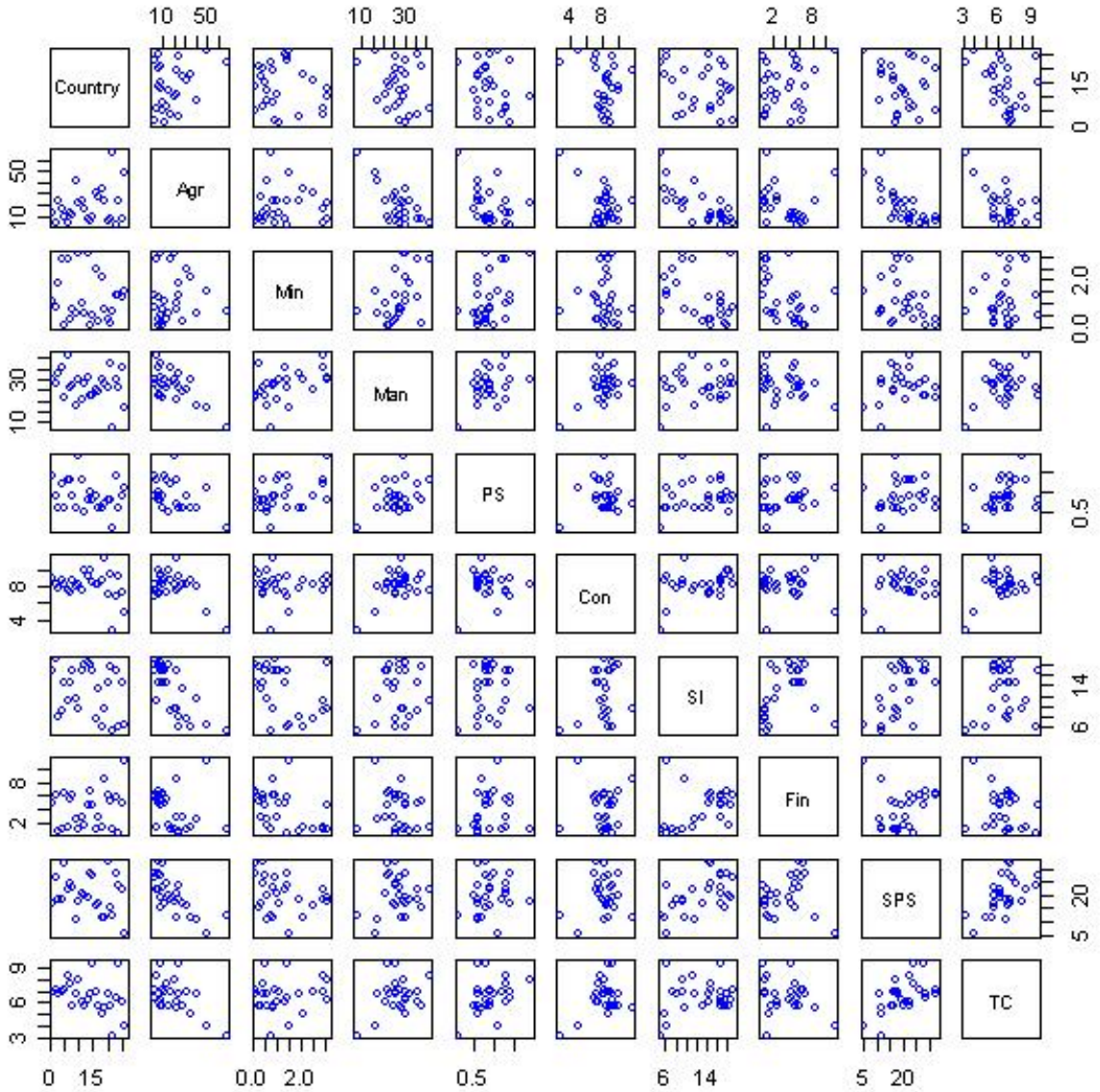
# Lecture 10: PCA, Model Selection

Nancy R. Zhang

Statistics 203, Stanford University

February 11, 2010

# European Jobs Data

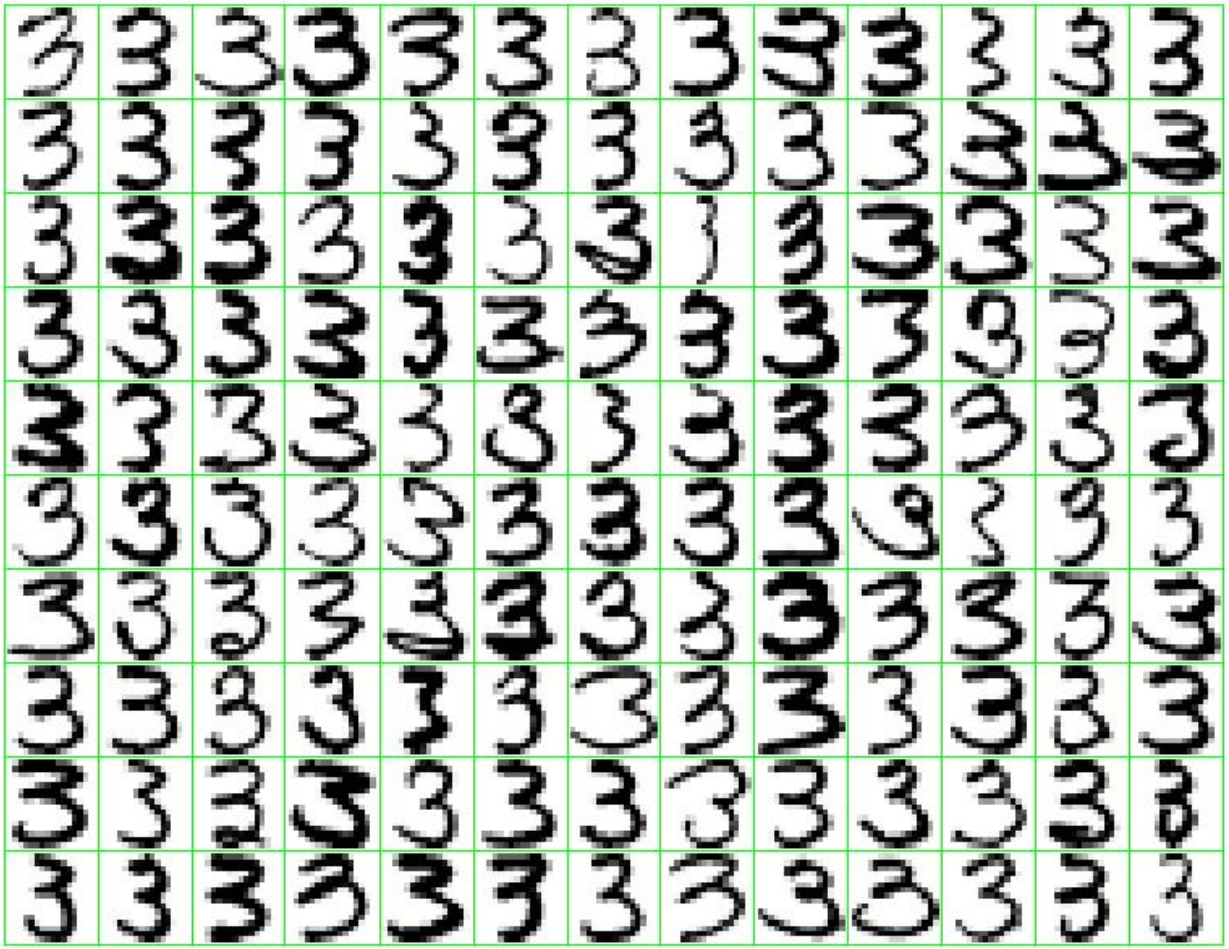



Percentage of jobs for 26 European countries in following industries:

1. Country: Name of country
2. Agr: agriculture
3. Min: mining
4. Man: manufacturing
5. PS: power supply industries
6. Con: construction
7. SI: service industries
8. Fin: finance
9. SPS: social and personal services
10. TC: transport and communications

Data collected in 1979.

