Multidimensional Data Analysis

Lin Cong

University of California, Riverside

gradquant@ucr.edu

November 4, 2019

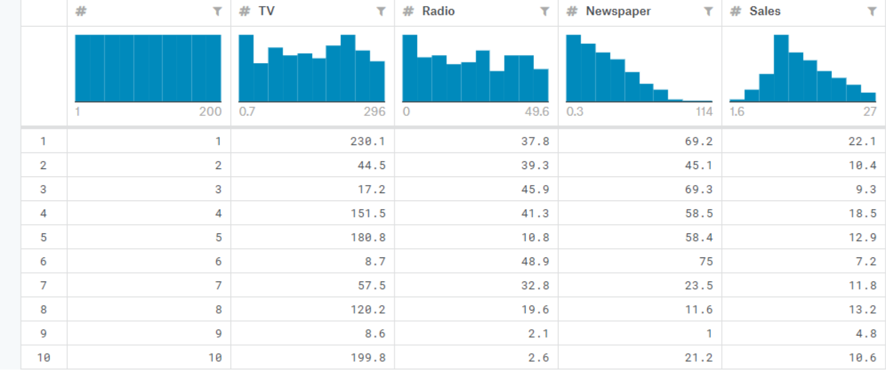
Contents

1. Descriptive Statistics
2. Multiple Linear Regression
3. Principle Component Analysis
4. Discussion

Descriptive Statistics

Introduction

Advertising data:
The dataset contains statistics about the sales of a product in 200 different markets, together with advertising budgets in each of these markets for different media channels: TV, radio and newspaper. The sales are in thousands of units and the budget is in thousands of dollars.



	#	# TV	# Radio	# Newspaper	# Sales
1	1	230.1	37.8	69.2	22.1
2	2	44.5	39.3	45.1	10.4
3	3	17.2	45.9	69.3	9.3
4	4	151.5	41.3	58.5	18.5
5	5	180.8	10.8	58.4	12.9
6	6	8.7	48.9	75	7.2
7	7	57.5	32.8	23.5	11.8
8	8	120.2	19.6	11.6	13.2
9	9	8.6	2.1	1	4.8
10	10	199.8	2.6	21.2	10.6

Descriptive Statistics

- Numerical summary of the descriptive statistics for all variables:
  summary() function
- Correlation matrix:
  corr() function
- Plots for univariate and bivariate cases can be applied for multidimensional data separately:
  hist() function, **plot()** function.
- Rotating Scatterplot:
  scatterplot3d() function
- Scatterplot Matrix:
  pairs() function

Multiple Linear Regression

Motivation

- Is there a relationship between advertising budget and sales?
- How accurately can we estimate the effect of different media on sales?
- Is the relationship linear?
- Is there synergy among the advertising media? Perhaps spending $50,000 on television advertising and $50,000 on radio advertising results in more sales than allocating $100,000 to either television or radio individually. In marketing, this is known as a synergy effect, while in statistics it is called an interaction effect.

Multiple Linear Regression

Model:

- Given the independent data $(x_{11}, x_{12}, ..., x_{1p}, y_1), (x_{21}, x_{22}, ..., x_{2p}, y_2), ..., (x_{n1}, x_{n2}, ..., x_{n3}, y_n)$, the multiple linear regression fits the data with the following model:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

- The least squares estimates(LSE) of the coefficients are:

$$\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p = argmin_{\beta_0,\beta_1,...,\beta_p} RSS = argmin_{\beta_0,\beta_1,...,\beta_p} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - ... - \hat{\beta}_p x_{ip})^2$$

Multiple Linear Regression

- Bias and Unbiasedness:
  The bias of an estimator means it might over or under estimate the truth averaging the corresponding estimates for a large number of data sets. An unbiased estimator does not systematically over- or under-estimate the true parameter. The property of unbiasedness holds for the least squares coefficient estimates.
- Standard error:
  The standard error of an estimator is standard deviation of the estimator, describing its variation due to repeated sampling. Denoted as $SE(\hat{\beta}_i)$.
- Confidence interval

Multiple Linear Regression

Hypothesis testing — t-test

- Hypothesis:

$$H_0 : \beta_i = 0, H_a : \beta_i \neq 0, i = 1, ..., p$$

- Test statistic:

$$t = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

- null distribution: Under H_0, the test statistic follows a t distribution with degrees of freedom $n - p - 1$.

