Efficiency of Bayesian procedures in some high dimensional problems

Natesh S. Pillai
Dept. of Statistics, Harvard University
pillai@fas.harvard.edu

May 16, 2013
DIMACS Workshop

Joint Work: Collaborators

- Anirban Bhattacharya, Debdeep Pati and David Dunson (Duke University and Florida State)
- Christian Robert, Jean-Michel Marin, Judith Rousseau (Paris 9)
- Jun Yin (University of Wisconsin)

Outline

- Goal: Understand Bayesian methods in high dimensions.
- Example 1: Covariance matrix estimation
- Example 2: Bayesian model choice via ABC
- Implications, Frequentist-Bayes connection in high dimensions.

Conversation with Peter E. Huybers

- Motivation: Time variability in covariance patterns: 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.

Covariance Matrix Estimation: Why Shrinkage?

- We observe
$$y_1, \ldots y_n \overset{\text{i.i.d}}{\sim} N_{p_n}(0, \Sigma_{0n})$$
and set $\mathbf{y}^{(n)} = (y_1, \ldots, y_n)$
- For $p_n = p$, fixed, the sample covariance estimator
$$\Sigma^{\text{sample}} = \frac{1}{n}\sum_{i=1}^{n} y_i y_i^T$$
is consistent for population eigenvalues.
- $\hat{\lambda}_i$ are consistent for population eigenvalues:
$$\sqrt{n}(\hat{\lambda}_i - \lambda_i) \Rightarrow \mathrm{N}(0, V(\lambda_i))$$

Covariance Matrix in high dimensions

- Simplest Case: $\Sigma_{0n} = I$
- Take $p = p_n = c\, n$, $c \in (0, 1)$.
- $\widehat{\lambda}_1, \widehat{\lambda}_{p_n}$ largest and smallest (non-zero) eigenvalues of

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

- Then as $n \to \infty$ (and thus p_n also grows), (Marcenko-Pastur, 1967) almost surely!

$$\lim_{n\to\infty} \widehat{\lambda}_1 = (1 + \sqrt{c})^2$$

$$\lim_{n\to\infty} \widehat{\lambda}_{p_n} = (1 - \sqrt{c})^2$$

- MLE is not consistent!

Covariance Matrix in high dimensions

- $\lim_{n\to\infty} \widehat{\lambda}_1 = (1+\sqrt{c})^2 = \lambda_+$.
- Confidence Interval:

$$n^{2/3}(\widehat{\lambda}_1 - \lambda_+) \Rightarrow \mathrm{TW}_1$$

  where TW_1 is the Tracy-Widom law (Johnstone 2000).
- Universality phenomenon: Results go beyond the case of Gaussian (Tao and Vu, 2009; P. and Yin, 2011)

Correlation Matrix

- Johnstone (2001): Correlation Matrices for PCA.

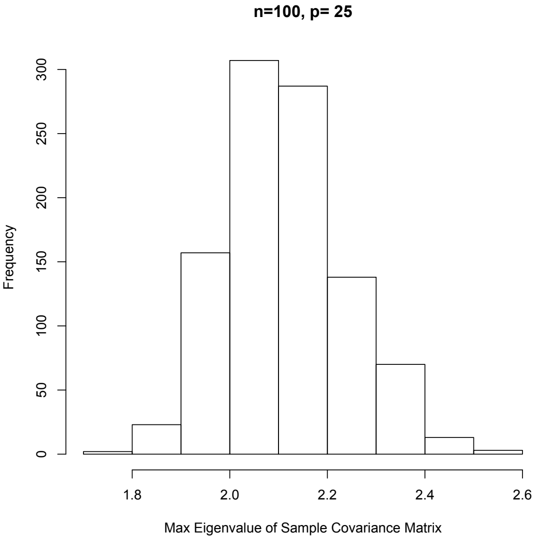
Theorem (P. and Yin, 2012, AoS)

Largest eigenvalue of sample correlation matrices still inconsistent. All of the problems from covariance matrices persist.

