Information-theoretically Optimal Sparse PCA

Authors: Yash Deshpande, Andrea Montanari

Publication: IEEE International Symposium on Information Theory, 2014

Speaker: Jang Jehyuk

Short summary:

Sparse Principal Conponent Analysis (PCA) is a dimensionality reduction technique wherein one seeks a low rank representation of a data matrix with additional sparity.

The authors consider two probabilistic models of sparse PCA: a spiked Wigner and spiked Wishart (or spiked covariance) model. They analyze an Approximate Message passing (AMP) algorithm to seek the underlying data and show that AMP estimates are information-theoretically optimal for the models.

I. INTRODUCTION

A. Two sparse PCA models:

Suppose we are given data $Y_\lambda \in \mathbb{R}^{n\times n}$ distributed according to the following models

1) spiked Wigner model

$$Y_\lambda = \sqrt{\frac{\lambda}{n}} x x^{\mathbf{T}} + Z \tag{1}$$

Here $x \in \mathbb{R}^n$, and each coordinate is denoted by $x_i \sim \mathrm{Ber}(\varepsilon)$. $Z \in \mathbb{R}^{n\times n}$ is a symmetric matrix where $(Z_{ij})_{i\le j}$ are i.i.d $\mathrm{N}(0,1)$ variables, independent of x.

2) spiked Wishart model

$$Y_\lambda = \sqrt{\frac{\lambda}{n}} u v^{\mathbf{T}} + Z \tag{2}$$

Here $u \in \mathbb{R}^m$, with i.i.d coordinates $u_i \sim \mathrm{N}(0,1)$ and $v \in \mathbb{R}^n$ with i.i.d coordinates $v_j \sim \mathrm{Ber}(\varepsilon)$. Further, $Z \in \mathbb{R}^{m\times n}$ is a matrix with $Z_{ij} \sim \mathrm{N}(0,1)$ i.i.d. random variables.

In either case, the given data consists of a sparse, rank-one matrix observed through Gaussian noise. The authors let X denote the clean, underlying signal ($xx^{\mathbf{T}}$ or $uv^{\mathbf{T}}$). The goal is to estimate the signal X from the data Y_λ in the high dimension asymptotic where $n \to \infty$, $m \to \infty$ with $m/n \to \alpha \in (0, \infty)$.

B. Information-theoretical approach and AMP:

This paper focuses on estimation in the sense of the mean squared error, defined for an estimator error $\widehat{X}(Y_\lambda)$ as:

$$\mathrm{mse}\left(\widehat{X}, \lambda\right) = \frac{1}{n^2} \mathrm{E}\left\{\left\|\widehat{X} - X\right\|_F^2\right\} \tag{3}$$

The mean squared error is minimized by the estimator $\widehat{X} = \mathrm{E}\{X \mid Y_\lambda\}$, i.e. the conditional expectation of the signal given the observations [1]. The estimator, which is often intractable to compute, can be obtained using a polynomial-time scheme by using AMP algorithm.

Consequently, the minimum mean squared error(MMSE) is given by:

$$\text{M-mmse}(\lambda, n) \equiv \frac{1}{n^2} \mathrm{E}\left\{\left\|X - \mathrm{E}\{X \mid Y_\lambda\}\right\|_F^2\right\} \tag{4}$$

The machinery of AMP reduces the high-dimensional *matrix* problem in model (1), (2) to the following simpler *scalar* denoising problem:

$$Y_\lambda = \sqrt{\lambda} X_0 + N \tag{5}$$

where $X_0 \sim \mathrm{Ber}(\varepsilon)$ and $N \sim \mathrm{N}(0,1)$ are independent. The scalar MMSE [2] in estimating X_0 from Y_λ is given by:

$$\text{S-mmse}(X_0, \lambda) = \mathrm{E}\left\{\left(X_0 - \mathrm{E}\{X_0 \mid Y_\lambda\}\right)^2\right\} \tag{6}$$

C. Main results:

The main results of this paper characterize the optimal mean squared error (4) by using scalar MMSE (6) in the large n asymptotic, when $\varepsilon > \varepsilon_c \approx 0.05$, and establish that AMP achieves this fundamental limit.

II. BACKGROUND

- PCA, which involves using the principal eigenvector is ineffective in estimating the underlying clean signal X when $\lambda < \lambda_c$. (Phase transient phenomenon)
- Since the sparsity assumptions are added to the model, this can be leveraged when λ is small.
- Estimating the support correctly for some conditions is impossible due to information-theoretic obstructions. [17] So they focus on another natural figure-of-merit: the mean squared error.

