Sparse PCA from incomplete data

Lekshmi Ramesh

Indian Institute of Science
Bangalore

April 14, 2018

Inference from data with missing values

- Missing data occur frequently in practice
- How to design good estimators/tests in the presence of missing data?
- Commonly used fix
  - Discard samples with missing values–can lead to loss of large amount of data
  - Imputation–usually ad hoc
- This presentation: Two approaches to do SPCA using incomplete data, guarantees

Approximating SPCA from incomplete data[1]

- Data matrix $\mathbf{X} \in \mathbb{R}^{d \times n}$–$n$ samples in d dimensions
- PCA: find solution to the the following variance maximization problem

$$\mathbf{V}_k = \underset{\mathbf{V} \in \mathbb{R}^{d \times k}, \mathbf{V}^\top \mathbf{V} = \mathbf{I}}{\arg\max} \operatorname{Tr} \mathbf{V}^\top \mathbf{X} \mathbf{X}^\top \mathbf{V}$$

- SPCA: additional sparsity constraint on columns of $\mathbf{V}$

$$\mathbf{S}_k = \underset{\mathbf{V} \in \mathbb{R}^{d \times k}, \mathbf{V}^\top \mathbf{V} = \mathbf{I}, \|\mathbf{V}_i\|_0 \leq r}{\arg\max} \operatorname{Tr} \mathbf{V}^\top \mathbf{X} \mathbf{X}^\top \mathbf{V}$$

[1] Abhisek Kundu, Petros Drineas, and Malik Magdon-Ismail. “Approximating Sparse PCA from Incomplete Data”. In: *Advances in Neural Information Processing Systems*. 2015, pp. 388–396.

Approximating SPCA from incomplete data

- Missing data case
  - Only a sparse sampling of entries of $\mathbf{X}$ available–use it to construct the "sketch" $\tilde{\mathbf{X}}$
  - Solve SPCA using the sketch–call the output $\tilde{\mathbf{S}}$
- How does $\tilde{\mathbf{S}}_k$ perform as an approximation to $\mathbf{S}_k$?
- Quality of approximation measured in terms of the objective

Main result–I

Theorem

Let $\mathbf{S}_k$ and $\tilde{\mathbf{S}}_k$ be solutions to the Sparse PCA problem with full data and with sketched data, respectively. Then,

$$\mathrm{Tr}(\tilde{\mathbf{S}}_k^\top \mathbf{X}\mathbf{X}^\top \tilde{\mathbf{S}}_k) \geq \mathrm{Tr}(\tilde{\mathbf{S}}_k^\top \mathbf{X}\mathbf{X}^\top \tilde{\mathbf{S}}_k) - 2k\|\mathbf{X}\mathbf{X}^\top - \tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top\|_{op}$$

- Doing SPCA using $\tilde{\mathbf{X}}$ is good if $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top$ closely approximates $\mathbf{X}\mathbf{X}^\top$
- Will see: $\|\mathbf{X}\mathbf{X}^\top - \tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top\|_{op}$ small if larger data entries sampled with higher probability

Forming the sketch $\tilde{\mathbf{X}}$

- Thresholding-based scheme

$$\tilde{\mathbf{X}}_{ij} = \begin{cases} \mathbf{X}_{ij}, & \text{if } |\mathbf{X}_{ij}| \geq \delta \\ 0, & \text{else} \end{cases}$$

- (ℓ_1, ℓ_2) element sampling: Sample index (i, j) w.p. p_{ij}

$$\tilde{\mathbf{X}}_{ij} = \begin{cases} \frac{1}{p_{ij}} \mathbf{X}_{ij}, & \text{w.p. } p_{ij} \\ 0, & \text{w.p. } 1 - p_{ij} \end{cases}$$

  - Note that $\mathbb{E}\tilde{\mathbf{X}}_{ij} = p_{ij} \frac{\mathbf{X}_{ij}}{p_{ij}} + (1 - p_{ij})0 = \mathbf{X}_{ij}$

(ℓ_1, ℓ_2) sampling based sketch

- Choose p_{ij} as follows ($\alpha \in (0,1)$)

$$p_{ij} = \alpha \frac{|\mathbf{X}_{ij}|}{\|\mathbf{X}\|_1} + (1-\alpha)\frac{\mathbf{X}_{ij}^2}{\|\mathbf{X}\|_F^2}$$

- Biases the sampling scheme towards larger elements
- Reasonable way to model sampling in some cases
  Recommendation systems: users more likely to rate items they like/dislike a lot (large positive/large negative)

More details: thresholding scheme

- Let $\mathbf{X} = \tilde{\mathbf{X}} + \Delta$. This gives $\|\Delta\|_F^2 = \sum_{|\mathbf{X}_{ij}|<\delta} \mathbf{X}_{ij}^2$
- Let $\tilde{r} = \frac{\|\mathbf{X}\|_F^2}{\|\mathbf{X}\|_{op}^2}$ be the stable rank of $\mathbf{X}$
- Then

$$\begin{aligned}\|\mathbf{X}\mathbf{X}^\top - \tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top\|_{op} &= \|\mathbf{X}\Delta + \Delta^\top\mathbf{X}^\top + \Delta^\top\Delta\|_{op} \\ &\leq 2\|\mathbf{X}\|_{op}\|\Delta\|_{op} + \|\Delta\|_{op}^2\end{aligned}$$