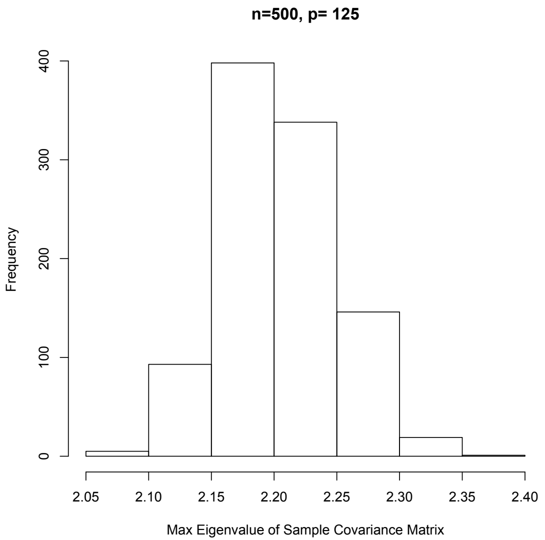
Understanding Asymptotics

- 20 century $n \to \infty$.
- Now: both $p, n \to \infty$.
- Why should we bother?
- Because the above asymptotics is remarkably accurate for 'small' n, 'small' p!

Sample covariance matrix plot, n = 100, p = 25



Sample covariance matrix plot, n = 500, p = 125



Factor Models: Motivation

- Interest in estimating dependence in high-dim obs. + prediction and classification from high-dim correlated markers such as gene expression, SNPs.
- Center prior on a "sparse" structure, while allowing uncertainty and flexibility.
- Latent factor methods (West, 2003; Lucas et al., 2006; Carvalho et al., 2008).
- Huge applications (economics, finance, signal processing..)

Gaussian factor models

- Explain dependence through shared dependence on fewer *latent factors*

$$y_i \sim \mathrm{N}(0, \Sigma_{p \times p}) , \quad 1 \leq i \leq n .$$

- Focus on the case $p = p_n \gg n$.
- Factor models assume the "decomposition"

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

- Λ is a $p \times k$ matrix, $k \ll n$.

Gaussian factor models

- Explain dependence through shared dependence on fewer *latent factors*

$$y_i = \mu + \Lambda\eta_i + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}_p(0, \Sigma), \quad i = 1, \ldots, n$$

- $\mu \in \mathbb{R}^p$, a vector of means, with $\mu = 0$.
- $\eta_i \in \mathbb{R}^k$, latent factors, Λ a $p \times k$ matrix of factor loadings with $k \ll p$.
- ϵ_i are i.i.d with $\mathrm{N}(0, \sigma^2)$.

Factor models for covariance estimation

- Unstructured Σ has $O(p^2)$ free elements
- Factor models $\Sigma = \Lambda\Lambda^T + \sigma^2 \mathrm{I}_p$.
- Still $O(p)$ elements to estimate!

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 Liu, 2011 ...)
- Estimator which achieves the 'minimax' rate:

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

- Unstable; Confidence intervals..

Sparse factor modeling

- A natural bayesian alternative: *sparse factor modeling* (West, 2003); also (Lucas et al., 2006; Carvalho et al., 2008) and many others
- Allow zeros in loadings through point mass mixture priors: Λ_{ij} given point mass priors or shrinkage priors.
- Prior assigns $\Lambda_{ij} = 0$ with non-zero probability.
- Why care about this prior? Bayesian analogue of thresholding.
- Assume k to be known (but easy to relax this).

Important questions

- Can Bayes methods produce estimators which are comparable to frequentist estimators?
- Can one do computation in reasonable time?
- How to address Statistical efficiency-Computational efficiency trade off?

Our objective

- Bayesian counterpart lacks a theoretical framework in terms of posterior convergence rates.
- A prior $\Pi(\Lambda \otimes \sigma^2)$ induces a prior distribution $\Pi(\Omega)$
- How does the posterior behave assuming data sampled from fixed truth?
- Huge literature on frequentist properties of the posterior distribution

Questions need to be addressed

- Does the posterior measure concentrate around the truth increasingly with sample size?
- What role does the prior play?
- How does the dimensionality affect the rate of contraction?

Preliminaries

- We consider the operator norm ($\| \cdot \|_2$)

$$\|A\|_2 = \sup_{x \in \mathcal{S}^{r-1}} \|Ax\|_2 = s_{(1)}$$

- Largest Eigenvalue of A, for symmetric A.