Multiple Linear Regression

- rejection rule:
  - critical value approach:
    If the test statistic derived from the observed data t is larger than $t_{critical}$, then the null hypothesis should be rejected.
  - P-value approach:
    p-value is the probability of observing any value equal to $|t|$ or larger, under the null hypothesis, which is $\beta_1 = 0$

    $$p - value_{two-sided} = 2 * P(T > |t||H_0)$$

    If p-value is less or equal to the predefined significance level, then the null hypothesis should be rejected.

Multiple Linear Regression

R implementation:

- Simply use the *lm* function.
- How do we interpret the results?

Multiple Linear Regression

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	9.312	0.563	16.54	< 0.0001
radio	0.203	0.020	9.92	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Note: cor(TV, radio) = 0.0548, cor(TV, newspaper) = 0.0567, cor(radio, newspaper) = 0.3541.

Multiple Linear Regression

Assessing the model fit(REVIEW):

- Partitioning Variation: Break down difference between observation and grand mean into two parts:

$$Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i) \tag{1}$$

  - $Y_i - \bar{Y}$: Total deviation.
  - $\hat{Y}_i - \bar{Y}$: Deviation of fitted value around ground mean.
  - $Y_i - \hat{Y}_i$: Deviation around fitted value.

Multiple Linear Regression

Sums of Squares:
Square both sides, and the cross-terms in $(\hat{Y}_i - \bar{Y}) * (Y_i - \hat{Y}_i)$ will cancel.

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (\hat{Y}_i - \bar{Y})^2 + \sum_i (Y_i - \hat{Y}_i)^2 \tag{2}$$

- $\sum_i (Y_i - \bar{Y})^2$: Sum of squares total.(SSTO)
- $\sum_i (\hat{Y}_i - \bar{Y})^2$: Sum of squares regression.(SSR)
- $\sum_i (Y_i - \hat{Y}_i)^2$: Residual sum of squares/RSS (Sum of squares error/SSE)

Multiple Linear Regression

Assessing the model fit

- 1. RSE

$$RSE = \sqrt{\frac{RSS}{n-p-1}}$$

- 2. R-square

Quantity	**Value**
Residual standard error(RSE)	1.686
R^2	0.897

Disadvantage: R-square can only increase as predictors are added to the regression model. This increase is artificial when predictors are not actually improving the model's fit.

Multiple Linear Regression

Assessing the model fit

- 3. Adjusted R-square

$$\text{Adjusted } R^2 = 1 - \frac{RSS/df_{RSS}}{SST/df_{SST}}$$

Advantage: Adjusted R-squared will decrease as predictors are added if the increase in model fit does not make up for the loss of degrees of freedom. Likewise, it will increase as predictors are added if the increase in model fit is worthwhile.

Assessing the model fit

- 4. F-test
  - The F-test evaluates the null hypothesis that all regression coefficients are equal to zero versus the alternative that at least one is not.
  - F-test determines whether the proposed relationship between the response variable and the set of predictors is statistically reliable and can be useful when the research objective is either prediction or explanation.

Multiple Linear Regression

Hypothesis test — F test:
We can test all the coefficients together.

- Hypothesis:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$$
$$H_a\text{: Not all } \beta_j = 0, j = 1, ..., p$$

- Test statistic:

$$F = \frac{(SST - RSS)/p}{RSS/(n-p-1)}$$

- null distribution: Under H_0, the test statistic follows a F distribution with degrees of freedom p and $n - p - 1$.

Multiple Linear Regression

Advertising Example:

	Estimate	**Std. Error**	**t value**	**Pr($>|t|$)**
(Intercept)	2.939	0.312	9.422	$< 2e - 16^{***}$
TV	0.046	0.001	32.809	$< 2e - 16^{***}$
Radio	0.189	0.009	21.893	$< 2e - 16^{***}$
newspaper	-0.001	0.006	-0.177	0.86
F-statistic	570.3			
p-value	$< 2.2e - 16^{***}$			

Multiple Linear Regression

Hypothesis test — F test:
We can also test part of the coefficients.

- Hypothesis:

$$H_0 : \beta_{p-q+1} = \beta_2 = ... = \beta_p = 0$$
$$H_a\text{: Not all the above } \beta_j = 0, j = p - q + 1, ..., p$$

- Test statistic:

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)}$$

  where RSS_0 is residual sum of squares of the model that uses all the variables except those last q.