III. ALGORITHM

AMP is a low complexity iterative algorithm that produces iterates $x^t, \hat{x}^t \in \mathbb{R}^n$ for a data matrix A.

Symmetric Bayes-optimal AMP algorithm

Input: Data Y_λ

Define $A = Y_\lambda / \sqrt{n}$ and $\hat{x}^0, \hat{x}^{-1} = 0$. For $t \geq 0$ compute

$$x^{t+1} = A\hat{x}^t - b_t \hat{x}^{t-1}$$

$$\hat{x}^{t+1} = f_t(x^{t+1})$$

$$\widehat{X}^{t+1} = \hat{x}^{t+1}(\hat{x}^{t+1})^{\mathrm{T}}$$

Here $f_t : \mathbb{R} \to \mathbb{R}$ are scalar functions and $\{b_t\}_{t \geq 0}$ is a sequence of scalars. For a scalar function f_t, the authors define its extension to $\mathbb{R}^n$ by applying it component-wise, i.e. $f_t : \mathbb{R}^n \to \mathbb{R}^n, v \mapsto f_t(v) = \left(f_t(v_1), f_t(v_2) \cdots, f_t(v_n)\right)^{\mathbf{T}}$

The key property of approximate message passing is that it admits an *asymptotically exact* characterization in the high dimensional limit where $n \to \infty$. The iterates x_i^t converge as $n \to \infty$ to Gaussian random variables with prescribed mean and variance. These prescribed

mean and variance parameters evolve according to deterministic recursion, jointly termed "state evolution". It is defined for $t \geq 0$ as:

$$\mu_{t+1} = \sqrt{\lambda}\mathrm{E}\left\{X_0 f_t\left(\mu_t X_0 + \sqrt{\tau_t}Z\right)\right\} \tag{7}$$

$$\tau_{t+1} = \mathrm{E}\left\{f_t\left(\mu_t X_0 + \sqrt{\tau_t}Z\right)^2\right\} \tag{8}$$

Where $X_0 \sim \mathrm{Ber}(\varepsilon)$ and $Z \sim \mathrm{N}(0,1)$ are independent. The recursion is initialized with $\mu_0 = \tau_0 = 0$.

The scalar b_t are computed as:

$$b_t = \frac{1}{n}\sum_{i=1}^{n} f_t'\left(x_i^t\right)$$

The authors choose $f_t(y) = \mathrm{E}\left\{X_0 \mid \mu_t X_0 + \sqrt{\tau_t}Z = y\right\}$, the posterior expectation of X_0, with observation corrupted by Gaussian noise and SNR μ_t^2 / τ_t.

IV. Main Result

I introduce the result for spiked Wigner case because the paper doesn't provide proof for spiked Wishart case. The authors say the proof for spiked Wishart case follows similar ideas and will be provided in the full version of the present paper.

Definition II.1 *Let $\varepsilon_* \in (0,1)$ be the smallest positive real number such that for every $\varepsilon > \varepsilon_*$ the following is true. For every $\lambda > 0$, the equation below has only one solution in $[0,\infty)$:*

$$\lambda^{-1} y = \varepsilon - \text{S-mmse}(X_0, y) \tag{9}$$

Here $X_0 \sim \mathrm{Ber}(\varepsilon)$

Theorem 2. *Under Model (1) we have* $\text{M-mmse}(\lambda) = \lim_{n\to\infty} \text{M-mmse}(\lambda,\text{n})$ *exists for every $\lambda \geq 0$. This limit satisfies, when $\varepsilon \geq \varepsilon_*$:*

$$\text{M-mmse}(\lambda) = \varepsilon^2 - \frac{y_*(\lambda)^2}{\lambda^2} \tag{10}$$

where $y_*(\lambda)$ *is the unique solution to Eq.* (9) *above. Further, the symmetric Bayes-optimal AMP algorithm satisfies the following almost surely:*

$$\lim_{n\to\infty}\lim_{t\to\infty}\mathrm{MSE}_{\mathrm{AMP}}(\lambda,t)=\text{M-mmse}(\lambda) \tag{11}$$

To sum up, Theorem 2 says the estimation of AMP achieves minimum mean squared error under limiting regimes and the machinery of AMP reduces the high-dimensional matrix problem to the simpler scalar problem.