Setup

- We observe
$$y_1, \ldots y_n \overset{\text{i.i.d}}{\sim} N_{p_n}(0, \Sigma_{0n})$$
and set $\mathbf{y}^{(n)} = (y_1, \ldots, y_n)$, $\Sigma_{0n} = \Lambda_0 \Lambda_0^t + \sigma^2 I_{p_n \times p_n}$
- Want to find a minimum sequence $\epsilon_n \to 0$ such that
$$\lim_{n \to \infty} \mathbb{P}\left[\|\Sigma - \Sigma_{0n}\|_2 > \epsilon_n \mid \mathbf{y}^{(n)} \right] = 0$$
- Can we find such ϵ_n even if $p_n \gg n$?
- What is the role of the prior?

Assumptions on truth

"Realistic Assumption:"

(A1) Sparsity: Each column of Λ_{0n} has at most s_n non-zero entries, with $s_n = O(\log p_n)$.

Prior choice & a key result

Prior

(PL) Let $\Lambda_{ij} \sim (1-\pi)\delta_0 + \pi g(\cdot)$, $\pi \sim \text{Beta}(1, p_n + 1)$. $g(\cdot)$ has Laplace like or heavier tails

Theorem (Pati, Bhattacharya, P. and Dunson, 2012)

For the high-dimensional factor model $r_n = \sqrt{\log^7(p_n)/n}$,

$$\lim_{n\to\infty} \mathbb{P}(\|\Sigma - \Sigma_0\|_2 > r_n \mid \mathbf{y}^{(n)}) = 0 \, .$$

Prior choice & a key result

Prior

(PL) Let $\Lambda_{ij} \sim (1-\pi)\delta_0 + \pi g(\cdot)$, $\pi \sim \text{Beta}(1, p_n + 1)$. $g(\cdot)$ has Laplace like or heavier tails

Theorem (Pati, Bhattacharya, P. and Dunson, 2012)

For the high-dimensional factor model $r_n = \sqrt{\log^7(p_n)/n}$,

$$\lim_{n\to\infty} \mathbb{P}(\|\Sigma - \Sigma_0\|_2 > r_n \mid \mathbf{y}^{(n)}) = 0\ .$$

Implication of the result

- Rate $\epsilon_n = \sqrt{\log^2(p_n)/n}$.
- We will get consistency if

$$\lim_{n\to\infty} \frac{\log^7 p_n}{n} = 0 .$$

- Ultra-High dimensions, $p_n = e^{n^{1/7}}$.

Important Implication for Asymptotics

- This rate we get is similar to the minimax rate for similar problems Cai and Zhou (2011), but not the same!
- $$r_n = \text{minimax rate} \times \sqrt{\log p_n}$$
- The above phenomenon is similar to what happens in mixture modeling!
- Ghosal (2001): Bayesian nonparametric modeling doesn't match frequentist rates.
- If true: Serious implications.

A couple of Implications

- Minimax theory will tell only half the story.
- Heuristics based on bayes.
- BIC?
- Frequentist-Bayes agreement/disagreement?

Interesting Challenges in Mathematical Statistics

- Need to have 2 things to show Bayesian methods work well.
- Show prior is not too "dogmatic".
- Likelihood is able to "separate points".
- Neymann-Pearson Lemma
- Separation of points: Traditional Likelihood Ratio doesn't work!

Example : Intuition and Tools from Random Matrix Theory

- Intuition from random matrix theory (RMT) - “tall” matrices properly normalized look like identity matrices.
- If entries of Λ_0 were drawn i.i.d. $N(0, 1)$, Vershynin (2011) tells us

$$\|\frac{1}{p}\Lambda_0^{\mathrm{T}}\Lambda_0 - \mathrm{I}_k\|_2 \leq C\frac{\sqrt{k}}{\sqrt{p}}$$

with high probability.

Computationally easier priors

- We need to construct prior distribution for a $p_n \times 1$ vector Λ.
- Conjugate priors – easier to update
- Many popular ones.
- Many 'loss functions' are prior distributions; thus point estimates are posterior modes.

