1)} & \rho\Lambda^{(1)}\Lambda^{(2)T} & \rho\Lambda^{(1)}\Lambda^{(3)T} \\ \rho\Lambda^{(2)}\Lambda^{(1)T} & \Lambda^{(2)}\Lambda^{(2)T} & \rho\Lambda^{(2)}\Lambda^{(3)T} \\ \rho\Lambda^{(3)}\Lambda^{(1)T} & \rho\Lambda^{(3)}\Lambda^{(2)T} & \Lambda^{(3)}\Lambda^{(3)T} + \Omega^{(3)} \end{pmatrix}$$





Numerical Experiments

Experiments performed on simulated data using Flux available at University of Michigan.

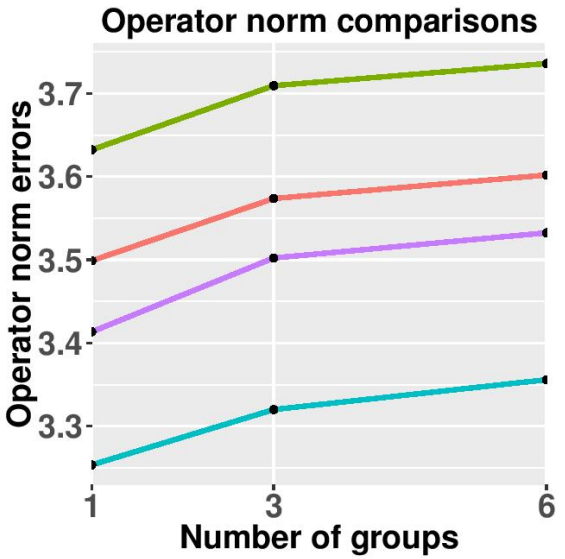
Setup

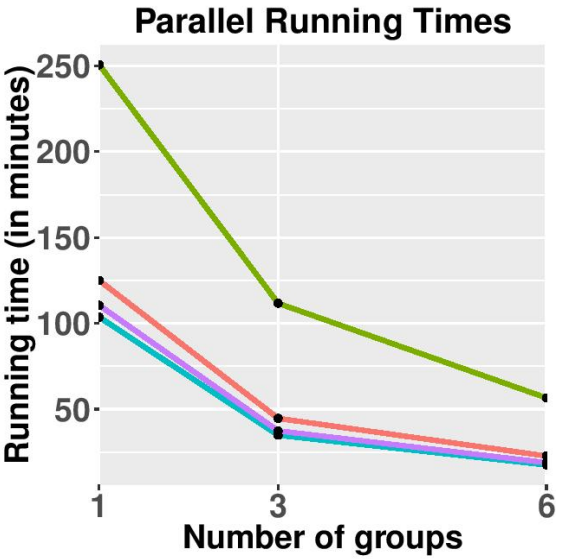
$$y_i = \Lambda \eta_i + \epsilon_i, \quad i = 1, \ldots, n$$

- $\epsilon \sim N(0, \Omega)$, $\Omega = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$, $\sigma^{-2} \sim \Gamma(1, 0.25)$
- $\Lambda_h \in \mathbb{R}^p$, $h = 1, \ldots, k$ is such that $|\Lambda_h| = \log(p)$
- $\Lambda_{jh} \sim U(0.1, 3)$
- $n = 100$ for all choices of $\{p, k\}$ and g
- Number of replications = 100
- Comparison in terms of
  - operator norm $\|\cdot\|_2$ where $\|A\|_2 = s_{\max}(A)$,
  - frobenius norm $\|\cdot\|_F$ where $\|A\|_F = \sqrt{\text{tr}(A^T A)}$
  - computation time, in minutes, per replication.

