Bayesian Sparse PCA

Harrison H. Zhou
Department of Statistics
Yale University

Jointly with Chao Gao
Outline

• Introduction
  – Sparse PCA
  – Bayesian Estimation

• Main Results
  – Adaptive Prior
  – Rate-optimal Posterior Contraction

• Rank-one Computation

• Summary
Introduction

What is PCA:

Given a r.v. $X \in \mathbb{R}^p$ with $\mathbb{E}X = 0$ and $\text{Cov}(X) = \Sigma$.

The first PC is defined as

$$\theta = \arg \max_{a \in S^{p-1}} \text{Var}(a^T X).$$

Equivalently, $\theta$ is the first eigen-vector of $\Sigma$. 
Consistency

Sample Version:

Given i.i.d. $X_1, X_2, ..., X_n$, compute sample covariance $\hat{\Sigma}$.

Estimate the first PC $\theta$ by the first eigen-vector of $\hat{\Sigma}$.

Consistency when $p/n \to 0$:

By Davis-Kahan matrix perturbation bound, consistency is obtained, since

$$\|\hat{\Sigma} - \Sigma\|^2 = O_P\left(\frac{p}{n}\right) = o_P(1).$$
Inconsistency

Inconsistency when $p/n \to c$:

For $\Sigma = I + \theta^T \theta$, fix the norm of first PC $||\theta||$ and let $p, n \to \infty$, then

$$\left| \cos \angle (\hat{\theta}, \theta) \right|^2 \to \frac{\max (||\theta||^4 - c, 0)}{||\theta||^4 + c||\theta||^2}, \quad \text{a.s.}$$

This motivates sparse PCA. See Johnstone and Lu (2009).
Sparse PCA – Spike

Spiked Covariance:

\[ \Sigma = \sum_{l=1}^{r} \theta_l \theta_l^T + I = V_0 \Lambda_0 V_0^T + I, \]

and \( X_i \sim N(0, \Sigma) \).

Equivalent Formulation:

\( X_i \) can be written as

\[ X_i = V_0 \Lambda_0^{1/2} W_i + Z_i, \]

where \( W_i \sim N(0, I_{r \times r}) \) and \( Z_i \sim N(0, I_{p \times p}) \) independently.

See Johnstone and Lu (2009) and Birnbaum et al. (2013).
Sparse PCA – Parameter Space

Parameter Space $\mathcal{G}(s, r)$:

1. 
   \[
   \max_{1 \leq l \leq r} |\text{supp}(\theta_l)| \leq s,
   \]

2. 
   \[
   ||\theta_l||^2 \in (K^{-1}, K),
   \]

where

\[
\Sigma = \sum_{l=1}^{r} \theta_l \theta_l^T + I, \text{ with } \theta_l^T \theta_k = 0 \text{ for } k \neq l.
\]
Sparse PCA – Loss Function

Loss Function:

Consider the orthonormal matrix

\[ V_0 = [||\theta_1||^{-1}\theta_1, \ldots, ||\theta_r||^{-1}\theta_r]. \]

The projection matrix to \(\text{span}(\theta_1, \ldots, \theta_r)\) is \(V_0 V_0^T\).

Consider

\[ \|\hat{V}\hat{V}^T - V_0 V_0^T\|_F \approx \inf_{Q \in \mathcal{U}(r, r)} \|\hat{V} - V_0 Q\|_F. \]
Sparse PCA – Minimax Rate

Minimax Rate:

Let \( r \leq m \log p \) and \( s \leq p^{1-c} \), for some \( c \in (0,1) \) and \( m > 1 \).

Vu and Lei (2013) proved

\[
\inf_{\hat{\Sigma}} \sup_{\Sigma \in \mathcal{G}(s,r)} P^n \left\| \hat{\Sigma} \hat{V}^T - V_0 V_0^T \right\|_F^2 \preceq \frac{rs \log p}{n}.
\]

Remark: Minimax rate for sparse PCA in other settings are proved in Birnbaum et al. (2013), Cai, Ma and Wu (2013a) and Cai, Ma and Wu (2013b).
Question:

Can we have an adaptive Bayesian solution to sparse PCA?
Bayesian Framework

Model:

\[ X^n \sim P^n_{\Gamma}, \]

where \( P_{\Gamma} = N(0, \Gamma) \).

Prior:

\[ \Gamma = V\Lambda V^T + I \sim \Pi. \]

Posterior:

\[ \Pi(A|X^n) = \frac{\int_A \frac{dP^n_{\Gamma}}{dP^n_{\Sigma}}(X^n)d\Pi(\Gamma)}{\int \frac{dP^n_{\Gamma}}{dP^n_{\Sigma}}(X^n)d\Pi(\Gamma)}, \]

where \( P_{\Sigma} = N(0, \Sigma) \).
Posterior Contraction

Construct a prior $\Pi$ such that for any $\Sigma \in \mathcal{G}(s,r)$,

$$P^n_{\Sigma} \Pi \left( \| V V^T - V_0 V_0^T \|_F > M \epsilon |X^n| \right) \to 0,$$

where $\epsilon^2 = \frac{rs \log p}{n}$, for some $M > 0$. 
Relations with Other Works

Pati et al. (2013):

Consider $r \leq m \log p$ and $s \leq p^{1-c}$, for some $c \in (0, 1)$ and $m > 1$ as in Vu and Lei (2013).