- null distribution: Under H_0, the test statistic follows a F distribution with degrees of freedom q and $n - p - 1$.

Multiple Linear Regression

Advertising Example:

	Estimate	**Std. Error**	**t value**	**Pr($>|t|$)**
(Intercept)	7.032	0.458	15.36	$< 2e - 16^{***}$
TV	0.048	0.003	17.67	$< 2e - 16^{***}$

Multiple Linear Regression

Model Selection:

- Criterions:
  Akaike information criterion(AIC), Bayesian information criterion(BIC), Mallow's C_p, adjusted R^2, etc.
- Procesures:
  - Forward Selection (RSS based)
  - Backward Selection (p-value based)
  - Stepwise Selection (combination of forward and backward)

Multiple Linear Regression

Model Selection:

- Forward Selection (RSS based):
  Begin with the null model-a model that contains an intercept but no predictors. We then fit p simple linear regressions and add to the null model the variable that results in the lowest RSS. Then add to that model the variable that results in the lowest RSS for the new two-variable model. This approach is continued until some stopping rule is satisfied.

Multiple Linear Regression

Model Selection:

- Backward Selection (p-value based):
  Start with all variables in the model, and remove the variable with the largest p-value-that is, the variable that is the least statistically significant. The new (p - 1)-variable model is fit, and the variable with the largest p-value is removed. This procedure continues until a stopping rule is reached.

Multiple Linear Regression

Model Selection:

- Stepwise Selection (combination of forward and backward):
  Start with no variables in the model, and as with forward selection, add the variable that provides the best fit. We continue to add variables one-by-one. If at any point the p-value for one of the variables in the model rises above a certain threshold, then remove that variable from the model. Continue to perform these forward and backward steps until all variables in the model have a sufficiently low p-value, and all variables outside the model would have a large p-value if added to the model.

Multiple Linear Regression

Qualitative variables:

- Also known as factors, discrete/categorical variables.
- Levels of qualitative predictors: possible values of the predictors

Dummy coding:

- 0-1 coding for two level predictors:

$$x_i = \begin{cases} 1, \text{if the } i\text{th person is female.} \\ 0, \text{if the } i\text{th person is male.} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i, \text{if the } i\text{th person is female.} \\ \beta_0 + \epsilon_i, \text{if the } i\text{th person is male.} \end{cases}$$

Multiple Linear Regression

Dummy coding:

- Sum to zero contrast:

$$x_i = \begin{cases} 1, \text{if the } i\text{th person is female.} \\ -1, \text{if the } i\text{th person is male.} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i, \text{if the } i\text{th person is female.} \\ \beta_0 - \beta_1 + \epsilon_i, \text{if the } i\text{th person is male.} \end{cases}$$

Multiple Linear Regression

Comparison of dummy coding:

- LSE of balance on gender using Credit data set(0-1 coding).

	Coefficient	Std. error	t-statistic	p-value
Intercept	509.80	33.13	15.389	< 0.0001
gender[Female]	19.73	46.05	0.429	0.6690

- LSE of balance on gender using Credit data set(sum to zero).

	Coefficient	Std. error	t-statistic	p-value
Intercept	519.665	23.026	22.569	< 0.0001
gender1	9.865	23.026	0.429	0.6690

Multiple Linear Regression

Three -level qualitative variables:
For example: ethnicity(Asian/Caucasian/African American.)
multiple 0-1 dummy coding:

$$x_{i1} = \begin{cases} 1, \text{if the } i\text{th person is Asian.} \\ 0, \text{if the } i\text{th person is not Asian.} \end{cases}$$

$$x_{i2} = \begin{cases} 1, \text{if the } i\text{th person is Caucasion.} \\ 0, \text{if the } i\text{th person is not Caucasion.} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i, \text{if the } i\text{th person is Asian.} \\ \beta_0 + \beta_2 + \epsilon_i, \text{if the } i\text{th person is Caucasion.} \\ \beta_0 + \epsilon_i, \text{if the } i\text{th person is African American.} \end{cases}$$

where β_0 is the baseline.

Multiple Linear Regression

synergy effect(interaction effect)

- Take the advertising data set as an example, given a fixed budget of \$100,000, spending half on radio and half on TV may increase sales more than allocating the entire amount to either TV or to radio.
- In statistical language, spending money on radio advertising actually increases the effectiveness of TV advertising, so that the slope term for TV should increase as radio increases.
- Let's first have a look of it.