# Handwritten Digits



Handwritten digits, automatically scanned from envelopes by the U.S. Postal Service in 16 x 16 grayscale images (Le Cun et al., 1990) Here is a sampling of 130 3's. A total of 638 3's analyzed.

# Principal Components

Principal components is a useful way to explore high dimensional data.

- Does not distinguish between "predictor" and "response".
- Look for "meaningful" linear projections of the data.

What do we mean by "meaningful"?

- "best fitting hyperplane":

$$\min_{\mu,\{\beta_i\},V_k} \sum_{i=1}^{N} \|x_i - \mu - V_k\beta_i\|^2.$$

- Direction of maximum variation (more on next slide).

# Direction of maximum variation

Your data is $n \times p$ matrix $X$, containing $n$ data points of dimension $p$. (For example, European jobs data has $p = 9$, $n = 26$. $X$ must first be *centered* to have columns of mean 0. Find $v \in \Re^p$, such that:

$$\|v\| = 1,$$

and

$$Var(Xv) \quad \text{is maximized.}$$

The vector that satisfies the above is called the first principal component. Since

$$\begin{aligned} Var(Xv) &= v'(X - \bar{X})'(X - \bar{X})v \\ &= v'\Sigma_X v, \end{aligned}$$

where $\Sigma_X$ is the sample covariance matrix of $X$, then the first principal component is simply the eigenvector of $\Sigma_X$ corresponding to its largest eigenvalue.

# The first k principal components

You may want to find the $k$ directions of maximum variation.
Let

$$v_1 = \text{argmax}_{\|v\|=1} Var(Xv)$$

be the first principal component. The second principal component is defined as:

$$v_2 = \text{argmax}_{\|v\|=1, v'v_1=0} Var(Xv),$$

that is, the direction of maximal variation that is orthogonal to $v_1$.

The $3, 4, \ldots, k$ principal components can be defined recursively in this way. They correspond to the $k$ eigenvectors corresponding to the $k$ largest eigenvalues of $\Sigma_X$.

# Practical Implementation

In R, and most other software, principal components are computed by the Singular Value Decomposition (SVD) of $X$, which gives:

$$X = UDV',$$

where

$$\begin{aligned} U: &\quad n \times p \quad \text{orthonormal columns} \\ D: &\quad p \times p \quad \text{diagonal,} \\ V: &\quad p \times p \quad \text{orthonormal.} \end{aligned}$$

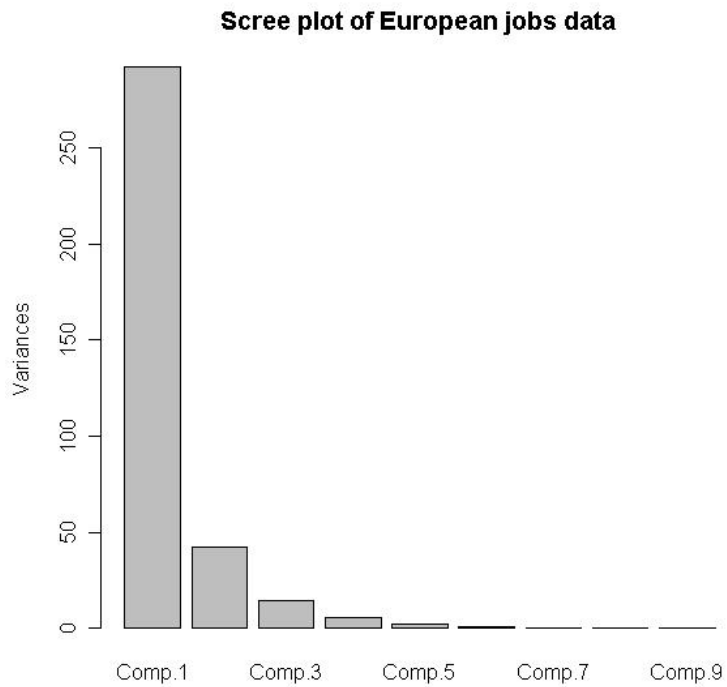
The columns of $V$ are the principal component vectors, also called "loadings". The columns of $U$ are sometimes called "scores". The magnitude of projection of $X$ on $V$ are in the columns of $UD$. The diagonal elements of $D$ are the variances along the principal component vectors.

