STATS 415: Dimension reduction for linear regression

Prof. Liza Levina

Department of Statistics, University of Michigan

Improving on Ordinary Least Squares

- Subset selection (of variables X)
- Shrinkage (of coefficients $\hat{\beta}$)
- Dimension reduction (of variables X) if p is large

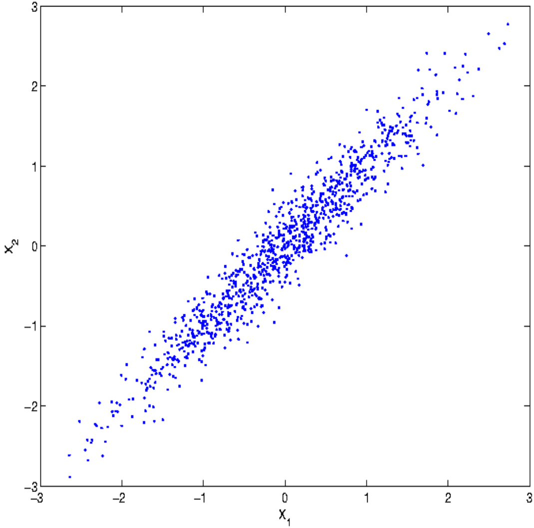
Dimension reduction

- Projecting all p predictors into a k-dimensional space where $k < p$, and then fitting a smaller linear regression model (with k predictors)
- E.g. principle components regression, partial least squares
- Advantage: a much smaller model, faster to fit, coefficients are stable
- Disadvantages: the relationship between y and X is not taken into account when performing dimension reduction; original variables are no longer in the model, therefore interpretation is lost

Principal component analysis

- The main objective: reduce dimensionality of the data set.
- Replaces the original p variables with $k < p$ linear combinations of the original variables that are a "good representation" of the data (a linear dimension reduction method)
- Belongs to the class of projection methods
- Useful for
  - visualization (project to 2-d or 3-d)
  - as a pre-processing step for other methods that do not deal well with an excessive number of variables (principle component regression (PCR), classification based on principle components)

A Toy Example:



Question: What is a good 1-dim projection of the data?

Some Possibilities

- Could use 1 (2,3,...) of the variables (e.g. X_1 in the toy example). But what if there are many thousands of variables?
- Better idea: use a linear combination of the variables; i.e. a weighted average of the variables. In the toy example,

$$Z_1 = w_1 X_1 + w_2 X_2.$$

- What is a good choice for the weights w_i? Need a criterion.

The Criterion for Principal Components

- PCA finds the direction vector w that maximizes

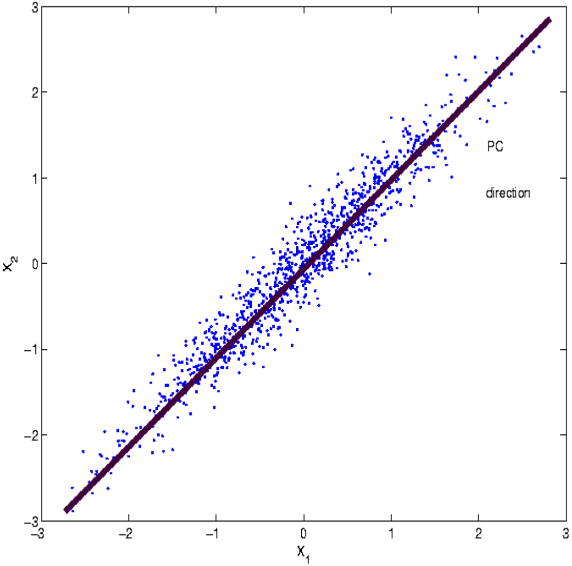
$$\max_{w:\|w\|=1} \mathrm{Var}\left(\sum_i w_i X_i\right)$$