Multiple Linear Regression

Linear regression with interactions:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Can you try to build this model in R?

Multiple Linear Regression

Linear regression with interactions:

	Estimate	**Std. Error**	**t value**	**Pr($>|t|$)**
(Intercept)	6.7500	0.2479	27.233	$< 2e - 16^{***}$
TV	0.0191	0.0015	12.699	$< 2e - 16^{***}$
Radio	0.0289	0.0089	3.241	0.0014
TVxRadio	0.0011	0.00005	20.727	$< 2e - 16^{***}$

Multiple Linear Regression

Estimated model:

$$E(y) = 6.75 + 0.0191 * x_1 + 0.0289 * x_2 + 0.0011 * x_1 x_2$$

- The R^2 for this model is 0.9678, while the one for model without interaction term is 0.8972.
- An increase in TV advertising of \$1,000 will result in increasing the sales by $(\beta_1 + \beta_3 * Radiounits) * 1000 = 19.1 + 1.1 * Radiounits$, an increase in Radio advertising of \$1,000 will result in increasing the sales by $(\beta_2 + \beta_3 * TVunits) * 1000 = 28.9 + 1.1 * TVunits$
- **hierarchical principle**: if we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.

Question?

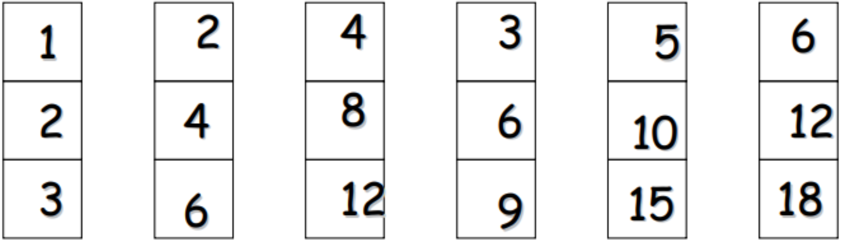
Principle Component Analysis

Principle Component Analysis

The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables.
[*Jolliffe*, *Pricipal Component Analysis*, 2^{nd} *edition*]

Principle Component Analysis

Toy Example:
Consider the following 3D points

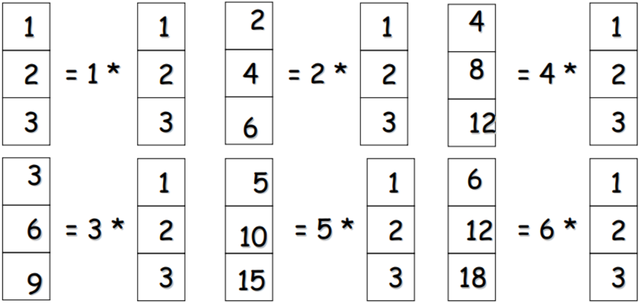


1	2	4	3	5	6
2	4	8	6	10	12
3	6	12	9	15	18

If each component is stored in a byte, we need 18 = 3 x 6 bytes.

Principle Component Analysis

Actually, they are all the same point, scaled by a factor.



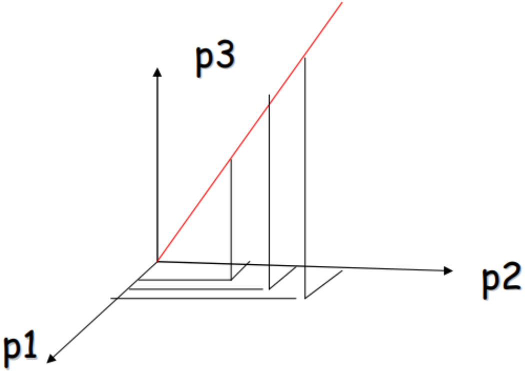
$$\begin{bmatrix}1\\2\\3\end{bmatrix} = 1 * \begin{bmatrix}1\\2\\3\end{bmatrix} \quad \begin{bmatrix}2\\4\\6\end{bmatrix} = 2 * \begin{bmatrix}1\\2\\3\end{bmatrix} \quad \begin{bmatrix}4\\8\\12\end{bmatrix} = 4 * \begin{bmatrix}1\\2\\3\end{bmatrix}$$

$$\begin{bmatrix}3\\6\\9\end{bmatrix} = 3 * \begin{bmatrix}1\\2\\3\end{bmatrix} \quad \begin{bmatrix}5\\10\\15\end{bmatrix} = 5 * \begin{bmatrix}1\\2\\3\end{bmatrix} \quad \begin{bmatrix}6\\12\\18\end{bmatrix} = 6 * \begin{bmatrix}1\\2\\3\end{bmatrix}$$