Every $n \times p$ matrix $X$ can be decomposed in this way. What is the maximum number of principal components?

# Interpretation of Principal components

1. If the variances of the principal components drop off quickly, then $X$ is highly colinear.
2. To reduce the dimensionality of the data, we keep only the principal components with highest $d_i$.
3. The principal vectors are derived projections of the data, and may not have a specific meaning.

The scree plot, which shows $d_i$ versus $i$, is useful:



# Some measures of collinearity

Important: must first scale $X$ so that all columns have variance 1. Why?

1. Condition number

$$\kappa = \sqrt{\frac{d_1}{d_p}},$$

   large $\kappa$ means strong collinearity.

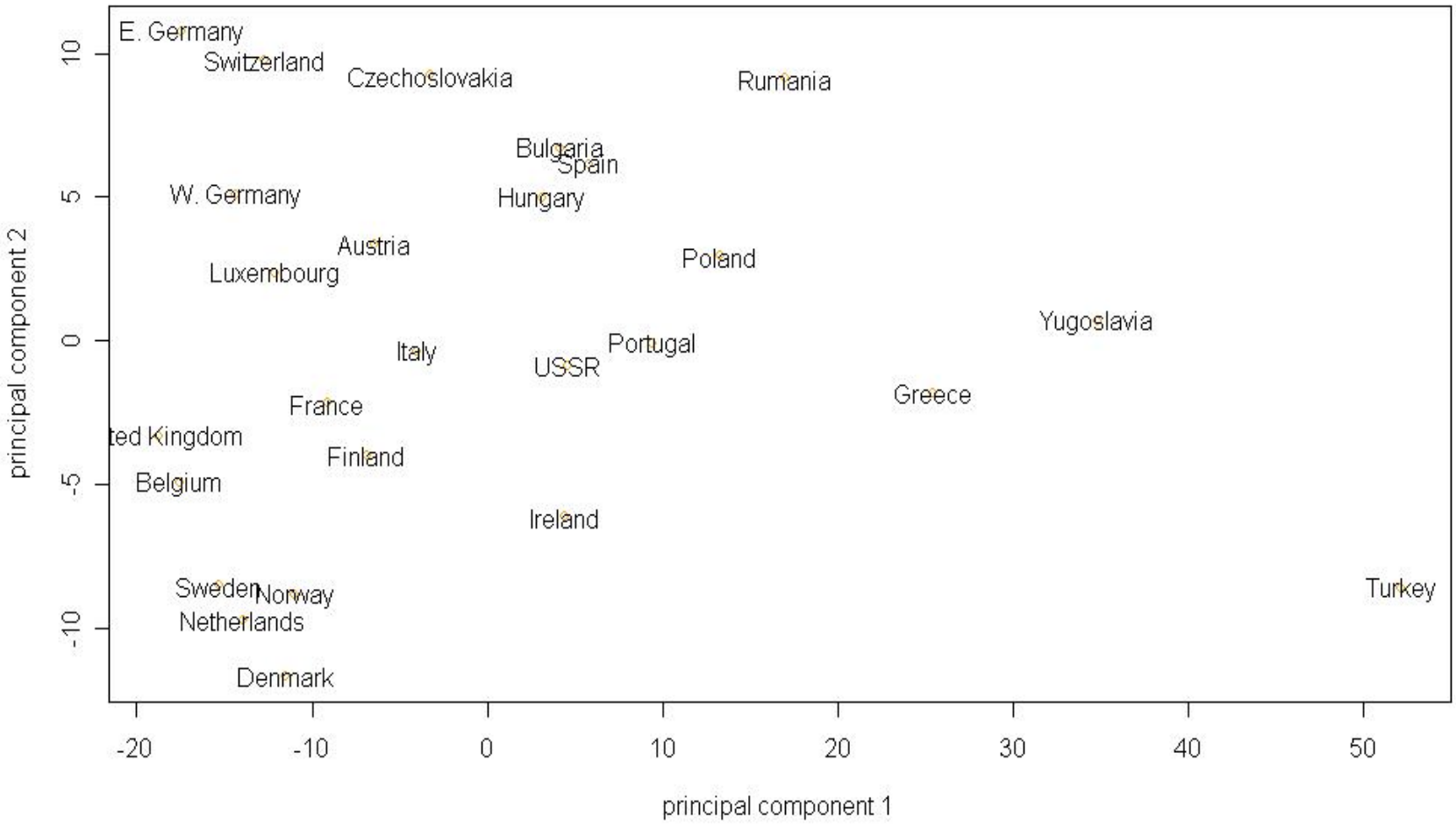
2. Less used: $\sum_{j=1}^{p} \frac{1}{d_j}$, large sum means strong collinearity.