Then

$$
P^n \sum \Pi \left( \| V V^T - V_0 V_0^T \|_F > M \epsilon |X^n| \right) \to 0,
$$

where $\epsilon^2 = \frac{r^3 s \log p \log n}{n}$, for some $M > 0$. 
Relations with Other Works

Inspired by Castillo and van der Vaart (2012):

- First work in theory of Bayes sparse estimation.
- Rate-optimal posterior contraction for sparse Gaussian sequence model.
- Efficient computation.

A Remark:
Banerjee and Ghosal (2013) considered Bayesian bandable precision matrix estimation using conjugate prior, which is different from PCA setting.
Main Results
Consider $r \leq m \log p$ and $s \leq p^{1-c}$, for some $c \in (0, 1)$ and $m > 1$ as in Vu and Lei (2013).

Let $\epsilon^2 = \frac{rs \log p}{n}$. We can construct an adaptive prior such that

**Rate-optimal Posterior Contraction:**

$$P^n_{\Sigma} \mathbb{P}\left(\|VV^T - V_0V_0^T\|_F > M\epsilon|X^n\right) \leq \exp\left(-Cn\epsilon^2\right).$$

**Rank Selection:**

$$P^n_{\Sigma} \mathbb{P}\left(\xi \neq r|X^n\right) \leq \exp\left(-Cn\epsilon^2\right).$$
**Point Estimation**

Adaptive Rate-Optimal Point Estimation:

\[ P^n_{\Sigma} \left\| \mathbb{E}_\Pi \left( VV^T | X^n \right) - V_0V_0^T \right\|_F^2 \leq 2M^2\epsilon^2. \]
Proofs
Le Cam-Schwartz Scheme – Conditions

Construct a prior $\Pi$ such that

**Sufficient Mass:**

$$\Pi \left( \frac{||\Gamma - \Sigma||_F^2}{\lambda_{\min}(\Gamma)^2} \leq \epsilon^2 \right) \geq \exp \left( - C_1 n \epsilon^2 \right).$$

**Testing Function $\phi$:**

$$P^n_{\Sigma} \phi \lor \sup_{\Gamma \in \mathcal{F} \cap \{||VV^T - V_0 V_0^T||_F > M \epsilon\}} P^n_{\Gamma} (1 - \phi) \leq \exp \left( - C_2 n \epsilon^2 \right).$$

**Prior Concentration:**

$$\Pi(\mathcal{F}^c) \leq \exp \left( - C_3 n \epsilon^2 \right).$$
Le Cam-Schwartz Scheme – Conclusion

Under three conditions, we have

\[ P^n_{\Sigma} \Pi \left( \| VV^T - V_0V_0^T \|_F > M\epsilon |X^n| \right) \leq \exp \left( -Cn\epsilon^2 \right). \]

Remark:

The version presented here is adapted from Barron, Schervish and Wasserman (1999) and Ghosal, Ghosh and van der Vaart (2000).
An Adaptive Prior

For some $\gamma > 0$, we describe $\Gamma \sim \Pi$ in the following sampling procedure.

1. A rank $\xi$ is chosen uniformly from $\{1,\ldots,\lfloor p^{\gamma/2} \rfloor\}$;

2. Given $\xi$, for each $l \in \{1,\ldots,\xi\}$, we randomly choose $S_l \subset \{1,\ldots,p\}$ by letting the indicator $\mathbb{I}\{i \in S_l\}$ for each $i = 1,\ldots,p$ follow a Bernoulli distribution with parameter $p^{-(1+\gamma)}$;

3. Given $(S_1,\ldots,S_\xi,\xi)$, we sample a $p \times \xi$ matrix $A$ from $G(s_1,\ldots,s_\xi,\xi)$ to be specified below, and then let $\Gamma = AA^T + I$. 
An Adaptive Prior

Sample $A|(S_1, \ldots, S_\xi, \xi)$:
Remarks:

1. **Adaptation:** The prior does not depend on $r$ or $s$.

