MS-E2112 Multivariate Statistical Analysis (5cr)
Lecture 3: Principal Component Analysis - part II

Lecturer: Pauliina Ilmonen
Slides: Ilmonen/Kantala

Contents

PCA Using Correlation Matrix

Correlation Structure in PCA

Multivariate Linear Regression

PCA in Regression Analysis

References

PCA Using Correlation Matrix

PCA Using Correlation Matrix

As was pointed out last week, PCA is highly sensitive for scaling of the variables. One can address this problem by standardizing the variables first. The data can be standardized by subtracting the sample mean $\bar{x}$, and then dividing each variable by the corresponding square root of the sample variance $\hat{\sigma}_{ii}$. PCA is then applied to this preprocessed data. Note that for standardized variables, the covariance matrix Σ turns into a correlation matrix.

PCA Using Correlation Matrix

If PCA is performed standardizing the variables first, it naturally becomes scale-invariant.

If variables do not have the same natural units, it is better to standardize the data first. For example, if the variables considered are weight, height, age, and IQ, it is a good idea to think about standardizing the data first. But if the variables do share the same units and if there are no large differences between the variances, then one can apply standard PCA.

PCA Using Correlation Matrix

One may address the problem of scale-sensitivity by standardizing the data first. However, this standardization does not make PCA fully invariant under all linear transformations.

Correlation Structure in PCA

Correlation Structure

Theorem

Let x denote a p-variate random vector with finite mean vector μ, and finite covariance matrix Σ. Let σ_{ii} denote the ith diagonal element of Σ. Let $y = \Gamma^T(x-\mu)$, where $\Gamma \in \mathbb{R}^{p \times p}$ is orthogonal, $\Gamma^T\Sigma\Gamma = \Lambda = diag(\lambda_1, \cdots, \lambda_p)$ and $\lambda_1 \geq \cdots \geq \lambda_p$. Let γ_j denote the jth column vector of Γ and let γ_{ij} denote the ith element of it (i.e. γ_{ij} denotes the ij element of Γ). Then

$$corr(x_i y_j) = \rho_{ij} = \frac{\gamma_{ij}\lambda_j}{\sqrt{\sigma_{ii}\lambda_j}}.$$

Correlation Structure

Proof.

Let x denote a p-variate random vector with finite mean vector μ, and finite covariance matrix Σ. Let σ_{ii} denote the ith diagonal element of Σ. Let $y = \Gamma^T(x-\mu)$, where $\Gamma \in \mathbb{R}^{p\times p}$ is orthogonal, $\Gamma^T\Sigma\Gamma = \Lambda = diag(\lambda_1, \cdots, \lambda_p)$ and $\lambda_1 \geq \cdots \geq \lambda_p$. Let γ_j denote the jth column vector of Γ and let γ_{ij} denote the ith element of it (i.e. γ_{ij} denotes the ij element of Γ). Now

$$E[(x-\mu)y^T] = E[(x-\mu)(\Gamma^T((x-\mu)))^T]$$

$$= E[((x-\mu))((x-\mu))^T\Gamma] = \Sigma\Gamma = \Gamma\Lambda.$$

Therefore the covariance between x_i and y_j is $\gamma_{ij}\lambda_j$. Since x_i and y_j have variances σ_{ii} and λ_j, respectively, the correlation between x_i and y_j is given by

$$\rho_{ij} = \frac{\gamma_{ij}\lambda_j}{\sqrt{\sigma_{ii}\lambda_j}}.$$

□

Correlation Structure

It can be said that "the proportion of the variation" of x_i explained by y_j is ρ_{ij}^2. Since the elements of y are uncorrelated, any set S of components explain a proportion

$$\rho_{iS}^2 = \sum_{j \in S} \rho_{ij}^2.$$

Note that when Σ is a correlation matrix, the variance $\sigma_{ii} = 1$ and thus $\rho_{ij} = \gamma_{ij}\sqrt{\lambda_j}$.

Multivariate Linear Regression

Multivariate Linear Regression

Regression analysis is used to predict the value of one or more responses from a set of predictors. Predictors can be continuous or categorical or a mixture of both.

Multivariate Regression Model

Let z be a p-variate random vector of dependent variables such that

$$z = B^T v + u,$$

where v is a q-variate fixed vector of predictors, B is a $q \times p$ matrix of regression parameters, and u is a p-variate vector of random errors with mean 0, and common covariance matrix C. The first element of v is assumed to be 1 (to allow a mean effect).

