Sparse PCA

Gauri Jagatap Vahid Daneshpajooh

Term project
EE 525 Data Analytics for ECpE
Iowa State University

April 24, 2017

Outline

Motivation

- Principal component analysis (PCA) is a widely used tool for dimensionality reduction; the first few principal components retain the most variation in the data.
- The main drawback of PCA is that the principal components are not interpretable, in the physical sense.
- This is because in most cases, the principal components are a linear combination of *all* of the variables or features of the data.

Sparse PCA

- In several applications of PCA, like biology, spectroscopy and financial econometrics, variables have specific physical meanings attached to them.
  - In the case of gene-expression data, each variable represents the expression-level of particular gene.
  - Goal: identify simple structures in the genome, that involve only a few genes but explaining the most variance of the data.

Sparse PCA

- Solution: sparse PCA!
  - a trade-off between the expressive power of the principal components (in terms of explaining the maximum variance of the data), and interpretability, by retaining only few variables.

- Apart from interpretability, sparse PCA can also be used to identify clusters using fewer variables.
  - Further dimensionality reduction.
  - More efficient clustering techniques.

Mathematical Model

- A data matrix X comprises of n samples, each corresponding to p-dimensional feature vectors, arranged as rows.

- Given a mean-subtracted data matrix $X \in \mathbb{R}^{n \times p}$, compute the first k-sparse principal direction:
  - Same as the $k-$sparse eigenvector corresponding to the maximum eigenvalue of $X^T X$.

$$v^* = \arg\max_{z \in \mathbb{R}^p} \quad z^T (X^T X) z \tag{1}$$
$$\text{subject to} \quad \|z\|_2 = 1, \quad \|z\|_0 \leq k.$$

Truncated Power Method

The truncated power method for sparse eigenvalue problems [1] uses an iterative thresholding based on sparsity factor k, as long as:

- the first principal direction v_1 is sparse with cardinality $k = \|v_1\|_0$, and,
- the largest eigenvalue λ_1 corresponding to v_1 of $X^T X$ is non-degenerate.

Truncated Power Method

Algorithm

Input: matrix $A = X^T X \in \mathbf{S}^{p \times p}$, initial vector $z_0 \in \mathbb{R}^p$, cardinality $k \in \{1, 2 \dots p\}$.

Output: z, top singular vector of X.

Repeat:

$$\text{Compute} \quad z_t' = \frac{A x_{t-1}}{\|A x_{t-1}\|_2},$$

$$\text{Let} \quad F_t = supp(x_t', k) \quad \text{largest k terms},$$

$$\text{Compute} \quad x_t = truncate(x_t', F_t),$$

$$\text{Normalize} \quad x_t = \frac{x_t}{\|x_t\|_2},$$

$$t \leftarrow t + 1.$$

Until: Convergence

*Subsequent J singular vectors can be obtained by repeating this procedure on the residual matrix $\left(A - \sum_{j=1}^{J-1} d_j u_j u_j^T\right)$ where $A = UDU^T = \sum_{i=1}^{p} d_i u_i u_i^T$ and $u_j = z$ from the j^{th} run of TPM.

Inverse Power Method

for non-linear eigenproblems

- Standard eigenproblem for symmetric $A \epsilon \mathbb{R}^{n \times n}$ is of form

$$Af - \lambda f = 0$$

- For symmetric matrix *A*, eigenvectors of *A* can be characterized as critical points of "*Rayleigh quotient*" ,

$$R(f) = \frac{< f, Af >}{||f||_2^2}$$

- The IPM for spectral clustering and SPCA, minimizes the *inverse* of the Rayleigh objective, enforcing sparsity using:
  - The convex combination of *L1* and *L2* norms
  - Sparsity controlling parameter α

$$f^* = arg \min_f \frac{||f||_2^2}{< f, Af >} \xrightarrow{enforce\ sparsity} F(f) = \frac{(1-\alpha)\|f\|_2 + \alpha||f||_1}{||Xf||_2}$$

Inverse Power Method

Algorithm

- **Input:** data matrix X, sparsity controlling parameter α, accuracy ε
- **Output:** f, top singular vector of X
- **Repeat:**
  - $g^{k+1} = sign(\mu^k)(\lambda^k|\mu^k| - \alpha)_+$,
  - $f^{k+1} = \frac{g^{k+1}}{\|Xg^{k+1}\|_2}$,
  - $\lambda^{k+1} = (1-\alpha)\|f^{k+1}\|_2 + \alpha\|f^{k+1}\|_1$,
  - $\mu^{k+1} = \frac{\mathrm{A}f^{k+1}}{||Xf^{k+1}||_2}$,
- **Until:**
  - $\frac{|\lambda^{k+1}-\lambda^k|}{\lambda^k} < \epsilon$

where,

- $g^{k+1} = arg\min_{||f||_2 \le 1}(1-\alpha)\|f\|_2 + \alpha\|f\|_1 - \lambda^k < f, \mu^k >$
- $\mu^k = \frac{\mathrm{A}f^k}{\sqrt{<f^k, \mathrm{A}f^k>}}$

TextMining

Experimental Results

Dataset: Original data: 9394 (n) x 36,771 (p).
Used subset: 2200 x 2000.

Cardinality of sparse loadings: [5 7 2]

Output using both SparsePCA - TPM and IPM:

```
topic 1:
united iraq weapons un iraqi
topic 2:
bill companies congress industry tobacco legislation
smoking
topic 3:
spkr voice
```

Proportion of explained variance: **0.122** in both cases.
Execution time: TPM=**1.5066 sec** , IPM=**0.4437 sec**.

