A tutorial on sparse principal components analysis

Keshav Motwani

December 10, 2020

Principal components analysis (PCA)

Idea: high-dimensional data lives in a lower dimensional subspace

- Find linear combinations of input features that contain directions of variability in the data

$$\underset{w_i:\,||w_i||_2^2=1}{\arg\max}\ \mathrm{Var}(Xw_i) \quad \text{s.t.} \quad w_j' w_k = 0 \text{ for all } j \neq k$$

- Solution is leading eigenvectors of $X'X$ or equivalently right singular vectors of X
- Turns out to also minimize reconstruction error

$$\underset{W:W'W=I}{\arg\min}\, ||X - XWW'||_F^2$$

since we can show

$$||X - XWW'||_F^2 = \mathrm{tr}(X'X) - \mathrm{tr}(W'X'XW)$$

Sparse PCA

Regular principal components are a weighted sum of ALL features in the data

- Not very interpretable
- Inducing sparsity in the loading vectors w_i can solve this
  - Each principal component is then only the weighted sum of a subset of features in the data
- Many proposed methods for this task
  - Zou, Hastie, and Tibshirani proposed a method based on reduced-rank regression in 2006
  - Witten, Hastie, and Tibshirani proposed a method based on a penalized matrix decomposition in 2010

PCA as reduced-rank regression

Consider the following regression problem:

$$X = XBA' + E, \quad X \in \mathbb{R}^{n \times p}, \quad B, A \in \mathbb{R}^{p \times r}, \quad E \in \mathbb{R}^{n \times p}, \quad A'A = I_r$$

- BA' is a rank-r matrix of regression coefficients
- E is an error matrix

To minimize the errors, we want the following:

$$\underset{A,B:\, A'A = I_r}{\arg\min} \; ||X - XBA'||_F^2$$

- Similar to minimizing reconstruction error in normal PCA, but here we have two separate matrices B and A
- Columns of B are related to loading vectors

Equivalence of regression form and PCA when $n > p$

Want to solve

$$\underset{A, B: A'A = I_r}{\arg\min} ||X - XBA'||_F^2$$

- With A fixed, we can construct a matrix $A_\perp \in \mathbb{R}^{p\times(p-r)}$ such that $[A, A_\perp]'[A, A_\perp] = [A, A_\perp][A, A_\perp]' = I_p$.
- With this, we can show:

$$\begin{aligned} ||X - XBA'||_F^2 &= ||(X - XBA')[A, A_\perp]||_F^2 \\ &= ||XA - XB||_F^2 + ||XA_\perp||_F^2. \end{aligned}$$

- Taking the gradient with respect to B and setting it to 0, we get

$$\begin{aligned} -X'(XA - XB) &= 0 \\ \implies B &= (X'X)^{-1}X'XA \\ \implies \hat{B} &= A \end{aligned}$$

- With B fixed at $\hat{B} = A$, we get the minimum reconstruction error formulation of regular PCA

Addition of ridge penalty when $p > n$

Want to solve

$$\underset{A,B:\,A'A=I_r}{\arg\min}\ ||X - XBA'||_F^2 + \lambda||B||_F^2$$

- With A fixed, we can construct a matrix $A_\perp$ as in last slide
- Again, we have

  $$||X - XBA'||_F^2 + \lambda||B||_F^2 = ||XA - XB||_F^2 + ||XA_\perp||_F^2 + \lambda||B||_F^2$$

- Taking the gradient with respect to B and setting it to 0, we get

$$\begin{aligned} -X'(XA - XB) + \lambda B &= 0 \\ \implies X'XA &= (X'X + \lambda I_p)B \\ \implies \hat{B} &= (X'X + \lambda I_p)^{-1}X'XA \end{aligned}$$

- We can also show

$$\begin{aligned} -X'(XA - XB) + \lambda B &= 0 \\ \implies \lambda||\hat{B}||_F^2 &= \mathrm{tr}(\hat{B}'X'(XA - X\hat{B})). \end{aligned}$$

