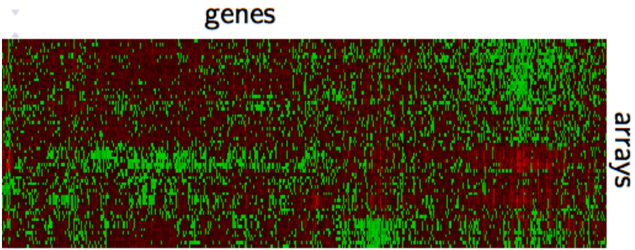
High-dimensional Bayes

Debdeep Pati
Florida State University

November 8, 2016

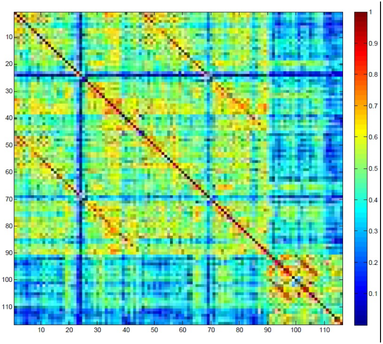
Motivating application - high-dim regression



- $y_i \in \mathbb{R}$ & $x_i = (x_{i1}, \ldots, x_{ip})' \in \mathcal{X} \subset \mathbb{R}^p$, $i = 1, \ldots, n$
- n = sample size, p=number of predictors & $p \gg n$
- $y_i = x_i^{\mathrm{T}}\beta + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2)$
- In big data problems, dimensionality reduction is crucial
- *sparsity* in β

Motivating application: Autism spectra-matrix

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



- Understand these patterns

Cov matrix estimation by Gaussian factor models

- $y_i = (y_{i1}, \ldots, y_{ip})^{\mathrm{T}}, i = 1, \ldots, n$ with $n \ll p$

$$y_i = \Lambda \eta_i + \epsilon_i, \quad \epsilon_i \sim N_p(0, \sigma^2 I_p), \quad i = 1, \ldots, n$$

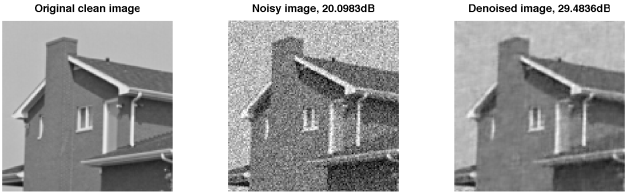
- $\eta_i \in \mathbb{R}^k$ latent factors, Λ a $p \times k$ matrix of factor loadings with $k \ll p$
- With $\eta_i \sim \mathrm{N}_k(0, \mathrm{I}_k)$, $y_i \sim \mathrm{N}_p(0, \Omega)$ with $\Omega = \Lambda\Lambda^{\mathrm{T}} + \sigma^2 I_p$.
- Unstructured Ω has $O(p^2)$ free elements
- Regularized estimation of Ω via parsimonious factorization
- Still $pk + 1$ parameters, crucial to assume $\Lambda(:, h)$ are sparse
- Connection to sparse PCA (Zou, Hastie & Tibshirani, 2006)

Image denoising using Dictionary learning

- Closely related to sparse factor modeling approach
- $x_i \in \mathbb{R}^D, i = 1, \ldots, N$ - image patches, functional data etc
- Instead of using a fixed basis - try to learn a dictionary

$$x_i = \Theta\eta_i + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2 \mathrm{I}_D)$$

- $\Theta \in \mathbb{R}^{D \times K}$ - unknown dictionary ($K \gg D$ usually)
- η_i *sparse* coefficient vector



How to do inference in $p \gg n$ setting ?

- Clearly classical methods such as maximum likelihood estimation break down (Stein, 1955; James & Stein, 1961)
- Most common choice is to use *regularization* or thresholding
- Focus is on obtaining a sparse point estimate
- There is a vast literature on Lasso/L1 regularization (Tibshirani, 1996) and variants
- In the regression setting, minimize

$$\sum_{i=1}^{n}(y_i - x_i^T\theta)^2 + \tau\sum_{j=1}^{p}|\theta_j|$$

- Resulting $\widehat{\theta}$ contains exact zeros

Regularization

- Bridge (FF 93), SCAD (FL 01), Elastic net (ZH 05), Adaptive Lasso (Z 06) and many others
- Very rich applied & theoretical literature
- Regularization approaches for large covariance estimation
- banding/tapering (BL 08, WP 10), thresholding (BL 08, RLZ 09, CL 11), banding/penalizing Cholesky factor (WP 03, RLZ 10), regularized PCA (JL 09, HT 06) and many others