PitProps

Experimental Results

The PitProps dataset contains 180 observations of 13 variables.
We only have access to the correlation matrix $A = X^T X \in \mathbb{R}^{13\times 13}$.
Cardinality of the first five principal components: $[6 \quad 5 \quad 5 \quad 4 \quad 4]$.
Proportion of explained variance: 0.734.

PCs	topd	length	moist	testsg	ovensg	ringt	ringb	bowm	bowd	whorls	clear	knots	diaknot
PC1	0.4788	0.4625	0.3296	0.3802	0	0.3815	0.3976	0	0	0	0	0	0
PC2	0	0	-0.2808	0	0	0	0	0.5208	0.4600	0.5527	0	-0.3643	0
PC3	0	0	0.2761	0	-0.5086	-0.4338	-0.4362	0	0	0	0	0	0.5355
PC4	0	0	0	0	0	0	0	0.1834	0.2284	-0.2940	0.9098	0	0
PC5	0	0	0	0.5595	0.7002	0	-0.2846	0	0	0	0	0	0.3401

Table: Sparse PCA - TPM

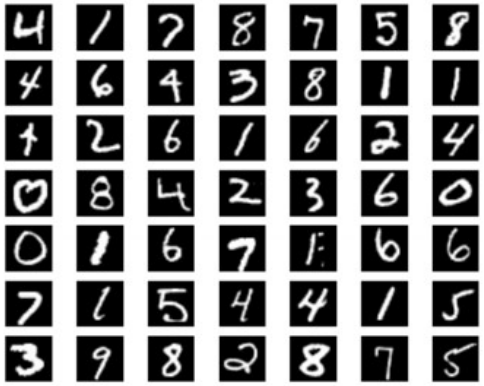
PCs	topd	length	moist	testsg	ovensg	ringt	ringb	bowm	bowd	whorls	clear	knots	diaknot
PC1	**-0.4038**	**-0.4055**	**-0.1244**	**-0.1732**	-0.0572	**-0.2844**	**-0.3998**	-0.2936	-0.3566	-0.3789	0.0111	0.1151	0.1125
PC2	-0.2179	-0.1861	**-0.5406**	-0.4556	0.1701	0.0142	0.1896	**0.1892**	**-0.0171**	**0.2485**	-0.2053	**-0.3432**	-0.3085
PC3	-0.2073	-0.2350	**0.1415**	0.3524	**0.4812**	**0.4753**	**0.2531**	-0.2431	-0.2076	-0.1188	-0.0705	0.0920	**-0.3261**
PC4	-0.0912	-0.1027	0.0784	0.0548	0.0491	-0.0634	-0.0650	**0.2855**	**0.0967**	**-0.2050**	**0.8037**	-0.3008	-0.3034
PC5	-0.0826	-0.1128	0.3498	**0.3558**	**0.1761**	-0.3158	**-0.2151**	0.1853	-0.1061	0.1564	-0.3430	-0.6004	**0.0799**

Table: PCA

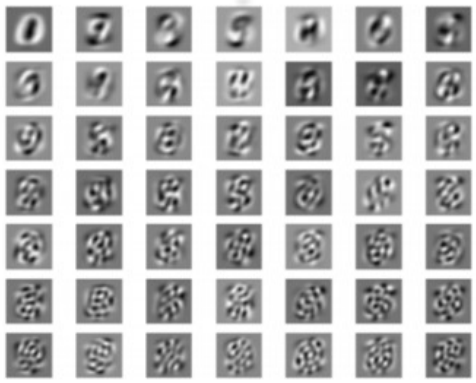
Classification of MNIST dataset

Experimental Results

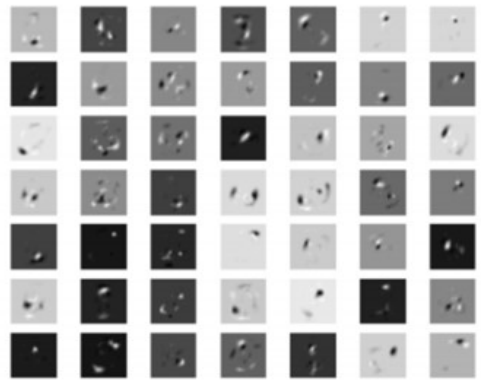
- MNIST is a dataset of handwritten digits.
- Classification using k-NN algorithm with Euclidean distance as the distance measure.



A random selection of 49 handwritten digit images from the MNIST training dataset



49 PCA loadings



49 SPCA loadings

Features	Classification error
49 PCA loadings	0.075
49 SPCA loadings	0.054

Comparison

So which method is better?

- Depends on the application and data available!
  - TPM requires just the correlation matrix $A = X^T X$, whereas IPM requires the actual data matrix X.
  - If $p >> n$, it is difficult to store the correlation matrix, so IPM should be preferred.
  - If $p << n$, TPM is computationally faster, so TPM should be preferred.

Further Study: Performance

Testing performance of TPM and IPM on toy dataset:

- Computational complexity/running time.
- Explained variance v/s cardinality trade-off.

Summary

So why Sparse PCA?

- Useful applications in data *analysis*.
- Picking out most commonly used important words in texts.
- Building financial models based on few important parameters.
- Spectral initialization in non-convex signal recovery algorithms (for specific random measurement schemes).
- Data compression.
- Efficient clustering in gene expression data.

For Further Reading I

X.-T. Yuan and T. Zhang, "Truncated power method for sparse eigenvalue problems," *Journal of Machine Learning Research*, vol. 14, no. Apr, pp. 899–925, 2013.

M. Hein and T. Bühler, "An inverse power method for nonlinear eigenproblems with applications in 1-spectral clustering and sparse pca," in *Advances in Neural Information Processing Systems*, pp. 847–855, 2010.