They can be stored using only 9 bytes (50% savings! Store one point (3 bytes) + the multiplying constants (6 bytes)

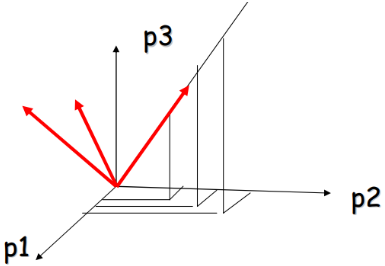
Principle Component Analysis

Geometrical Interpretation:



Principle Component Analysis

Geometrical Interpretation:
Consider a new coordinate system where one of the axes is along the direction of the line:



In this coordinate system, every point has only one only one non-zero coordinate: we only need to store the direction of the line (a 3 bytes image) and the nonzero coordinate for each of the points (6 bytes).

Principle Component Analysis

Introduction:

Suppose that we have a random vector X:

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$$

With population variance-covariance matrix:

$$Var(X) = \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{pmatrix}$$

Principle Component Analysis

- Consider the linear combinations:

$$\begin{aligned} C_1 &= e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p \\ C_2 &= e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p \\ &\vdots \\ C_p &= e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p \end{aligned}$$

  Which can be generalized by $C_i = e_i^T X, i = 1, ..., p$

- So, Y_i has a population variance:

$$Var(Y_i) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{ik} e_{il} \sigma_{kl} = e_i^T \Sigma e_i$$

- And Y_i and Y_j have population covariance:

$$Cov(Y_i, Y_j) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{ik} e_{jl} \sigma_{kl} = e_i^T \Sigma e_j$$

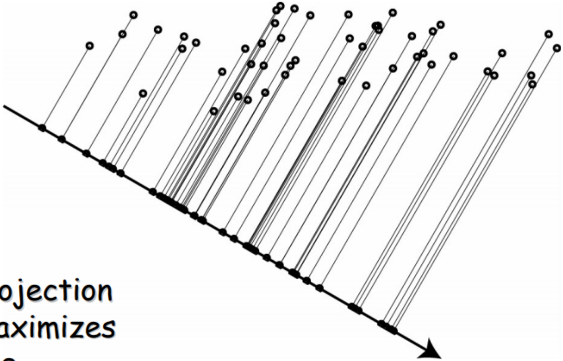
Principle Component Analysis

- The coefficients e_{ij} can be collected into a vector:

$$e_i = \begin{pmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{ip} \end{pmatrix}$$

Principle Component Analysis

Goal:



find projection
that maximizes
variance

Principle Component Analysis

Procedure:

- Find linear function of X, $e_1^T X$ with maximum variance.
- Next find another linear function of X, $e_2^T X$, uncorrelated with $e_1^T X$ maximum variance.
- Iterate until the pth linear function $e_p^T X$.

Principle Component Analysis

Details: First principal component

- The first principal component is the linear combination of x-variables that has maximum variance (among all linear combinations). It accounts for as much variation in the data as possible.
- Specifically we define coefficients for the first component in such a way that its variance is maximized, subject to the constraint that the sum of the squared coefficients is equal to one. This constraint is required so that a unique answer may be obtained.
- Mathematically,
  Maximize
  $$Var(C_1) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{1k} e_{1l} \sigma_{kl} = e_1^T \Sigma e_1$$
  subject to the constraint
  $$e_1^T e_1 = \sum_{i=1}^{p} e_{1j}^2 = 1$$

Principle Component Analysis

Details: Second principal component

- The second principal component is the linear combination of x-variables that accounts for as much of the remaining variation as possible, with the constraint that the correlation between the first and second component is 0.
- Mathematically,
  Maximize
  $$Var(C_2) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{2k} e_{2l} \sigma_{kl} = e_2^T \Sigma e_2$$
  subject to the constraint
  $$e_2^T e_2 = \sum_{i=1}^{p} e_{2j}^2 = 1$$
  Along with the additional constraint that these two components are uncorrelated.
  $$Cov(C_1, C_2) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{1k} e_{2l} \sigma_{kl} = e_1^T \Sigma e_2 = 0$$