- Can rewrite this in matrix form:

$$\max_{w:\|w\|=1} w^T \Sigma w$$

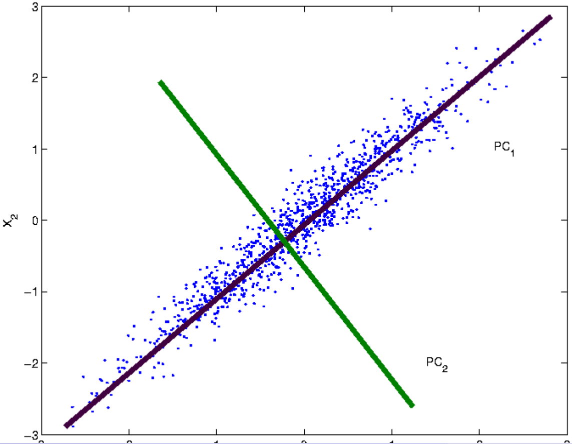
where Σ is the covariance matrix of the data X.
- The “interesting” direction in the data according to the PCA criterion is the one that captures the most variance in the data.

Toy Example: 1-dim PCA solution



Toy Example: 2-dim PCA solution

- What if we wanted a second linear combination, i.e. $Z_2 = v_1 X_1 + v_2 X_2 = v^T X$?
- Require subsequent linear combinations to be orthogonal to previous ones.



Mathematical Formulation of PCA

- Assume that the variables have been centered
- The problem: Find p new variables $Z_1, Z_2, \cdots, Z_p$, such that $Z_i = \sum_{j=1}^{p} w_{ij} X_j$ and the weights w maximize

$$w_i^T \Sigma w_i \text{ subject to } w_i^T w_i = 1, w_i^T w_j = 0.$$

- In matrix form: find $Y = XW$ where W solves

$$\max_{W: W^T W = I} W^T \Sigma W.$$

- Solution (proof omitted): the columns of W are given by the eigenvectors of Σ.

Properties of PCs

- New variables Z_j have mean 0
- Var$(Z_j) = \lambda_j$, where λ_j is the jth largest eigenvalue of Σ.
- Cor$(Z_j, Z_j') = 0$ for all $j \neq j'$: PCs are uncorrelated.

PCA Terminology

- The vectors w_i are called PC directions
- Vectors $Z_i = Xw_i$ are projections of the data onto the PC directions
- Components of Xw_i are called scores
- The coordinates w_{ij} are called loadings; sometimes loadings are defined as $\sqrt{\lambda_j} w_{ij}$.

Covariance vs Correlation

- Should we standardize the variables first (mean 0, sd 1)?
- This is equivalent to applying PCA to the correlation matrix instead of the covariance matrix
- The PCs from covariance and from correlation are not the same
- Reason to standardize: makes the analysis independent of units; generally recommended.
- Reason to not standardize: there is information in the variance, particularly if all variables are measured on the same scale

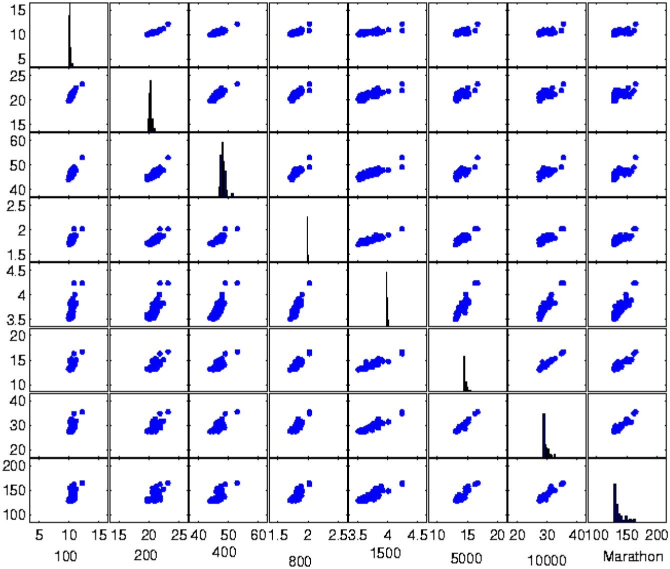
Example: Athletic Performance Data

- Records for 55 countries in the following men's track events: 100, 200, 400, 800, 1500, 5000, 10000 meters and the marathon
- The data are in seconds for the first three events and in minutes for the rest.