- Choose δ so that $\|\Delta\|_F \leq \frac{\epsilon}{\sqrt{\tilde{r}}}\|\mathbf{X}\|_F$
- Then, previous theorem gives

$$\mathrm{Tr}(\tilde{\mathbf{S}}_k^\top \mathbf{X}\mathbf{X}^\top \tilde{\mathbf{S}}_k) \geq \mathrm{Tr}(\tilde{\mathbf{S}}_k^\top \mathbf{X}\mathbf{X}^\top \tilde{\mathbf{S}}_k) - 2k\epsilon(1+\epsilon)\|\mathbf{X}\|_{op}^2$$

Main result–II

- Sample complexity for (ℓ_1, ℓ_2) scheme

  Sample s entries from $\mathbf{X} \in \mathbb{R}^{d \times n}$ to form the sparse sketch $\mathbf{X}$ using the (ℓ_1, ℓ_2) scheme. Let $\mathbf{S}_k$ and $\tilde{\mathbf{S}}_k$ be solutions to the Sparse PCA problem using $\mathbf{X}$ and $\tilde{\mathbf{X}}$, respectively. Then, if the number of samples satisfies

$$s \geq 2k^2 \epsilon^{-2} (\rho^2 + \frac{\epsilon}{3k}) \log(\frac{d+n}{\delta})$$

  with $\rho^2 = \tilde{r} \max(d, n) f(\alpha, \mathbf{X})$, we have that

$$\mathrm{Tr}(\tilde{\mathbf{S}}_k^\top \mathbf{X}\mathbf{X}^\top \tilde{\mathbf{S}}_k) \geq \mathrm{Tr}(\tilde{\mathbf{S}}_k^\top \mathbf{X}\mathbf{X}^\top \tilde{\mathbf{S}}_k) - 2k\epsilon(1+\epsilon)\|\mathbf{X}\|_{op}^2$$

  w.p. at least $1 - \delta$.

Sparse PCA with Missing Observations[2]

- Samples $\mathbf{X}_1, \ldots, \mathbf{X}_n$ in $\mathbb{R}^d$ with mean zeros and covariance Σ
- Goal is to estimate the first principal component in the high dimensional ($d > n$) and missing data regime
- Covariance matrix of the data represented as

$$\mathbf{\Sigma} = \sigma_1 \theta_1 \theta_1^\top + \sigma_2 \Gamma,$$

where θ_1 is the first principal component, $\sigma_1, \sigma_2 \geq 0$ and $\Gamma \succeq 0$

[2] Karim Lounici. "Sparse Principal Component Analysis with Missing Observations". In: *High Dimensional Probability VI*. ed. by Christian Houdré et al. Basel: Springer Basel, 2013, pp. 327–356.

Missingness model

- For each sample $\mathbf{X}_i$, we observe its j^{th} entry $\mathbf{X}_{ij}$, independently of other entries, w.p. δ
- Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be the observations where

$$\mathbf{Y}_{ij} = \delta_{ij} \mathbf{X}_{ij}$$

  and $\delta_{ij} \overset{iid}{\sim} \mathrm{Ber}(\delta)$
- In practice, δ can be approximated using the fraction of observed entries per sample

Estimation procedure

- Recovering the first principal component: Let Σ_n denote the sample covariance matrix
  - Sparsity level known

$$\hat{\theta}_1 = \underset{\theta \in S^d:\ \|\theta\|_0 \leq s}{\arg\max}\ \theta^\top \Sigma_n \theta$$

  - Sparsity level unknown

$$\hat{\theta}_1 = \underset{\theta \in S^d}{\arg\max}\, \theta^\top \Sigma_n \theta - \lambda \|\theta\|_0$$

Estimation procedure

- Missing data case
  - Sample covariance $\Sigma_n^\delta = \frac{1}{n}\mathbf{Y}\mathbf{Y}^\top$ formed using incomplete samples is biased
  - Can apply the following correction to get an unbiased estimate of Σ

$$\tilde{\Sigma}_n = \frac{1}{\delta^2}\Sigma_n^\delta + \left(\frac{1}{\delta} - \frac{1}{\delta^2}\right)\mathrm{diag}(\Sigma_n^\delta)$$

- Final estimator

$$\hat{\theta}_1 = \arg\max_{\theta \in S^d} \theta^\top \Sigma_n \theta - \lambda \|\theta\|_0$$

Guarantees

- For $\mathbf{X}_1, \ldots, \mathbf{X}_n$ subgaussian and $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ defined as before, let

$$\lambda = \frac{\sigma_1^2}{\sigma_1 - \sigma_2} \frac{\log(ed)}{\delta^2 n}.$$

  Then the estimate $\hat{\theta}_1$ satisfies

$$\|\hat{\theta}_1 \hat{\theta}_1^\top - \theta_1 \theta_1^\top\|_F^2 \leq c \|\theta\|_0 \tilde{\sigma}^2 \frac{\log(ed)}{\delta^2 n}$$

  w.p. at least $1 - \frac{1}{d}$, where $\tilde{\sigma} = \frac{\sigma_1}{\sigma_1 - \sigma_2}$

- Bound increases as $\sigma_1 - \sigma_2$ decreases–problem harder when separation between first and second singular values decreases

Remarks

- Choice of λ depends on (unknown) singular values of Σ; a choice of λ based on singular values of $\tilde{\Sigma}_n$ is also given
- For fully observed case, bound shows that roughly $\|\theta\|_0 \log(d)$ samples suffice
- For missing data case, additional $\frac{1}{\delta^2}$ penalty
- The δ^{-2} penalty is shown to be tight in the lower bound result

Thank you