Principle Component Analysis

Details: The ith principal component

- All subsequent principal components are linear combinations that account for as much of the remaining variation as possible and they are not correlated with the previous principal components.
- Mathematically,
  Maximize

$$Var(C_i) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{ik} e_{il} \sigma_{kl} = e_i^T \Sigma e_i$$

subject to the constraint

$$e_i^T e_i = \sum_{i=1}^{p} e_{ij}^2 = 1$$

Along with the additional constraint.

$$Cov(C_1, C_i) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{1k} e_{il} \sigma_{kl} = e_1^T \Sigma e_i = 0$$

$$\vdots$$

$$Cov(C_{i-1}, C_i) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{(i-1)k} e_{il} \sigma_{kl} = e_{(i-1)}^T \Sigma e_i = 0$$

Principle Component Analysis

- Let λ_1 through λ_p denote the eigenvalues of the variance-covariance matrix Σ. These are ordered so that λ_1 has the largest eigenvalue and λ_p is the smallest.

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$$

- Let the vectors e_1 through e_p

$$e_1, e_2, \cdots, e_p$$

denote the corresponding eigenvectors. It turns out that the elements for these eigenvectors are the coefficients of our principal components.

Principle Component Analysis

Methodology:
Constrained maximization - Lagrange multipliers
We maximize the function $e_i^T \Sigma e_i - \lambda(e_i^T e_i - 1)$ with respect to e_i by differentiating with respect to e_i.

$$\text{This will result in } \Sigma e_i = \lambda_i e_i$$

This can be recognizable as an eigenvector equation where e_i is an eigenvector of Σ and λ_i is the associated eigenvalue.

Principle Component Analysis

Methodology:

$$Var(C_i) = e_i^T \Sigma e_i = e_i^T \lambda_i e_i = \lambda_i e_i^T e_i = \lambda_i$$

So the varaince for C_i is just the eigenvalue λ_i, then we should choose λ_i to be as big as possible.

Then the solution to:

$$\Sigma e_1 = \lambda_1 e_1$$

is the 1st principal component of X.

Principle Component Analysis

Methodology:

- The second PC, $e_2^T X$ maximizes $e_2^T \Sigma e_2$ subject to being uncorrelated with $e_1^T X$.
- The uncorrelation constraint can be expressed using any of these equations:

$$Cov(e_1^T X, e_2^T X) = e_1^T \Sigma e_2 = e_2^T \Sigma e_1 = e_2^T \lambda_1 e_1 = \lambda_1 e_2^T e_1 = 0$$

Principle Component Analysis

Methodology:
By combining the uncorrelation constraint to the previous constraint, then we just need to maximize the following function to get the e_2:

$$e_2^T \Sigma e_2 - \lambda_2(e_2^T e_2 - 1) - \phi e_2^T e_1$$

By differentiating the function w.r.t the e_2, and set the derivative equals to 0, then we have ϕ must equal to 0, then what is left is:

$$\Sigma e_2 = \lambda_2 e_2$$

So similar to the first PC, we choose e_2 to be the eigenvector associated with the second largest eigenvalue to get the second PC of X, namely $e_2^T X$.

Principle Component Analysis

Methodology:
This process can be repeated for $k = 1, ..., p$ yielding up to p different eigenvectors of Σ along with the corresponding eigenvalues $\lambda_1, ..., \lambda_p$.
Furthermore, the variance of each of the PC's are given by:

$$Var(e_k^T X) = \lambda_k, k = 1, ..., p$$

Principle Component Analysis

Dimension reduction

Now, think of the PC as projections of the of X, since the projections are uncorrelated, the percentage of variance accounted for by retaining the first q PC's is given by:

$$\frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k} * 100$$

Principle Component Analysis

Practical Principle Component Analysis Procedure

- Sample covariance matrix:
  An unbiased estimator for the covariance matrix of X:
  $$S = \frac{1}{n-1} X^T X$$
  where X is $n * p$ data matrix, with (i, j)th element being $x_{ij} - \bar{x}_j$ (centered matrix).
- Construct the matrix A by combining the p eigenvectors of S (or eigenvectors of $X^T X$), then we can define a matrix of PC scores:
  $$Z = XA$$
- For dimension reduciton:
  Selecting the q eigenvectors corresponding to the q largest eigenvalues of S when forming A, then Z is an "optimal" q-dimensional projection of X.