2. **Sparseness:** The prior samples a random sparse spiked covariance
   \[ \Gamma = \sum_{l=1}^{\xi} \eta_l \eta_l^T. \]

3. **Orthogonality:** The imposed orthogonality of $\{\eta_l\}_{l=1}^{\xi}$ is for controlling the
   spectrum of $\Gamma$. We impose boundedness on $||\eta_l||$ for each $l$, thus bounding
   the spectrum.
Properties of $\Pi$

Sufficient Mass:

If $r \lor \log n \leq m \log p$ for some $m > 0$,

$$\Pi\left( \frac{||\Gamma - \Sigma||^2}{\lambda_{\min}(\Gamma)^2} \leq \epsilon^2 \right) \geq \exp\left(-C_1n\epsilon^2\right).$$

Prior Concentration:

Let $\mathcal{F} = \{|S_1 \cup \ldots \cup S_\xi| \leq Ars\}$. If $r \lor \log n \leq m \log p$ for some $m > 0$,

$$\Pi\left(\mathcal{F}^c\right) \leq \exp\left(-C_3n\epsilon^2\right),$$

for some large $A > 0$. 
Properties of $\Pi$

Testing Function $\phi$:

There exists $\phi$ such that

$$P^n_\Sigma \phi \lor \sup_{\{\Gamma \in \text{supp}(\Pi) \cap \mathcal{F} : \|VV^T - V_0V_0^T\|_F > M\epsilon\}} P^n_\Gamma (1 - \phi) \leq \exp \left( - C_2 n \epsilon^2 \right).$$

Construction of $\phi$:

- We first test $\Sigma$ against the complement of its spectral neighborhood.
- Inside the neighborhood, we divide the set into pieces in a delicate way and a likelihood ratio test can be applied on each piece.
Computation on Rank-one Case
The Rank-one Setting

Single Spike Model:

\[ X_{ij} = W_i \theta_j + Z_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, p, \]

with \( Z_{ij} \) and \( W_i \) i.i.d. \( N(0, 1) \) for all \( i \) and \( j \).
A Prior

1. Given $\kappa > 0$, sample $q$ according to $\pi(q) \propto \exp \left( -\kappa q \log p \right)$ from $\{1, 2, ..., p\}$.

2. Given $q$, sample $S \subset \{1, 2, ..., p\}$ uniformly from all subsets with $|S| = q$.

3. Given $(q, S)$, sample $\eta_S \sim N(0, I_{q \times q})$. Let $\eta^T = (\eta^T_S, \eta^T_{S^c}) = (\eta^T_S, 0^T)$. The covariance is $\Gamma = \eta \eta^T + I$. 
Some Remarks:

1. The new prior does not restrict the $l^2$ norm of the spike $\eta$, because we do not consider rank-adaptation.

2. The posterior contraction is proved for this prior under the single-spike model.
Computation of Posterior Mean

Under the single spike model, the posterior mean ($j$-th coordinate) has formula

$$\mathbb{E}_\Pi(\eta_j | X^n) = \frac{\int \eta_j \int \prod_{i=1}^n \prod_{j=1}^p \phi(X_{ij} - W_i \eta_j) \phi(W^n) dW^n d\Pi(\eta)}{\int \int \prod_{i=1}^n \prod_{j=1}^p \phi(X_{ij} - W_i \eta_j) \phi(W^n) dW^n d\Pi(\eta)}$$

$$= \frac{\int N_{n,j}(W^n) \phi(W^n) dW^n}{\int D_n(W^n) \phi(W^n) dW^n},$$

where $\phi(W^n) dW^n = \prod_{i=1}^n \phi(W_i) dW_1...dW_n$ and $\phi$ is the density function of $N(0, 1)$. 

Computation of Posterior Mean

The denominator is $\int D_n(W^n)\phi(W^n)dW^n$.

Remembering the structure of the prior, $D_n(W^n)$ can be written as

$$D_n(W^n) = \sum_{q=1}^{p} \frac{\pi(q)}{(p)_q} \sum_{|S|=q \notin S} \left\{ \prod_{i=1}^{n} \phi(X_{ij}) \right\} \prod_{j \in S} \left\{ \int \prod_{i=1}^{n} \phi(X_{ij} - W_i\eta_j)\phi(\eta_j)d\eta_j \right\}$$

$$= \sum_{q=1}^{p} \frac{\pi(q)}{(p)_q} C(q, W^n).$$
Key: Castillo and van der Vaart (2012) observed that $C(q, W^n)$ is the coefficient of $Z^q$ of the polynomial

$$Z \mapsto \prod_{j=1}^{p} (f(X_j) + h(X_j, W^n)Z),$$

where

$$f(X_j) = \prod_{i=1}^{n} \phi(X_{ij}), \quad h(X_j, W^n) = \int \prod_{i=1}^{n} \phi(X_{ij} - W_i\eta_j)\phi(\eta_j)\,d\eta_j.$$
Computation of Posterior Mean

Step 1

Sample $W_1^n, W_2^n, ..., W_T^n$ i.i.d. from $N(0, I_{n \times n})$.

Step 2

For each $W_t^n$, compute coefficients of a polynomial with degree $p$, denoted as $\{C(q, W_t^n)\}_q$.

Step 3

Approximate $\int D_n(W^n) \phi(W^n) dW^n$ by

$$
\frac{1}{T} \sum_{t=1}^{T} D_n(W_t^n) = \frac{1}{T} \sum_{t=1}^{T} \left( \sum_{q=1}^{p} \frac{\pi(q)}{\binom{p}{q}} C(q, W_t^n) \right).
$$

Similar strategy can be used to compute the numerator.
Summary

- An adaptive prior is constructed.
- Posterior contraction is rate-optimal.
- A computation strategy is proposed for the rank-one case.