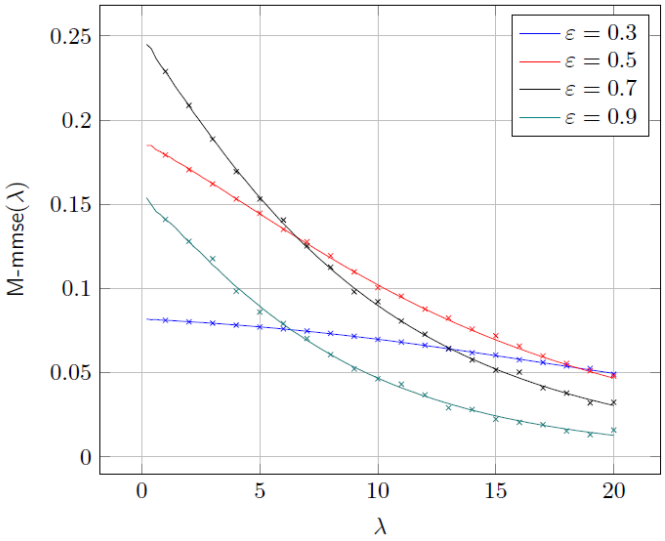

Figure 1. The solid curves $\text{M-mmse}(\lambda)$ above are computed analytically using Theorem 2. The crosses mark median MSE incurred by AMP in 100 Monte Carlo runs with $n=2000$ for the spiked Winger model (1).

V. Sketch of Proof

The paper provides proofs of two propositions. The first proposition is that mean squared error of MSE estimation, $\lim_{t\to\infty}\lim_{n\to\infty}\mathrm{MSE}_{\mathrm{AMP}}(\lambda,t)$ can be characterized as the terms of scalar MMSE, $\text{S-mmse}(X_0,\lambda)$. The second proposition says $\int_0^\infty \text{M-mmse}(\lambda)d\lambda \geq \int_0^\infty \lim_{t\to\infty}\lim_{n\to\infty}\mathrm{MSE}_{\mathrm{AMP}}(\lambda,t)\,d\lambda$.

Since the posterior expectation minimizes the mean squared error, we have that $\text{M-mmse}(\lambda)\leq \lim_{t\to\infty}\lim_{n\to\infty}\mathrm{MSE}_{\mathrm{AMP}}(\lambda,t)$. These imply that $\lim_{t\to\infty}\lim_{n\to\infty}\mathrm{MSE}_{\mathrm{AMP}}(\lambda,t)=\text{M-mmse}(\lambda)$

The sketch of proof for the first proposition is as follows.

Note that:

$$\begin{aligned}\mathrm{MSE}_{\mathrm{AMP}}(\lambda,t) &= \frac{1}{n^2}\left\|\widehat{X}^t - X\right\|_F^2 \\ &= \frac{1}{n^2}\left\|\hat{x}^t\left(\hat{x}^t\right)^{\mathbf{T}} - xx^{\mathbf{T}}\right\|_F^2 \\ &= \frac{1}{n^2}\left(\left\|\hat{x}^t\right\|^4 + \|x\|^4 - 2\left\langle \hat{x}^t, x\right\rangle^2\right)\end{aligned}$$

By the strong law of large numbers, $\left(\|x\|^2/n\right)^2 \to \left(\mathrm{E}\left\{x^2\right\}\right)^2 = \varepsilon^2$ almost surely.

The functions $f_t(y)$ are $\sqrt{\lambda}$-Lipschitz continuous. Hence, it is a direct consequence of Theorem 1 of [21] that the following limits hold almost surely:

$$\lim_{n\to\infty}\left(\frac{1}{n}\left\|\hat{x}^t\right\|^2\right)^2 = \mathrm{E}\left\{\left(\hat{x}^t\right)^2\right\}^2 = \mathrm{E}\left\{f_t\left(\mu_t X_0 + \sqrt{\tau_t}Z\right)^2\right\}^2$$

$$\lim_{n\to\infty}\left(\frac{1}{n}\left\langle \hat{x}^t, x\right\rangle\right)^2 = \mathrm{E}\left\{x\cdot\hat{x}^t\right\}^2 = \mathrm{E}\left\{X_0 f_t\left(\mu_t X_0 + \sqrt{\tau_t}Z\right)\right\}^2$$