Country codes

AG=Argentina AL=Australia AR=Austria BG=Belgium BM=Bermuda BZ=Brazil BU=Burma CD=Canada CL=Chile CH=China CO=Colombia CI=Cook.Islands CR=Costa.Rica CS=Czechoslovakia DR=Denmark DM=Dominican.Rep FL=Finland FR=France GD=German.Dem.Rep GF=German.Fed.Rep GB=Great.Britain.NI GC=Greece GT=Guatemala HU=Hungary IN=India IO=Indonesia IL=Ireland IS=Israel IT=Italy JA=Japan KY=Kenya KS=Korea KN=Korean.DP.Rep LX=Luxemburg MA=Malaysia MR=Mauritius MX=Mexico NL=Netherlands NZ=New.Zealand NW=Norway PN=Papua.New.Guinea PH=Philippines PL=Poland PR=Portugal RO=Romania SI=Singapore SP=Spain SW=Sweden SZ=Switzerland TP=Taipei TH=Thailand TU=Turkey US=USA UR=USSR WS=Western.Samoa

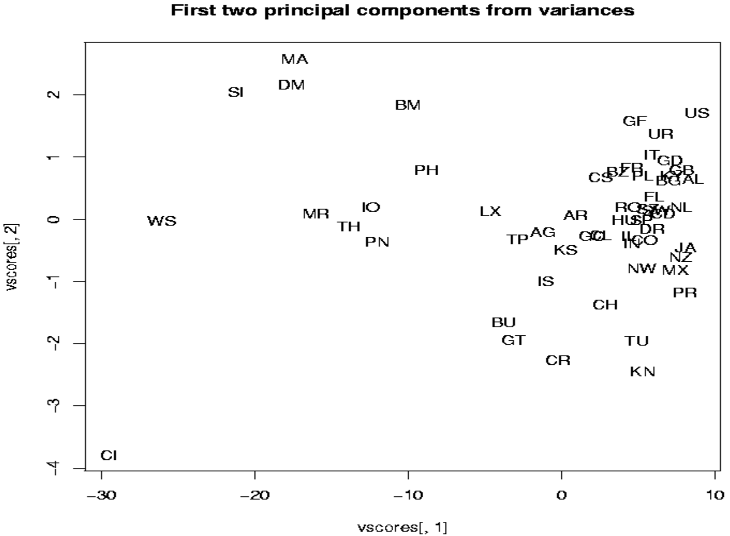
Pairwise scatterplots and histograms



Loadings from the covariance matrix (variables not standardized)

Variable	Comp1	Comp2
X100	-0.020	-0.211
X200	-0.042	-0.359
X400	-0.111	-0.828
X800	-0.005	-0.023
X1500	-0.014	-0.045
X5000	-0.079	-0.130
X10000	-0.181	-0.299
Marathon	-0.973	0.181

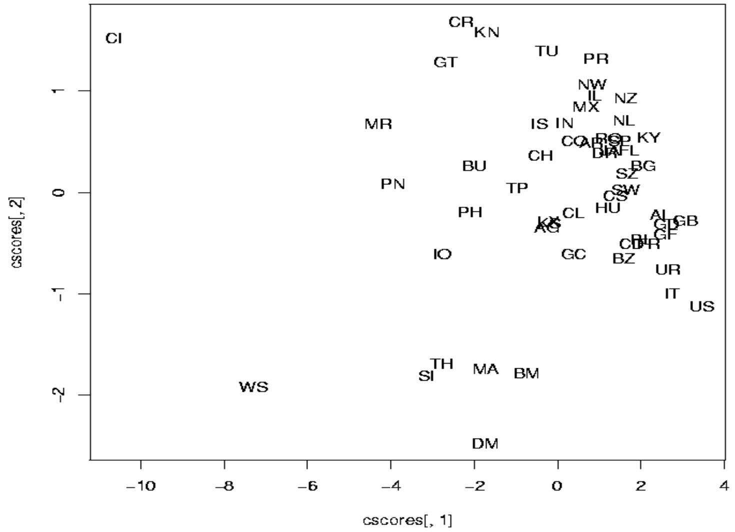
Projection onto the first two PCs



Loadings from the correlation matrix

Variable	Comp1	Comp2
X100	-0.318	0.567
X200	-0.337	0.462
X400	-0.356	0.248
X800	-0.369	0.012
X1500	-0.373	-0.140
X5000	-0.364	-0.312
X10000	-0.367	-0.307
Marathon	-0.342	-0.439

