Sparse principal component Methods

Xiong Zhao

2021 年 12 月 22 日

中国人民大學
RENMIN UNIVERSITY OF CHINA

目录

Introduction

- Principal component analysis (PCA) is widely used in data processing and dimensionality reduction.
- However, PCA suffers from the fact that each principal component is a linear combination of all the original variables, thus it is often difficult to interpret the results.

Introduction

- Principal component regression (PCR) is a two-stage procedure that selects some principal components and then constructs a regression model regarding them as new explanatory variables.
- Note that the principal components are obtained from only explanatory variables and not considered with the response variable. For this, the prediction accuracy of the PCR could be low, if the response variable is related to principal components having small eigenvalues.

Introduction

- Without loss of generality, assume the column means of $\mathbf{X}$ are all 0 . Let the SVD of $\mathbf{X}$ be

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

- $\mathbf{Z} = \mathbf{U}\mathbf{D}$ are the principal components (PCs), and the columns of $\mathbf{V}$ are the corresponding loadings of the principal components.

Principal Component and Loss Function

- PCA can be written as a regression-type optimization problem, with a quadratic penalty:

$$\min_{B} \sum_{i=1}^{n} \left\| \boldsymbol{x}_i - BB^T \boldsymbol{x}_i \right\|^2$$

$$\text{subject to } B^T B = I_k,$$

- where $B = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)$ is a $p \times k$ loading matrix, k denotes the number of principal components, and I_k is the $k \times k$ identity matrix. The solution is given by

$$\hat{B} = V_k Q^T$$

where $V_k = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$ and Q is a $k \times k$ arbitrary orthogonal matrix.

Sparse Principal Component Analysis

- We first discuss a simple sparse approach to PCA–direct sparse approximations:
- Theorem 1. For each i, denote by $Z_i = \mathbf{U}_i\mathbf{D}_{ii}$ the ith principal component. Consider a positive λ and the ridge estimates $\hat{\beta}_{\text{ridge}}$ given by

$$\hat{\beta}_{\text{ridge}} = \arg\min_{\beta} \|Z_i - \mathbf{X}\beta\|^2 + \lambda\|\beta\|^2.$$

Let $\hat{v} = \frac{\hat{\beta}_{\text{ridge}}}{\|\hat{\beta}_{\text{ridge}}\|}$, then $\hat{v} = V_i$.

Sparse Principal Component Analysis

- Theorem 1 depends on the results of PCA, so it is not a genuine alternative.We now present a "self-contained" regression-type criterion to derive sparse PCs:
- Theorem 2. For any $\lambda > 0$, let

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} \left\| \mathbf{x}_i - \alpha\beta^T \mathbf{x}_i \right\|^2 + \lambda \|\beta\|^2$$

subject to $\|\alpha\|^2 = 1$.
Then $\hat{\beta} \propto V_1$.

Sparse Principal Component Analysis

- For the whole sequence of PCs:
- Theorem3. Suppose we are considering the first k principal components. Let $\mathbf{A}_{p\times k} = [\alpha_1, \ldots, \alpha_k]$ and $\mathbf{B}_{p\times k} = [\beta_1, \ldots, \beta_k]$. For any $\lambda > 0$, let

$$(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}) = \arg\min_{\mathbf{A},\mathbf{B}} \sum_{i=1}^{n} \left\|\mathbf{x}_i - \mathbf{A}\mathbf{B}^T\mathbf{x}_i\right\|^2 + \lambda \sum_{j=1}^{k} \left\|\beta_j\right\|^2$$

  subject to $\mathbf{A}^T\mathbf{A} = I_{k\times k}$.
- Then $\hat{\beta}_j \propto V_j$ for $j = 1, 2, \ldots, k$.

Sparse Principal Component Analysis

- We carry on the connection between PCA and regression, and use the lasso approach to produce sparse loadings. For that purpose, we add the lasso penalty into the criterion and consider the following optimization problem:

$$(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}) = \arg\min_{\mathbf{A},\mathbf{B}} \sum_{i=1}^{n} \left\| \mathbf{x}_i - \mathbf{A}\mathbf{B}^T\mathbf{x}_i \right\|^2 + \lambda \sum_{j=1}^{k} \left\| \beta_j \right\|^2 + \sum_{j=1}^{k} \lambda_{1,j} \left\| \beta_j \right\|_1$$

$$\text{subject to } \mathbf{A}^T\mathbf{A} = I_{k\times k}.$$

- Whereas the same λ is used for all k components, different $\lambda_{1,j}$'s are allowed for penalizing the loadings of different principal components.

Supervised Principal Component Regression

- 1.choose the explanatory variables that are related to the response variable with respect to correlation.
- 2.make PCA in the explanatory variables that are related to the response variable with respect to correlation.
- 3.make regression analysis in the principle components.

Principal Component Regression by Principal Component Selection

- This method selects principal components using variable selection procedures instead of a small subset of major principal components in principal component regression:
- 1.reduce the number of principal components using the conventional principal component regression to yield the set of candidate principal components.
- 2.Select principal components among the candidate set using sparse regression techniques.

Sparse Principal Component Regression

- However, none of them integrated the two loss functions for ordinary regression analysis and PCA along with the L 1 type regularization.
- In general, principal components are irrelevant with the response variables. Therefore, PCR might fail to predict the response if the response is associated with principal components corresponding to small eigenvalues.