Posterior uncertainty

- Simply obtaining a point estimate is insufficient in many applications
- In small n, large p there will typically be substantial uncertainty in $\widehat{\theta}$
- We would like to characterize uncertainty in inferences about the impact of predictors & in predictions
- We start with a prior distribution $\pi(\theta)$
- Posterior distribution $\pi(\theta \mid y^n)$ provides a probabilistic characterization of uncertainty in θ

Bayesian sparsity priors

- Prior belief about sparsity in high-dim $\theta = (\theta_1, \ldots, \theta_p)^T$:

$$\boxed{\theta_j \sim (1 - \pi_0)\delta_0 + \pi_0 g(\cdot)}$$

- δ_0 = point mass at zero, so $\text{pr}(\theta_j = 0) = 1 - \pi_0$
- $g(\cdot)$ = prior density on the 'signal' coefficients
- Empirical Bayes to estimate π_0 (Johnstone & Silverman, 2004)
- $\pi_0 \sim \text{beta}(a, b)$ to allow uncertainty in model size (sparsity) (Scott & Berger, 2010)
- Minimax optimality of empirical Bayes & full posterior (Johnstone & Silverman, 2004, Castillo & van der Vaart, 2012)

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)

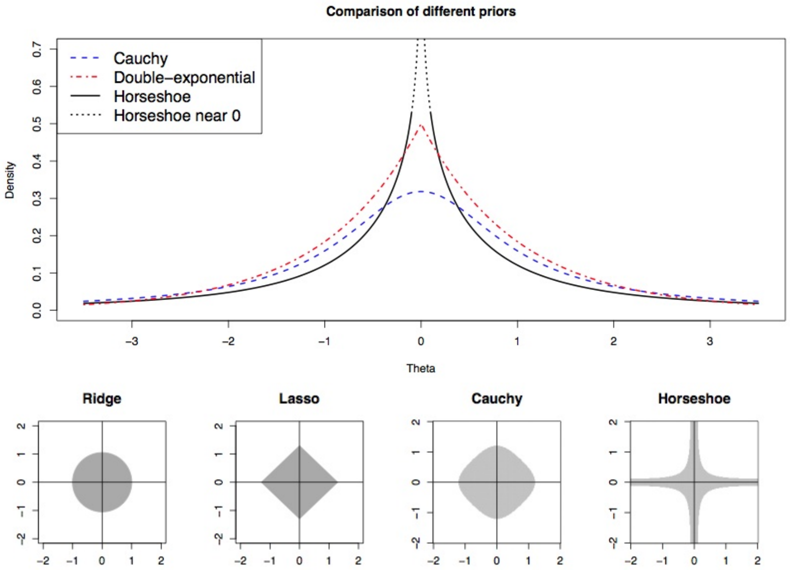
Global-local scale mixtures of Gaussians (Polson & Scott, 2010)

- Essentially all shrinkage priors can be represented as

$$\theta_j \overset{ind}{\sim} N(0, \psi_j \tau), \quad \psi_j \sim g, \quad \tau \sim f$$

- τ - global shrinkage toward zero, ψ_j's - avoid over-shrinking signals locally
- g exponential (Bayesian Lasso, Park & Casella, 2008; Hans, 2009)
- g gamma (normal-gamma, Griffin & Brown, 2010)
- g inverse-gamma (RVM, Tipping, 2001)
- g square root of half-Cauchy (Horseshoe, Carvalho et al., 2009)

Global-local priors



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

Horseshoe (Carvalho, Polson and Scott, 2010, Biometrika)

- $\theta_j \mid \lambda_j, \tau \sim N(0, \lambda_j^2 \tau^2), \lambda_j \sim \text{Ca}^+(0,1), \tau \sim \text{Ca}^+(0,1)$
- The horseshoe prior has two interesting features that make it particularly useful as a shrinkage prior for sparse problems.
- Its flat, Cauchy-like tails allow strong signals to remain large (that is, un-shrunk) a posteriori.
- Yet its infinitely tall spike at the origin provides severe shrinkage for the zero elements of θ.