Another measure (more details in book):

$$VIF_j = \frac{1}{1 - R_j^2},$$

where $R_j$ is the multiple correlation coefficient from regressing $X_j$ on all other variables.

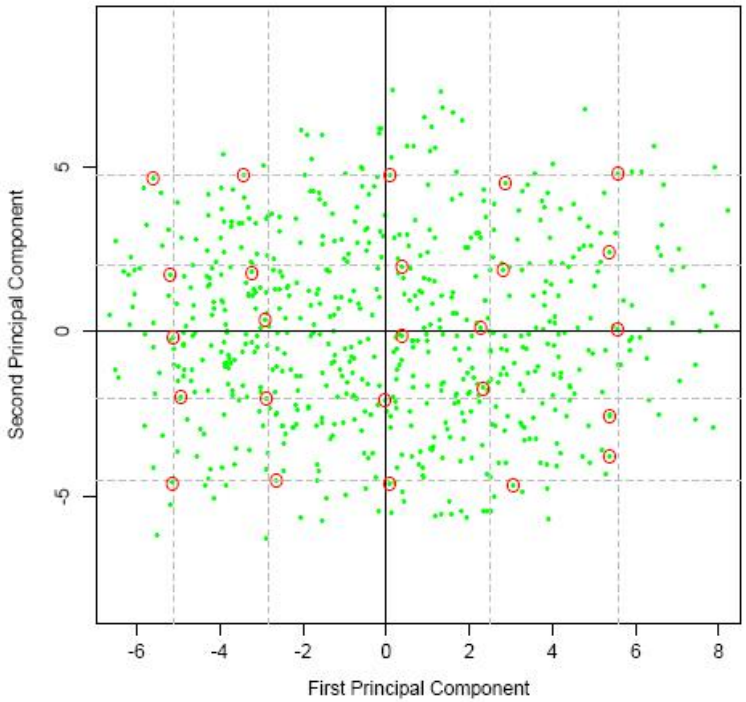
# European Jobs Data

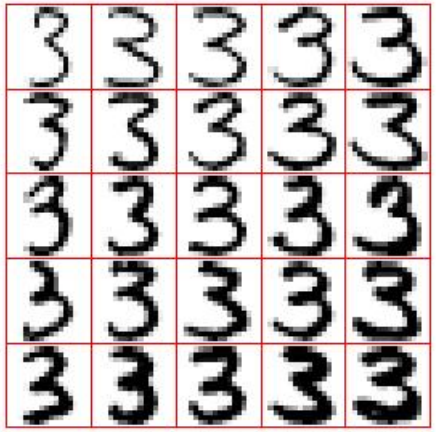



PC1: $0.89 \times \text{AGR} - 0.27 \times \text{Man} - 0.192 \times \text{SI} - 0.298 \times \text{SPS}$