Principle Component Analysis

- How to find the eigenvectors & eigenvalues — singular value decomposition(SVD):

  $$S = A^T \Lambda A$$

  where A is a matrix consisting of the eigenvectors of S and Λ is a diagonal matrix whose diagonal elements are the eigenvalues corresponding to each eigenvector.
- Based on the SVD, how to do dimension reduciton?

Principle Component Analysis

Sample:

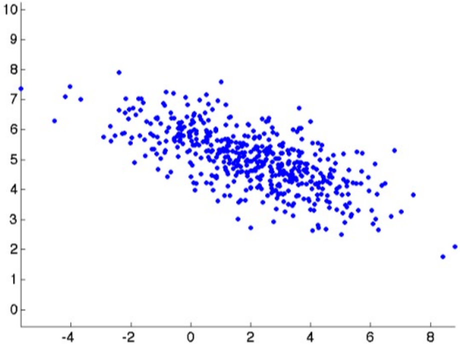


Figure: Gaussian sample

Principle Component Analysis

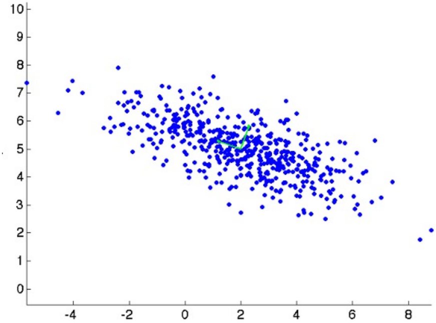


Figure: Gaussian sample with eigenvectors of sample covariance matrix

Principle Component Analysis

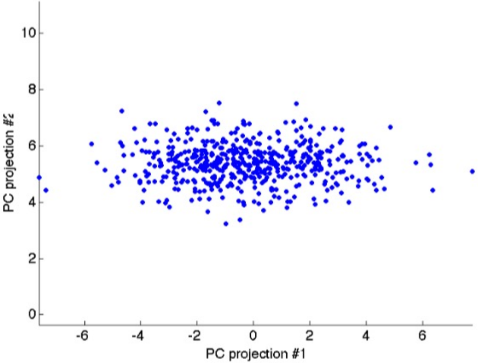


Figure: Projected sample

Principle Component Analysis

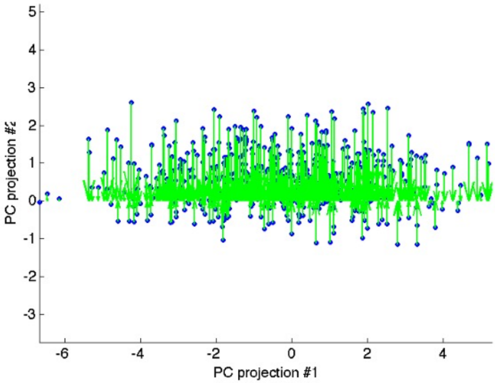


Figure: PC dimensionality reduction

Principle Component Analysis

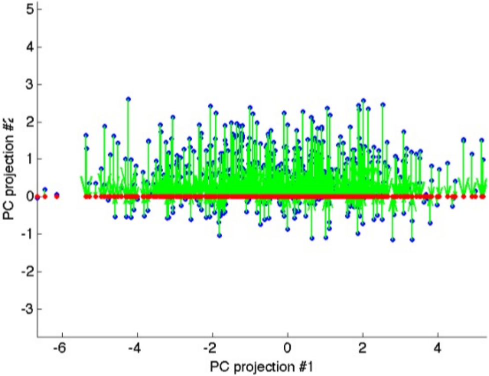


Figure: PC dimensionality reduction

Principle Component Analysis

PCA in linear regression:
Advantages:

- Identification and elimination of multicolinearities in the data.
- Reduction in the dimension of the input space leading to fewer parameters and easier regression.
- The variance of the regression coefficient estimator is minimized by the PCA choice of basis.