Horseshoe for fixed τ

- Let $y_i = \theta_i + \epsilon_i, i = 1, \ldots, n, \epsilon_i \sim N(0,1)$.
- Assume for now that $\tau = 1$, and define $\kappa_i = 1/(1+\lambda_i^2)$.
- κ_i is a random shrinkage coefficient, and can be interpreted as the amount of weight that the posterior mean for θ_i places on 0 once the data y have been observed.

$$E(\theta_i \mid y_i, \lambda_i) = \frac{\lambda_i^2}{1+\lambda_i^2} y_i + \frac{1}{1+\lambda_i^2} 0 = (1-\kappa_i) y_i$$

- Since $\kappa_i \in [0,1]$, this is clearly finite, and so by Fubini's Theorem

$$\begin{aligned} E(\theta_i|y) &= \int_0^1 (1-\kappa_i) y_i \pi(\kappa_i \mid y_i) d\kappa_i \\ &= (1 - E(\kappa_i \mid y_i)) y_i. \end{aligned}$$

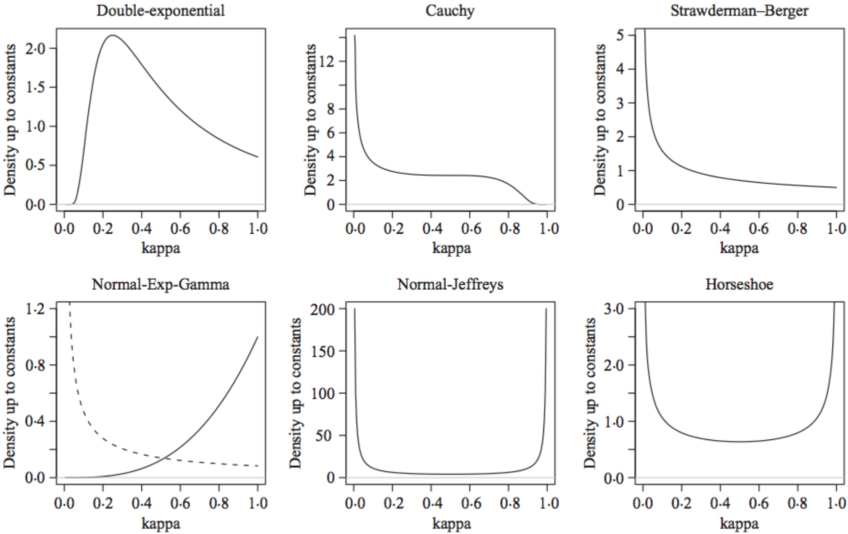
- If $\lambda_i \sim \mathrm{Ca}^+(0,1)$, $\kappa_i \sim \mathrm{Beta}(1/2, 1/2)$.

κ_i for various priors

Table 1. *Priors for λ_i and κ_i associated with some common local shrinkage rules. For the normal-exponential-gamma prior, it is assumed that $d = 1$. Densities are given up to constants.*

Prior for θ_i	Density for λ_i	Density for κ_i
Double-exponential	$\lambda_i \exp(-\lambda_i^2/2)$	$\kappa_i^{-2} \exp\{-1/(2\kappa_i)\}$
Cauchy	$\lambda_i^{-2} \exp\{1/(2\lambda_i^2)\}$	$\kappa_i^{-1/2}(1-\kappa_i)^{-3/2} \exp[-\kappa_i/\{2/(1-\kappa_i)\}]$
Strawderman–Berger	$\lambda_i (1+\lambda_i^2)^{-3/2}$	$\kappa_i^{-1/2}$
Normal-exponential-gamma	$\lambda_i (1+\lambda_i^2)^{-(c+1)}$	κ_i^{c-1}
Normal-Jeffreys	λ_i^{-1}	$\kappa_i^{-1}(1-\kappa_i)^{-1}$
Horseshoe	$(1+\lambda_i^2)^{-1}$	$\kappa_i^{-1/2}(1-\kappa_i)^{-1/2}$

Distribution of κ_i for various priors



Strengths of the Horseshoe prior

- It is highly adaptive both to unknown sparsity and to unknown signal-to-noise ratio.
- It is robust to large, outlying signals.
- It exhibits a strong form of multiplicity control by limiting the number of spurious signals.
- The horseshoe shares one of the most appealing features of Bayesian and empirical-Bayes model-selection techniques: after a simple thresholding rule is applied, the horseshoe exhibits an automatic penalty for multiple hypothesis testing.