First two principal components from correlations



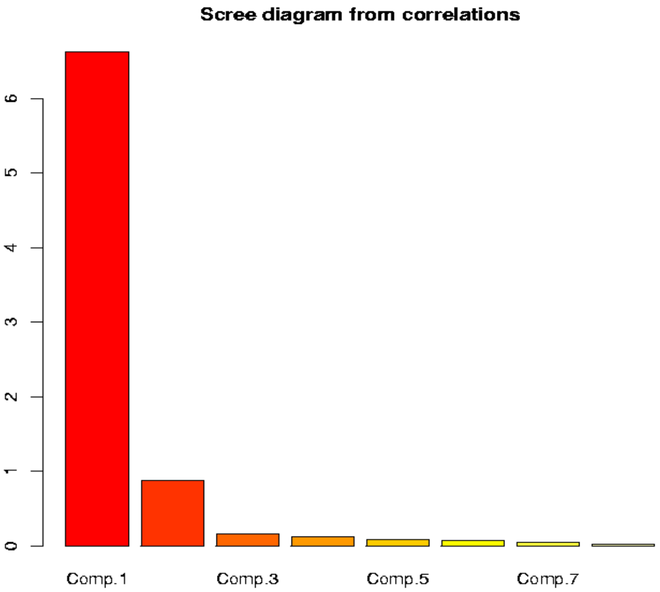
How many PCs should we use?

- For visualization, can only use 2 or 3
- For general dimension reduction, want to keep enough to represent the data "well"
- Scree plot: plot λ_i or $\sqrt{\lambda_i}$ against i and look for an "elbow"
- Percentage of variance explained: component i "explains" $\frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}$, so pick the first k such that

$$\frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j} \geq 1 - \alpha$$

  for some pre-specified small α (e.g. 0.1)
- Some hypothesis tests have been proposed, but no universal rule

Scree plot for athletic data



Some other issues

- PCs with equal variance: if k eigenvalues coincide, their eigenspace is a unique k-dimensional subspace, but within that subspace PC directions cannot be distinguished
- Outliers: PCA is not a robust method, some robust versions exist
- Subsets of variables: sometimes want to express a PC in terms of just a few original variables (for ease of interpretation). Sparse variants of PCA are available (shrink many loadings to exactly 0).
- Singular Σ: then only r PCs are defined, where $r = \text{rank}(\Sigma)$. Always the case when $p > n$, since the sample covariance matrix has rank $\min(p, n-1)$.

Principal Components Regression (PCR)

- PCR replaces the regression model

$$y = \beta_0 + \beta_1 x_1 + \ldots \beta_p x_p + \varepsilon$$

  with

$$y = \beta_0 + \beta_1 z_1 + \cdots + \beta_k z_k + \varepsilon$$

- Note: PCs are centered, so $\hat{\beta}_0 = \bar{y}$
- Potentially, $k \ll p$
- New predictors are orthogonal
- Interpretation and inference are no longer in terms of the original variables

Partial Least Squares

- PCR ignores y when building z's.
- Partial least squares (PLS) chooses z's that are best at predicting y.
- PLS does not solve a well-defined modeling problem; it's just an algorithm.
- Also need to select number of components
- No interpretation

Partial Least Squares

Algorithm:

1. Center y, center and standardize each x_j
2. Regress y on each x_j separately to get α_j
3. Construct $z_1 = \sum \alpha_j x_j$, which is the first PLS component
4. Regress y on z_1 to get $\hat{\beta}_1$
5. Regress each x_j on z_1 and replace it with the residual ("orthogonolize" x_j to the first component)
6. Return to step 2 and continue until the final model is fit:

$$\hat{y} = \bar{y} + \hat{\beta}_1 z_1 + \cdots \hat{\beta}_k z_k$$

Summary

- PCA is a useful and popular dimension reductino method
- Easy to use in regression, via PCR and PLS
- Need to choose K
- PCR is not interpretable in the original variables