Sparse Principal Component Regression

- To overcome this drawback, we propose SPCR using the principal components $B^T\boldsymbol{x}$ as follows:

$$\min_{A,B,\gamma_0,\boldsymbol{\gamma}} \left\{ (1-w)\sum_{i=1}^{n}\left(y_i-\gamma_0-\boldsymbol{\gamma}^T B^T\boldsymbol{x}_i\right)^2 + w\sum_{i=1}^{n}\left\|\boldsymbol{x}_i - AB^T\boldsymbol{x}_i\right\|^2 \right.$$
$$\left. +\lambda_\beta(1-\zeta)\sum_{j=1}^{k}\|\boldsymbol{\beta}_j\|_1 + \lambda_\beta\zeta\sum_{j=1}^{k}\|\boldsymbol{\beta}_j\|^2 + \lambda_\gamma\|\boldsymbol{\gamma}\|_1 \right\}$$

subject to $A^TA = I_k$

- where γ_0 is an intercept, $\gamma = (\gamma_1,\ldots,\gamma_k)^T$ is a coefficient vector, λ_β and λ_γ are regularization parameters with positive value, and w and ζ are tuning parameters whose values are between zero and one.

Sparse Principal Component Regression

- The first term means the least squares loss between the response and the principal components $B^T x$.
- The second term induces PCA loss of data X.
  The tuning parameter w controls the trade-off between the first and second terms, and then the value of w can be determined by users for any purpose.

Sparse Principal Component Regression

- The third and fifth terms encourage sparsity on B and γ, respectively. The sparsity on B enables us to easily interpret the loadings of the principal components. Meanwhile, the sparsity on γ induces automatic selection of the number of principal components.
- The tuning parameter ζ controls the trade-off between the L_1 and L_2 norms for the parameter B.

Adaptive sparse principal component regression

- we observe that SPCR does not produce enough sparse solution for the loading matrix B. Let us consider the weighted sparse principal component regression, given by:

$$\min_{A,B,\gamma_0,\boldsymbol{\gamma}} \left\{ (1-w)\sum_{i=1}^{n}\left(y_i-\gamma_0-\boldsymbol{\gamma}^T B^T \boldsymbol{x}_i\right)^2 + w\sum_{i=1}^{n}\left\|\boldsymbol{x}_i - AB^T\boldsymbol{x}_i\right\|^2 \right.$$
$$\left. +\lambda_\beta(1-\zeta)\sum_{j=1}^{k}\sum_{l=1}^{p}\omega_{lj}\left|\beta_{lj}\right| + \lambda_\beta\zeta\sum_{j=1}^{k}\left\|\boldsymbol{\beta}_j\right\|^2 + \lambda_\gamma\|\boldsymbol{\gamma}\|_1 \right.$$

subject to $A^TA = I_k$

- where $\omega_{lj}(>0)$ is an incorporated weight for the parameter β_{lj}. We call this procedure the adaptive sparse principal component regression (aSPCR).

Sparse Principal Component Regression

- For the SPCR, it is unclear whether the PCA loss function in Zou et al. (2006) is the best choice for building SPCR, as there exist several formulae for PCA.
- This paper proposes a novel formulation for SPCR. As a PCA loss for SPCR, we adopt a loss function based on a singular value decomposition approach (Shen and Huang 2008).

Sparse Principal Component Regression

- We consider the following minimization problem:

$$\min_{\beta_0,\boldsymbol{\beta},Z,V} \left\{ \frac{1}{n} \|\boldsymbol{y} - \beta_0 \mathbf{1}_n - XV\boldsymbol{\beta}\|_2^2 + \frac{w}{n} \left\|X - ZV^\top\right\|_F^2 + \lambda_V \|V\|_1 + \lambda_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 \right\}$$
$$\text{subject to } V^\top V = I_k,$$

- where β_0 is an intercept, k is the number of PCs, $\boldsymbol{\beta}$ is a k-dimensional coefficient vector, Z is an $n \times k$ matrix of PCs, V is a $p \times k$ PC loading matrix, and $\mathbf{1}_n$ is an n-dimensional vector of ones.
- In addition, w is a positive tuning parameter and λ_V, λ_β are non-negative regularization parameters.

Sparse Principal Component Regression

- The first term is the regression squared loss function relating the response and the PCs XV. The second term is the PCA loss function in the SVD approach in Shen and Huang (2008). The third and fourth terms constitute the lasso penalty that induces zero estimates of the parameters V and β,respectively.
- We call this method SPCRsvd. We will observe that SPCRsvd is competitive with or better than SPCR through numerical studies.

Sparse Principal Component Regression

- We remark on two points here. First, it is possible to use Z in the first term instead of XV, since Z is also the PCs. However, the formulation with Z instead of XV did not perform well in numerical studies, so we adopt the formulation with XV here.
- Second, SPCR imposes a ridge penalty for the PC loading but SPCRsvd does not. The ridge penalty basically comes from SPCA in Zou et al. (2006). Because SPCRsvd is not based on SPCA in Zou et al. (2006), a ridge penalty does not appear. It is possible to add a ridge penalty and replace the lasso penalty with other penalties that induce sparsity, e.g., the adaptive lasso penalty, the SCAD penalty, or minimax concave penalty (Zhang 2010).