The authors' choice $f_t(y) = \mathrm{E}\left\{X_0 \mid \mu_t X_0 + \sqrt{\tau_t}Z = y\right\}$ yields:

$$\mathrm{E}\left\{X_0 f_t\left(\mu_t X_0 + \sqrt{\tau_t}Z\right)\right\} = \mathrm{E}\left\{X_0 \mathrm{E}\left\{X_0 \mid \mu_t X_0 + \sqrt{\tau_t}Z\right\}\right\}$$

$$= \int_y \sum_{k=0}^{1} k\mathrm{E}\left\{X_0 \mid \mu_t X_0 + \sqrt{\tau_t}Z = y\right\} f_{X_0,\mu_t X_0+\sqrt{\tau_t}}\left(X_0 = k, \mu_t X_0 + \sqrt{\tau_t}Z = y\right)dy$$

$$= \int_y \mathrm{E}\left\{X_0 \mid \mu_t X_0 + \sqrt{\tau_t}Z = y\right\}\left\{\sum_{k=0}^{1} k\cdot f_{X_0,\mu_t X_0+\sqrt{\tau_t}}\left(X_0 = k \mid \mu_t X_0 + \sqrt{\tau_t}Z = y\right)\right\} f_{\mu_t X_0+\sqrt{\tau_t}}\left(\mu_t X_0 + \sqrt{\tau_t}Z = y\right)dy$$

$$= \int_y \mathrm{E}\left\{X_0 \mid \mu_t X_0 + \sqrt{\tau_t}Z\right\}^2 f_{\mu_t X_0+\sqrt{\tau_t}}\left(\mu_t X_0 + \sqrt{\tau_t}Z = y\right)dy$$

$$= \mathrm{E}\left\{\mathrm{E}\left\{X_0 \mid \mu_t X_0 + \sqrt{\tau_t}Z\right\}^2\right\} = \tau_{t+1} = \sqrt{\lambda}^{-1}\mu_{t+1}$$

Thus,

$$\lim_{n\to\infty}\mathrm{MSE}_{\mathrm{AMP}}(\lambda,t) \xrightarrow{\text{a.s.}} \varepsilon^2 - \tau_{t+1}^2$$

Let $\tau_*(\lambda)$ denote the smallest non-negative fixed solution of the equation(convergence point):

$$\tau_* = \mathrm{E}\left\{\mathrm{E}\left\{X_0 \mid \sqrt{\lambda}\tau_* X_0 + \sqrt{\tau_*}Z\right\}^2\right\} \tag{12}$$

Since the right had side of the equation (12) equals to $\varepsilon(1-\varepsilon)$ at $\tau_* = 0$ and ε at $\tau_* = \infty$, at least one fixed point must exist. Hence $\tau_*(\lambda)$ is well defined. It follows that

$$\lim_{t\to\infty}\lim_{n\to\infty}\mathrm{MSE}_{\mathrm{AMP}}\left(\lambda,t\right)\underset{\text{a.s.}}{\to}\varepsilon^2-\tau_*^2$$

Now, note that:

$$\begin{aligned}\tau_* &= \mathrm{E}\left\{\mathrm{E}\left\{X_0\,|\,\sqrt{\lambda}\tau_* X_0+\sqrt{\tau_*}Z\right\}^2\right\}\\ &= \mathrm{E}\left\{\mathrm{E}\left\{X_0\,|\,\sqrt{\lambda\tau_*}X_0+Z\right\}^2\right\}\\ &= \mathrm{E}\left\{X_0^2\right\}-\mathrm{E}\left\{\left(X_0-\mathrm{E}\left\{X_0\,|\,\sqrt{\lambda\tau_*}X_0+Z\right\}\right)^2\right\}\\ &= \varepsilon-\text{S-mmse}\left(X_0,\lambda\tau_*\right)\end{aligned}$$

VI. CONCLUSION

This paper introduces two sparse PCA models and AMP algorithm to obtain the optimal estimator in the sense of the mean squared error for underlying signals from the given data. AMP reduces the high-dimensional matrix problem to simpler scalar problem. They characterize mean squared error of AMP in terms of scalar mean squared error. And, the paper provides proof about the relationship between mean squared error of AMP and Matrix minimum mean squared error for one model, spiked Wigner model.

Reference

[1] T. M. Cover and J. A. Thomas, *Elements of information theory*, John Wiley & Sons, 2012

[2] D. Guo, S. Shamai, and S. Verdu, "Mutual information and minimum mean-square error in Gaussian channels," Information Theory, IEEE transactions on, vol. 51, no. 4, pp. 1261-1282, 2005

[17] A. A. Amini and M. J. Wainwright, "High-dimensional analysis of semidefinite relaxations for sparse principal components," The Annals of Statistics, vol. 37, no. 5B, pp. 2877-2921, 2009

[21] A. Javanmard and A. Montanari, "State evolution for general approximate message passing algorithms, with applications to spatial coupling," arXiv preprint: 1211.5164, 2012