Addition of ridge penalty when $p > n$

Want to solve

$$\underset{A,B:\, A'A=I_r}{\arg\min} \; ||X - XBA'||_F^2 + \lambda ||B||_F^2$$

- With B fixed at $\hat{B}$, we need to minimize

$$\begin{aligned} C_\lambda(A, \hat{B}) &= ||X - X\hat{B}A'||_F^2 + \lambda ||\hat{B}||_F^2 \\ &= \mathrm{tr}(X'X) - \mathrm{tr}(A'X'X(X'X + \lambda I_p)^{-1} X'XA) \end{aligned}$$

  such that $A'A = I_r$
- By similar argument as that in maximum variance formulation of PCA, solution is the first r eigenvectors of $X'X(X'X + \lambda I_p)^{-1} X'X$
- Letting $UDV' = X$ be the SVD of X, we can show

$$X'X(X'X + \lambda I_p)^{-1} X'X = VD^2(D^2 + \lambda I_p)^{-1} D^2 V'$$

  so $\hat{A} = V_{\cdot,1:r}$

Addition of ridge penalty when $p > n$

- In summary, we have

$$\hat{B} = (X'X + \lambda I_p)^{-1} X'XA$$

$$\hat{A} = V_{\cdot,1:r}$$

- Letting $UDV' = X$ be the SVD, we can show

$$\hat{B} = V(D^2 + \lambda I_p)^{-1} D^2 V'A$$

  and thus with $A = \hat{A}$, we have

$$\hat{B} = V_{\cdot,1:r} \left[(D^2 + \lambda I_p)^{-1} D^2 \right]_{1:r,1:r}$$

- In other words, $\hat{B}$ is simply a scaled version of V, the loading vectors from regular PCA
- Since regular loading vectors are unit vectors, we can recover them with $w_i = \frac{B_{\cdot,i}}{||B_{\cdot,i}||_2}$

Extension to sparsity

We want sparsity in columns of B, so we can add another penalty

$$\underset{A,B:\,A'A=I_r}{\arg\min}\ ||X - XBA'||_F^2 + \lambda||B||_F^2 + ||B\,\mathrm{diag}(\lambda_{1,1}, \ldots, \lambda_{1,r})||_1$$

where $\lambda_{1,1}, \ldots, \lambda_{1,r}$ are tuning parameters to control the degree of sparsity in each loading vector separately

Extension to sparsity

Once again, with A fixed, we want to minimize

$$\begin{aligned} C_{\lambda,\lambda_1}(A,B) &= ||X - XBA'||_F^2 + \lambda||B||_F^2 + ||B\,\mathrm{diag}(\lambda_{1,1},\ldots,\lambda_{1,r})||_1 \\ &= ||XA - XB||_F^2 + ||XA_\perp||_F^2 + \lambda||B||_F^2 \\ &+ ||B\,\mathrm{diag}(\lambda_{1,1},\ldots,\lambda_{1,r})||_1. \end{aligned}$$

The terms with B can be rewritten as

$$\sum_{i=1}^{r} ||XA_{\cdot,i} - XB_{\cdot,i}||_2^2 + \lambda||B_{\cdot,i}||_2^2 + \lambda_{1,r}||B_{\cdot,i}||_1$$

- Same as solving r independent elastic net regression problems

Extension to sparsity

With B fixed, we want to minimize

$$\begin{aligned} ||X - XBA'||_F^2 &= \mathrm{tr}((X - XBA')'(X - XBA')) \\ &= \mathrm{tr}(X'X) - 2\,\mathrm{tr}(X'XBA') + \mathrm{tr}(AB'X'XBA') \\ &= \mathrm{tr}(X'X) - 2\,\mathrm{tr}(X'XBA') + \mathrm{tr}(B'X'XB) \end{aligned}$$

subject to $A'A = I_r$. Thus we need to maximize

$$\mathrm{tr}(X'XBA')$$

If we let $UDV' = X'XB$, we get

$$\mathrm{tr}(X'XBA') = \mathrm{tr}(UDV'A') = \mathrm{tr}(V'A'UD)$$

Since D is diagonal, we want to maximize the elements of the diagonal of $V'A'U$. Since $V'A'U$ is orthogonal, this is maximized by letting $V'A'U = I_r$, so $V'A' = U'$ or $A = UV'$

Algorithm

- Initialize A to regular right singular vectors and B to solution of elastic net regression with the fixed A
- Until convergence
  - Update A with B fixed
  - Update B with A fixed

Sparse PCA as penalized matrix decomposition

- From SVD of $X = UDV'$, loading vectors are columns of V
- Well known result by Eckart and Young that best rank-r approximation to X is

$$\arg\min_{\hat{X} \in M(r)} ||X - \hat{X}||_F^2 = U_{\cdot,1:r} D_{1:r,1:r} V'_{\cdot,1:r}$$