Multivariate Linear Regression

Assume that we have a size n sample from the multivariate regression model. Then Z is a $n \times p$ data matrix such that

$$Z = VB + U,$$

where V is a known $n \times q$ matrix, B is a $q \times p$ matrix, and U is a $n \times p$ matrix of unobserved random disturbances. The elements of the first column of V are all assumed to be 1, and the rows of U are assumed to be uncorrelated.

Estimation

Assume that Z is a $n \times p$ data matrix such that

$$Z = VB + U,$$

where V is a known $n \times q$ matrix, B is a $q \times p$ matrix, and the $n \times p$ error matrix U is independent of V. The elements of the first column of V are all assumed to be 1. Assume that the rows of the error matrix U are independent and identically distributed with the mean vector $\mu = 0$ and the covariance matrix C. Assume that the inverse of $V^T V$ exists.

Let

$$P = I - V(V^T V)^{-1} V^T.$$

Now, the generalized least squares estimators of B and C can be given as

$$\hat{B} = (V^T V)^{-1} V^T Z$$

and

$$\hat{C} = \frac{1}{n} Z^T P Z.$$

Estimation

The estimate $\hat{B}$ can be used in estimating/predicting the values of the matrix Z,

$$\hat{Z} = V\hat{B}.$$

The estimate of the error matrix is obtained by taking the difference between Z and $\hat{Z}$

$$\hat{U} = Z - V\hat{B}.$$

Trace Correlation and Determinant Correlation

Assume that the matrix Z is centered so that the columns of Z have zero mean. Define now

$$D = (Z^T Z)^{-1} \hat{U}^T \hat{U}.$$

The matrix $\hat{U}^T \hat{U}$ ranges between zero, when all the variation of Z is explained by the regression model, and $Z^T Z$, when no part of the variation in Z is explained by V. Therefore $I - D$ varies between the identity matrix and the zero matrix. It can be shown that all the eigenvalues of $I - D$ lie between 1 and 0.

Trace Correlation and Determinant Correlation

It would be desirable that a measure of multivariate correlation would range between zero and one. This property is satisfied by two often used coefficients, the trace correlation r_T and the determinant correlation r_D,

$$r_T^2 = \frac{1}{p} tr(I - D),$$

and

$$r_D^2 = det(I - D).$$

Note that the coefficient r_D is zero if at least one of the eigenvalues of $I - D$ is zero, and r_T is zero if and only if all the eigenvalues of $I - D$ are zero.

Some Comments

- We assumed that the inverse of $V^T V$ exists. If it does not (or if some of the columns of V are nearly collinear), consider using smaller number of variables.
- One should not use the regression model for predicting outside of the range of the Z values. Behavior of extreme points may be different!
- Traditional L_2 regression is very sensitive to outlying observations.

PCA in Regression Analysis

PCA in Regression Analysis

Linear regression analysis is unstable in the presence of multicollinearity, or near multicollinearity, of the predictors. In this situation, PCA can be used to preprocess the data. Instead of performing regression analysis using the original variables, one can perform it using new variables obtained from PCA

PCA in Regression Analysis

Linear regression analysis is unstable in the presence of highly linearly dependent predictors. This problem is often solved simply by disregarding some of the predictors. Alternatively, PCA can be used to preprocess the data. Instead of performing regression analysis using the original variables, one can perform it using new variables obtained from PCA.

PCA in Regression Analysis

In general, when PCA is used, the principal components with the largest variance are chosen in order to explain as much of the total variation of x as possible. In regression settings, the choice of the components is somewhat different. In the context of regression, it is sensible to choose the components having the largest correlation with the most interesting dependent variables, because the purpose is to use the components in explaining the dependent variables. Fortunately, there is often a tendency in data for the components with largest variances to best explain the dependent variables.

PCA in Regression Analysis

If the original regression equation is given by

$$z = B^T v + u,$$

then also

$$z = A^T w + u,$$

where $w = \Gamma^T z$, Γ^T is the principal component transformation matrix, and $A = \Gamma^T B$. For the corresponding sample version it also holds that if

$$Z = VB + U,$$

then

$$Z = WA + U,$$

where $W = VG$, and $A = G^T B$.

One can now reduce dimension by deleting some of the columns of W.

Next Week

Next week we will talk about robust principal component analysis.

References

References I

K. V. Mardia, J. T. Kent, J. M. Bibby, Multivariate Analysis, Academic Press, London, 2003 (reprint of 1979).

References II

- R. V. Hogg, J. W. McKean, A. T. Craig, Introduction to Mathematical Statistics, Pearson Education, Upper Sadle River, 2005.
- R. A. Horn, C. R. Johnson, Matrix Analysis, Cambridge University Press, New York, 1985.
- R. A. Horn, C. R. Johnson, Topics in Matrix Analysis, Cambridge University Press, New York, 1991.