Principle Component Analysis

A simulation example:

- $X \sim N(\begin{pmatrix} 2 \\ 5 \end{pmatrix}, \begin{pmatrix} 4.5 & -1.5 \\ -1.5 & 1.0 \end{pmatrix})$
- Model with no noise(no colinearity):
  $Y = X \begin{pmatrix} 5 & -1 \\ 2 & \end{pmatrix}$,
  which means $Y = 5 + -1 * X_1 + 2 * X_2$
- Model with noise(colinearity):
  Add another predictor $X_3 = 0.8 * X_1 + 0.5 * X_2$

Principle Component Analysis

Linear Relationship with No Colinearity:

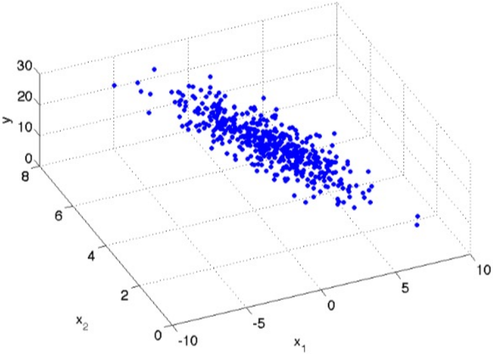


Figure: Noiseless Linear Relationship

Principle Component Analysis

Linear Relationship with No Colinearity:

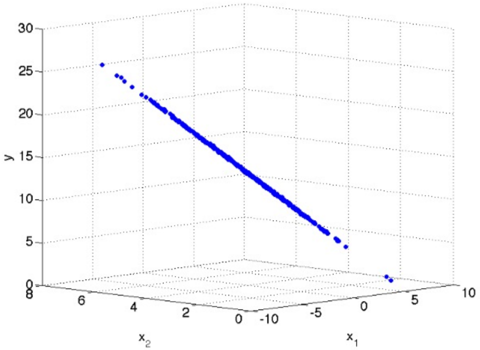


Figure: Noiseless Planar Relationship

Principle Component Analysis

- For the colinear data, it is not possible to plot it.
- When PCA is applied to the design matrix of rank q less than p (the number of positive eigenvalues discovered) is equal to q (the true rank of the design matrix).
- In this example, the rank of design matrix is rank 2, so the resulting projection will be in two dimensions.

Principle Component Analysis

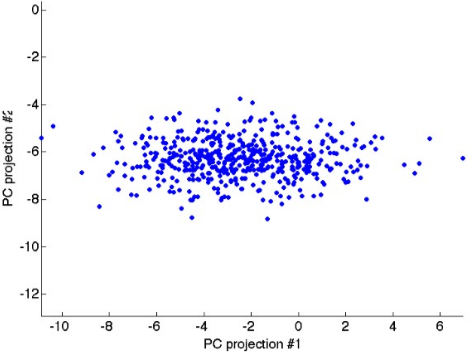


Figure: Projection of multi-colinear data onto first two PC's

Principle Component Analysis

If we take the standard regression model:

$$Y = X * \beta + \epsilon$$

The PCA rotation of X:

$$Z = X * A$$

Rewrite the regression model in terms of the PC's:

$$Y = Z * \gamma + \epsilon$$

Consider the reduced model:

$$Y = Z_q * \gamma_q + \epsilon_q$$

Principle Component Analysis

A is orthogonal, then rewrite:

$$X\beta = XAA^T\beta = Z\gamma$$

where $\gamma = A^T\beta$.
Using least squares (or ML) to learn $\hat{\beta} = A\hat{\gamma}$ is equivalent to learning $\hat{\beta}$ directly.
So the least square estimate $\hat{\gamma}$ is:

$$\hat{\gamma} = (Z^TZ)^{-1}Z^TY$$

And,

$$\hat{\beta} = A(Z^TZ)^{-1}Z^TY$$

Principle Component Analysis

Disadvantage:

- PCA assumes that the input data is real and continuous.
- PCA assumes approximate normality of the input space distribution.

Principle Component Analysis

Simple implementation:
We will use the *Breast Cancer Wisconsin* dataset. There are 32 variables in total: ID, diagnosis and ten distinct (30) features.
The mean, standard error, and “worst” or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, Variable 3 is Mean Radius, Variable 13 is Radius SE, Variable 23 is Worst Radius.

Principle Component Analysis

For PCA, simply use the *prcomp*() function!

Question?

Discussion

Discussion

1. Dimension reduction: Factor analysis

2. High-dimensional data analysis:

   - Multivariate Analysis of Variance (MANOVA)
   - Repeated Measures Analysis
   - Discriminant Analysis
   - Cluster Analysis

Question?

- For hands on help with your analyses, stop by our drop in hours or sign up for a consultation.
- Welcome to the workshops in the Fall quarter!
- If you have any workshop requests, now is the time to ask! We will be setting our fall schedule soon.
- For more details, visit our website:
  GradQuant.ucr.edu

Thank You
Welcome to GradQuant