Regularization: Statistical flavor of the decade

- Estimates of the form

$$\hat{\Lambda} = \arg\min_{\Lambda} \sum_{i=1}^{n} (Y_i - \Lambda_i)^2 + \theta \sum_{i=1}^{n} |\Lambda_i|^k .$$

- Gazillion papers; not a SINGLE one constructs confidence intervals or uncertainty estimation.
- Two special cases: $k = 2$: (Ridge regression, James-Stein type)

$$\hat{\Lambda} = \arg\min_{\Lambda} \sum_{i=1}^{n} (Y_i - \Lambda_i)^2 + \theta \sum_{i=1}^{n} |\Lambda_i|^2 .$$

- $k = 1$: (LASSO)

$$\hat{\Lambda} = \arg\min_{\Lambda} \sum_{i=1}^{n} (Y_i - \Lambda_i)^2 + \theta \sum_{i=1}^{n} |\Lambda_i| .$$

Prior choice & another key result

Prior

- Let the columns Λ_i = LASSO or RIDGE prior.

Theorem (Pati, Bhattacharya, P. and Dunson, 2012)

For a large class of models, the above, the convergence rate is strictly slower than the point mass priors.

Prior choice & another key result

Prior

- Let the columns $\Lambda_i =$ LASSO or RIDGE prior.

Theorem (Pati, Bhattacharya, P. and Dunson, 2012)

For a large class of models, the above, the convergence rate is strictly slower than the point mass priors.

Intuition?

- Independence!
- Stein phenomenon.

Dirichlet Laplace prior & properties

- We propose a simple dependent modification leading to optimal concentration & efficient computation

$$\Lambda_j \sim \mathrm{DE}(\phi_j \tau), \quad \phi = (\phi_1, \ldots, \phi_p)^{\mathrm{T}} \in \mathcal{S}^{p-1}, \quad \tau > 0$$

- DE = Double exponential
- Constraining ϕ to the simplex crucial - allows for dependence
- We let $\phi \sim \mathrm{Diri}(\alpha, \ldots, \alpha)$ - $\alpha < 1$ favors small # dominant values with remaining ≈ 0
- Computation easy! Take advantage of Conjugacy

Dirichlet-Laplace prior - motivation

Theorem (Pati, Bhattacharya, P. and Dunson, 2013)

The Dirichlet-Laplace priors produce convergence rates identical to that of the point mass priors.

ABC algorithm

- ABC: Approximate Bayes Computation.
- Rubin(1984)
- Generate $\theta^* \sim \pi$
- Generate pseudo-data Y_{pseudo} from f_{θ^*}.
- Accept θ^* as posterior, if

$$Y_{\text{pseudo}} = Y_{\text{obs}} \ .$$

- Repeat.

ABC algorithm

- Exactly matching the observed data - Impossible, even in 1 dimension!
- Key Idea: Approximately match.
- Choose a distance d, and tolerance ϵ.
- Accept θ^* if
$$d(Y_{\text{pseudo}}, Y_{\text{obs}}) < \epsilon \ .$$
- For a given d, accuracy of the procedure can be improved by choosing ϵ smaller and smaller and smaller...

ABC algorithm: Twist

- In real examples, it is still expensive/impossible to compute $d(Y_{\text{pseudo}}, Y_{\text{obs}})$.
- Twist: Use some function η of the data: called the “summary statistic" and accept if

$$d\Big(\eta(Y_{\text{pseudo}}), \eta(Y_{\text{obs}})\Big) < \epsilon \,.$$

- Why no sufficient statistics?
- Recall the Pitman-Koopman-Darmois theorem, for exponential families.
- Dimension of the sufficient statistic necessarily increases with the sample size!

ABC algorithm

- The above version, re-discovered in population genetics (Tavare et.al, 1997).
- Literally 100's of papers!
- How to choose *d* and ϵ?
- Fearnhead and Prangle, 2012, JRSS-B discussion.

ABC algorithm for Model Selection

- Compare 2 models: compute the Bayes factors.
- Bayes Factor $\propto$ Ratio of Marginal Likelihoods.
- Jeffreys' interpretation, as strength of evidence.
- Easy to perform, using the ABC algorithm!