- Considering rank-1 approximation first, we want to solve

$$\arg\min_{d,u,v} ||X - duv'||_F^2$$

$$u'u = 1, \quad v'v = 1, \quad ||v||_1 \leq c, \quad d \geq 0$$

where L_1 constraint induces sparsity in v

Rank-1 approximation with sparse v

- Problem equivalent to

$$\arg\min_{d,u,v} -2du'Xv + d^2$$

$$u'u = 1, \quad v'v = 1, \quad ||v||_1 \leq c, \quad d \geq 0$$

- u and v that solve the above must also solve

$$\arg\max_{u,v} u'Xv$$

$$u'u = 1, \quad v'v = 1, \quad ||v||_1 \leq c$$

with $d = u'Xv$

- To make it a biconvex problem, relax equality constraints

$$\arg\max_{u,v} u'Xv$$

$$u'u \leq 1, \quad v'v \leq 1, \quad ||v||_1 \leq c$$

Rank-1 approximation with sparse v

- With v fixed, we want to solve

$$\arg\min_u -u'Xv \quad \text{subject to} \quad u'u \leq 1$$

- Can show that

$$u = \frac{Xv}{||Xv||_2}$$

  satisfies the KKT conditions

Rank-1 approximation with sparse v

- With u fixed, we want to solve

$$\arg\min_v -u'Xv \quad \text{subject to} \quad v'v \leq 1, \quad ||v||_1 \leq c$$

- Can show that

$$v = \frac{S(X'u, \Delta)}{||S(X'u, \Delta)||_2}$$

  satisfies the KKT conditions, where S is the soft-thresholding operator $S(a, \Delta) = \text{sgn}(a)(|a| - \Delta)_+$ and Δ is chosen by binary search to satisfy the constraint on v

Algorithm

- Initialize v to regular first right singular vector
- Until convergence
  - Update u with v fixed
  - Update v with u fixed
  - Set $d = uXv'$

Extension to multiple sparse principal components

- For $k > 1$, to obtain d_k, u_k, and v_k, repeat algorithm for rank-1 approximation on

$$X - \sum_{i=1}^{k-1} d_i u_i v_i'$$

Extension to multiple sparse principal components with orthogonal U

- If we let $U_{k-1} = [u_1, \ldots, u_{k-1}]$, we want to solve

$$\arg\max_{u_k, v_k} u_k' X v_k$$

$$u_k' u_k = 1, \quad v_k' v_k = 1, \quad ||v_k||_1 \leq c, \quad U_{k-1}' u_k = 0$$

- In other words, solution to u_k in the column space of $U_{k-1}^{\perp}$, where $U_{k-1}^{\perp}$ is a matrix with columns as basis vectors orthogonal to U_{k-1}
- We want $u_k = U_{k-1}^{\perp}\theta$ for some θ

$$\arg\max_{\theta} \theta' U_{k-1}^{\perp'} X v_k \quad \text{subject to} \quad \theta'\theta \leq 1$$

$$\theta = \frac{U_{k-1}^{\perp'} X v_k}{||U_{k-1}^{\perp'} X v_k||_2}$$

Extension to multiple sparse principal components with orthogonal U

- From the last slide, we have

$$\theta = \frac{U_{k-1}^{\perp\prime} X v_k}{||U_{k-1}^{\perp\prime} X v_k||_2}$$

- Therefore, since $u_k = U_{k-1}^{\perp}\theta$,

$$u_k = \frac{U_{k-1}^{\perp} U_{k-1}^{\perp\prime} X v_k}{||U_{k-1}^{\perp\prime} X v_k||_2}$$

- How do we compute this?

Extension to multiple sparse principal components with orthogonal U

- First note that

$$||U^{\perp}_{k-1}U^{\perp'}_{k-1}Xv_k||_2 = ||U^{\perp'}_{k-1}Xv_k||_2$$

so if we can solve for

$$u^*_k = U^{\perp}_{k-1}U^{\perp'}_{k-1}Xv_k$$

then $u_k = \frac{u^*_k}{||u^*_k||_2}$

- We can see that $U^{\perp}_{k-1}U^{\perp'}_{k-1}$ is an orthogonal projection through connection with least squares regression

$$U^{\perp}_{k-1}U^{\perp'}_{k-1}Xv_k = U^{\perp}_{k-1}(U^{\perp'}_{k-1}U^{\perp}_{k-1})^{-1}U^{\perp'}_{k-1}Xv_k.$$

and thus we have

$$U_{k-1}(U'_{k-1}U_{k-1})^{-1}U'_{k-1} = I_r - U^{\perp}_{k-1}(U^{\perp'}_{k-1}U^{\perp}_{k-1})^{-1}U^{\perp'}_{k-1}$$

Conclusion

- Many interesting formulations of PCA with extensions to sparsity
- Detailed proofs for claims mentioned here are in write-up
- Implementations for these methods also provided and explained in write-up