PC2: $0.77 \times \text{Man} - 0.234 \times \text{SI} - 0.13 \times \text{FIN} - 0.567 \times \text{SPS}$

# Handwritten Digits



$n = 638$, $p = 256$ for 638 gray scale images, each of size $16 \times 16$.
Here are the first 2 principal components.

# Model Selection Outline

1. How to compare two non-nested models?
   1. Bias - variance trade-off
   2. $C_p$
   3. AIC
   4. Cross-validation
   5. BIC
2. How to search the space of possible models?
   1. Step-wise search
   2. Best subsets
   3. LASSO
   4. Bayesian methods

# Training versus Test Performance

Given $n \times p$ matrix $X$ and response $Y$, we obtain a fitted model:

$$\hat{Y} = \hat{f}_{X,Y}(X).$$

For example, in the linear regression models we have been studying,

$$\hat{f}_{X,Y}(X) = X\hat{\beta}_{X,Y} = X(X'X)^{-1}X'Y.$$

The predicted values for $Y$ are based on regression parameters that were fit using $X$, $Y$.

How would the model perform on unseen data?

$$EPE \equiv E[Y_{\text{new}} - \hat{Y}_{\text{new}}]^2 \quad ?$$

The expectation is taken over everything that is random:

$$X, Y, X_{\text{new}}, Y_{\text{new}}.$$

Good training performance does not imply good test performance.

# Bias-Variance Trade-off

$X$, $Y$ used to fit the model are called "training data", and "unseen" data used to estimate prediction error are called '"test data".

$$\text{Truth:} \quad y = f(x) + \epsilon, \quad Var(\epsilon) = \sigma^2.$$

$$\text{Estimate } f(\cdot) \text{ using } \hat{f}_{X,Y}.$$

$$\begin{aligned}
EPE &\equiv E[Y_{\text{new}} - \hat{Y}_{\text{new}}]^2 \\
&= E[Y_{\text{new}} - \hat{f}_{X,Y}(X_{\text{new}})]^2 \\
&= E[(Y_{\text{new}} - f(X_{\text{new}})) + (f(X_{\text{new}}) - E\hat{f}_{X,Y}(X_{\text{new}})) + (E\hat{f}_{X,Y}(X_{\text{new}}) - \hat{f}_{X,Y}(X_{\text{new}}))]^2 \\
&= E[Y_{\text{new}} - f(X_{\text{new}})]^2 + E[f(X_{\text{new}}) - E\hat{f}_{X,Y}(X_{\text{new}})]^2 \\
&\quad + E[E\hat{f}_{X,Y}(X_{\text{new}}) - \hat{f}_{X,Y}(X_{\text{new}})]^2 \\
&= \sigma^2 + (\text{Model bias})^2 + \text{Model variance}
\end{aligned}$$

As model complexity increases, bias *decreases* while variance *increases*. How to achieve a balance?

# Bias-Variance Trade-off

If you care more about $\hat{\beta}$:

$$\begin{aligned} MSE \equiv E[\beta - \hat{\beta}]^2 &= [E(\hat{\beta}) - \beta]^2 + E[\hat{\beta} - E(\hat{\beta})]^2 \\ &= Bias(\hat{\beta})^2 + Var(\hat{\beta}). \end{aligned}$$

For linear regression, assuming that you've got the correct model, the least squares estimates had 0 bias:

$$E[\hat{\beta}] = \beta,$$

so

$$MSE = Var(\hat{\beta}).$$

In a multiple regression, for any $\hat{\beta}_i$:

$$MSE(\hat{\beta}_i) = Var(\hat{\beta}_i) = \frac{\sigma^2}{\|r_{i,\ i}\|^2},$$

where $r_{i,\ i}$ are the residuals of regression $X_i$ on all other predictors.