ABC algorithm for Model Selection

- Choose Model 1 or 2 according to the prior.
- Given the model, generate $(\theta^*, Y_{\text{pseudo}})$ from the prior distribution of the corresponding model.
- Accept θ^*, and the Model, if

$$d(Y_{\text{pseudo}}, Y_{\text{obs}}) < \epsilon \ .$$

- Estimate for Bayes Factor = $\frac{\#\ \textit{of timesModel}\ 1\ \textit{is accepted}}{\#\ \textit{of timesModel}\ 2\ \textit{is accepted}}$

ABC algorithm for Model Selection using η

- The above algorithm = Recipe for Disaster!
- High Profile papers!
- Miller, N. et al, (2005) Science.
- Multiple transatlantic introductions of the Western corn rootworm.

Lots of popular software

- Donoho (2002).
- DIY-ABC
- ABCToolbox
- PopABC
- ABC-SysBio

Result

Theorem (Robert, Jean-Marie, Jean-Michel, P., 2011, PNAS)

Bayes Model selection based on a summary statistic η can be INCONSISTENT.

ABC algorithm for Model Selection using η

- “Popular beliefs" in the field.
- Accuracy can be increased with choosing ϵ very small: thus increase in computing power leads to more accurate results.
- If gives reasonable answers for parameter estimation, no reason why it should go wrong for model selection!

ABC algorithm for Model Selection

- What goes wrong for model selection?
- Marginal likelihood based on $\eta(Y) := \int_\Theta f(\eta(Y)|\theta)\pi(\theta)d\theta$.
- $\mathrm{BF}(\eta(Y))$:= Bayes Factor based on the single observation $\eta(Y)$.
- Sufficiency vs. Ancilliarity!

Example

- A statistic can be sufficient for two models, but cannot be "sufficient" across the models.
- Ancilliarity......?
- Suppose, we observe $Y = (y_1, y_2, \cdots, y_n)$ integer valued data.
- Two competing models: Poisson(λ) vs. Geometric(p).
- Statistic $\eta(Y) = \sum_{i=1}^{n} y_i$.

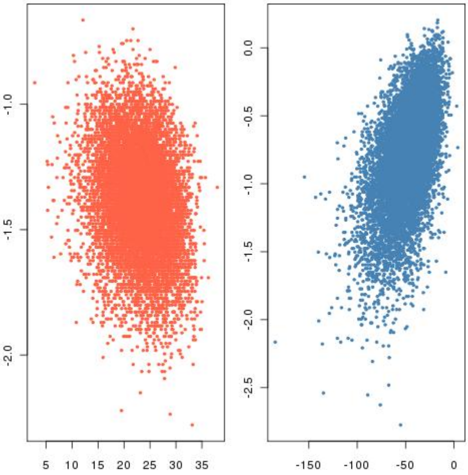
Example

- Almost surely, as the sample size goes to infinity, the Bayes Factor based on η converges to

$$\theta_0^{-1}(\theta_0 + 1)^2 e^{-\theta_0} \ ,$$

where $\theta_0 = \mathbb{E}(y_i) > 0$.

Ilustration



$\sum y_i$ vs. BF plot.

Another Example

- Consider two models:
- Model 1: $\mathrm{N}(\theta_1, 1)$, Model 2: $\mathrm{Laplace}(\theta_2, \frac{1}{\sqrt{2}})$
- $\bar{Y}$
- Median(Y)
- Sample variance
- mad(Y) = Median(|Y - Median(Y)|)

Conclusions

- Shrinkage priors = serious business in high dimensions.
- Innocent looking priors may look “dogmatic".
- Frequentist-Bayes agreement may not hold, implications?
- Ad-hoc methods often don’t work, but opportunity for statistical theory.
- Lots of open problems, virtually nothing is known!

References

- Universality of Correlation matrices (P., Yin, J., 2012), Annals of Statistics.
- Lack for confidence in ABC model selection, (Robert, Jean-Marie, Jean-Michel, P., 2011),PNAS.
- Bayesian Shrinkage, (Pati, Bhattacharya, P., Dunson, 2012) (2012)
- Bayesian high dimensional covariance estimation using factor models (Pati, Bhattacharya, P., Dunson, 2012)
- Universality of Covariance matrices (P., Yin, J., 2013), Annals of Applied Probability

Remarks

Thank you!