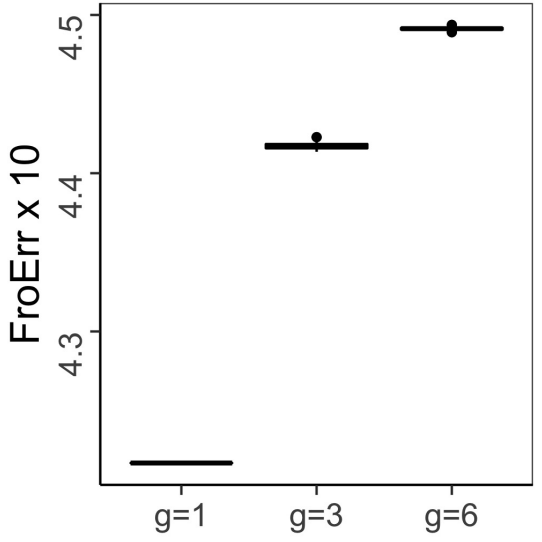
Statistical Accuracy vs Computational Gain

- $(p, k) = \{(252, 6), (504, 12), (1008, 24), (2016, 36)\}$
- $g = \{1, 3, 6\}$

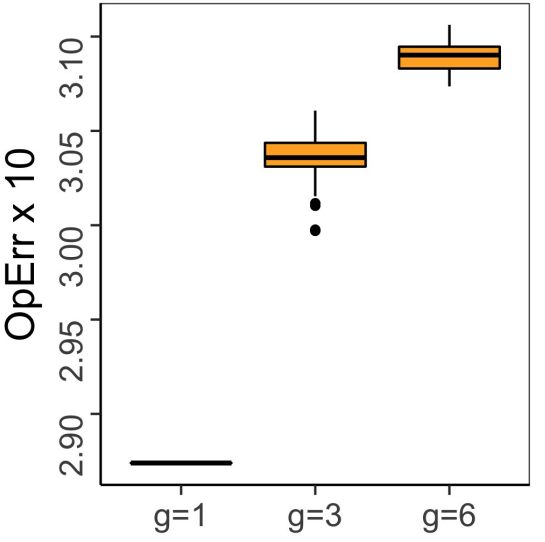




Sensitivity to Random splitting

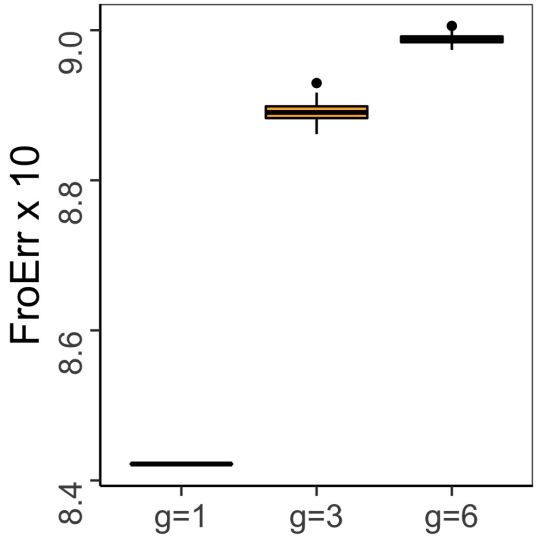


Frobenius norm errors for p=1008

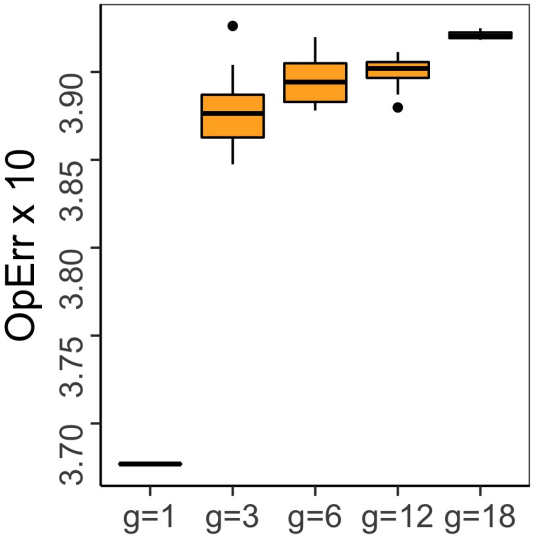


Operator norm errors for p=1008

Sensitivity to Random splitting



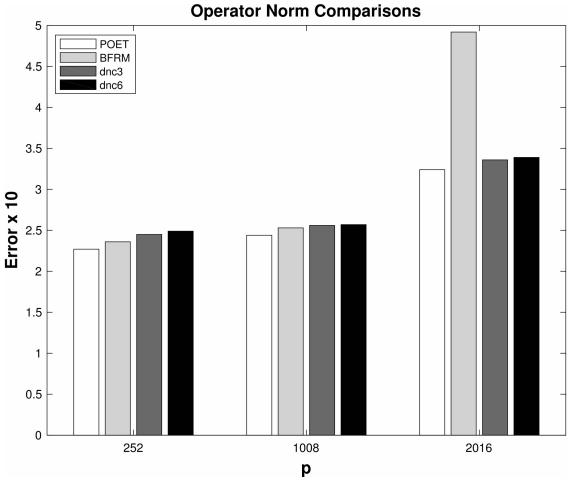
Frobenius norm errors for p=2016

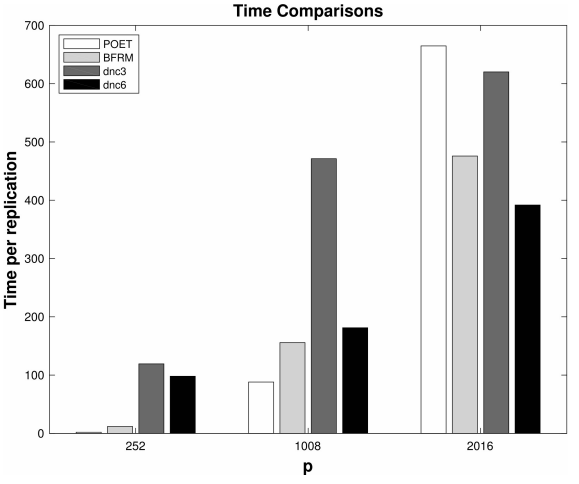


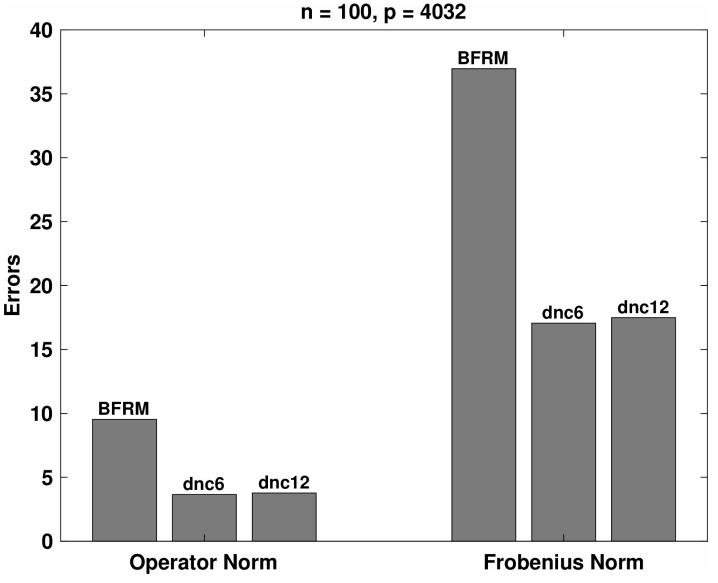
Operator norm errors for p=2016

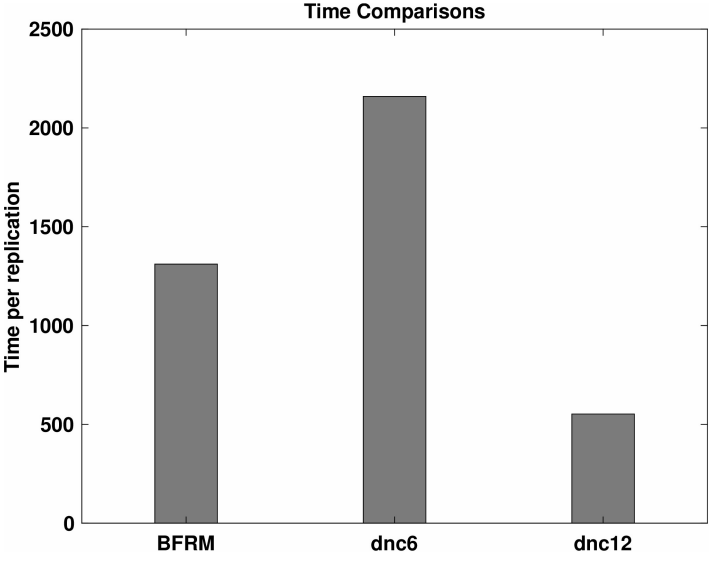
Results

- Compare Divide and Conquer **(DNC)** with Principal Orthogonal Complement Thresholding **(POET)** and Bayesian Factor Regression Modeling **(BFRM)** in terms of statistical accuracy.
- $n = 100, (p, k) = \{(252, 6), (1008, 24), (2016, 36)\}, g = \{3, 6\}$









- **POET** runs into memory issues for $p = 4032$.

p	k	Method	Error	Time
10^4	100	DNC_1	Fail	Fail
10^4	100	POET	Fail	Fail
10^4	100	BFRM	Inaccurate	Inaccurate
10^4	100	DNC_{10}	66.28 (0.09)	5 days
10^4	100	DNC_{20}	69.55 (0.16)	3 days

Application to Thyroid Gene Expression Data