# Estimating the prediction error

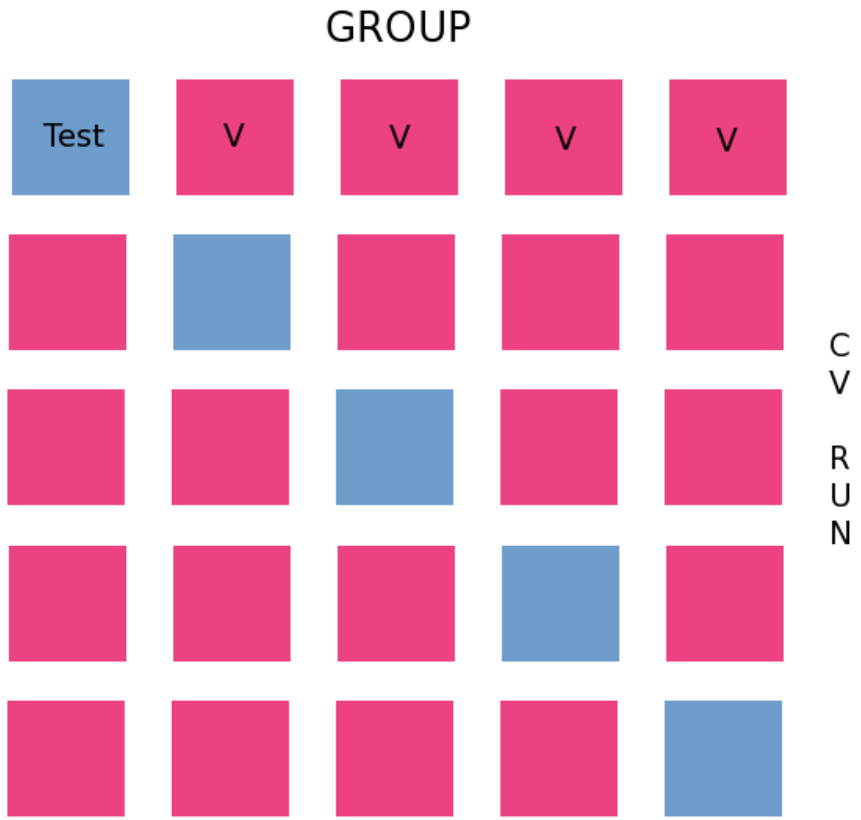
1. Asymptotic approximations
   - Mallows $C_p$:
     $$C_p = SSE + 2p\hat{\sigma}^2,$$
     where $p$ is the number of predictors. This is an unbiased estimate for $EPE$.
   - Akaike's Information Criterion:
     $$AIC = -2loglik + 2p.$$
     Reduces to $C_p$ for linear models. Will be useful later for non-linear models.
2. Cross-validation (next slide)

Select the model that *minimizes* the $C_p$ or AIC.

The BIC is another useful model selection criterion, but is *not* based on prediction error (more later).

# Cross-Validation

Idea: Use a part of the data for training, the other part for testing.