Dysregulation of gene expressions in the thyroid is associated with differnt types of metabolic and autoimmune diseases and thyroid cancer.

Data

- 17,908 gene expressions of thyroid tissue samples from 278 donors
- Access to additional information such as time of death and sample collection

Goal

Estimate the covariance matrix in order to understand the second order structure among these gene expressions i.e. how they co-vary.

Analysis

- Data matrix: $Y \in \mathbb{R}^{278 \times 17{,}908}$
- We estimate the covariance matrix to obtain $\hat{\Sigma}$ using $g = 11$ with $p_g = 1,628$ genes.
- We ran the Gibbs sampler for $15,000$ iterations with $5,000$ burn-in and collected every first sample after burn-in to thin the chain.
- The number of factors were set to $k = 110$.
- The 10 estimated factors across 11 groups are nearly identical, highlighting that each group finds similar estimates of the latent subspace despite using disjoint groups of genes.

Analysis

- Entries in Λ were thresholded to cluster genes within factors
- A Gene ontology enrichment analysis (Eden et al., 2009) showed that
  - many of the clustered genes in the first factor are related to cellular death processes
  - clustered genes contained in many other factors are involved in cellular metabolic processes (factors 3-6, 15, 22)

Further analysis showed correlation of factors with ischemic time. This along with enrichment of genes involved in processes related to cellular metabolism and death suggested that covariation in the data is driven by the rapid decay of tissue samples postmortem.

Summary

- A Divide and Conquer strategy is proposed to accelerate posterior inference for high-dimensional covariance matrix estimation in Bayesian sparse factor models is proposed.
- The approach is different from the existing divide and conquer strategies in that each subproblem includes all n samples and, instead, splits problem across dimensions.
- Numerical studies demonstrate the computational speedups offered at the cost of minimial loss is accuracy.
- Application to gene expression data reveals certain interesting biological processes but more research is needed to understand what is jointly regulating the expressed genes.
- Theoretical guarantees in terms of some common functionals of the covariance matrix.

Acknowledgements

- Debdeep Pati (Texas A&M University)
- Barbara Engelhardt (Princeton University)
- Natesh Pillai (Harvard University)

Thanks for your patience and attention!