# More on the $C_p$

$$C_p = SSE + 2p\hat{\sigma}^2.$$

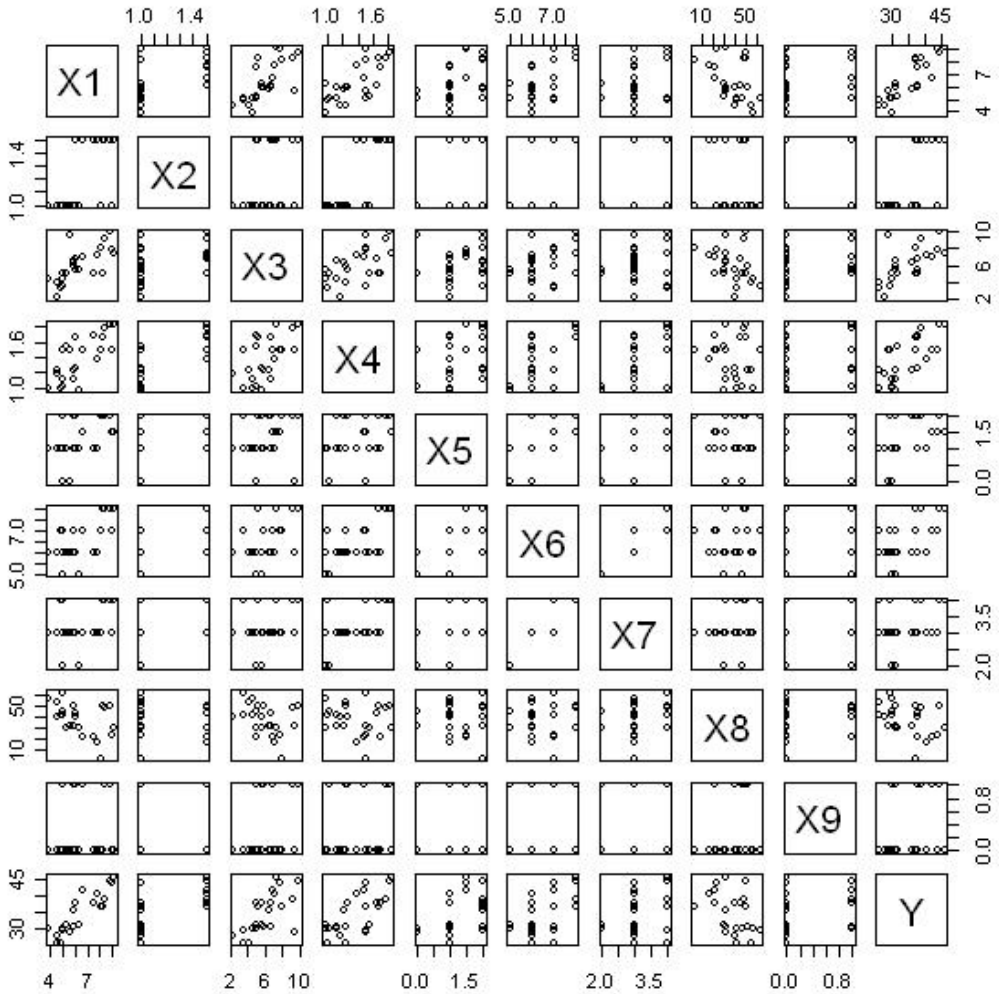
Sometimes also written as:

$$C_p = SSE + 2p\hat{\sigma}^2 - n\hat{\sigma}^2,$$

but the last term doesn't change across models.

1. $\hat{\sigma}^2$ is usually estimated from the largest model.
2. Usual practice: plot $C_p$ versus $p$, choose model with minimum.

# Exploring the model space




## Property values data:

$Y$: Sales price
$X_1$: Local taxes
$X_2$: # bathrooms
$X_3$: Lot size
$X_4$: Living space
$X_5$: Garage
$X_6$: # rooms
$X_7$: # bedrooms
$X_8$: Property age
$X_9$: # fireplaces

$2^9 = 512$ possible models!

# Exploring the model space

1. Forward selection:
   1. Start with null model.
   2. Repeat: add variable with the most significant F-test.
   3. End when no variable has F-test p-value $< \alpha$.
2. Backward elimination:
   1. Start with full model.
   2. Repeat: delete variable with the least significant F-test.
   3. End when all variables have F-test p-value $< \alpha$.
3. Forward + Backward: Same as forward procedure, with option of deleting a variable at each step.
4. All subsets: possible when number of possible predictors is small ($< 20$).