HKU BUSINESS SCHOOL
港大經管學院

Business Statistics

Topic Review of MSBA7002

Weichen Wang

Assistant Professor

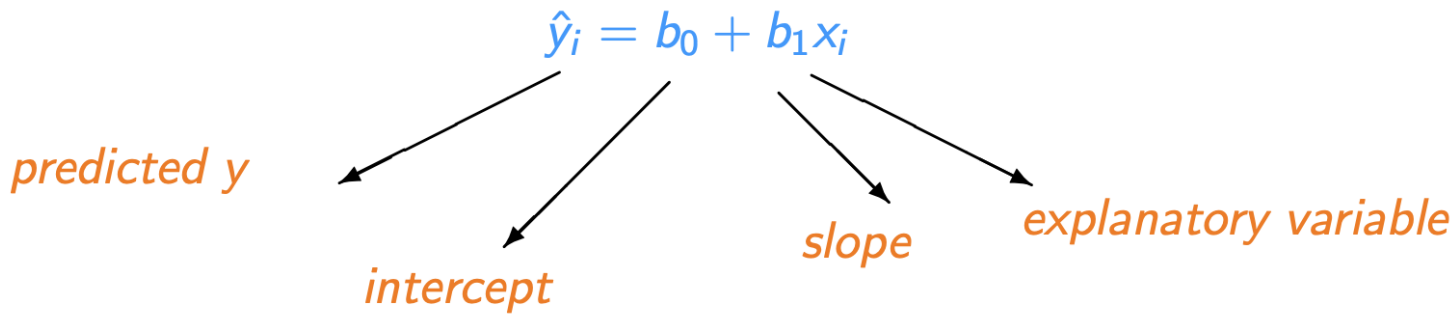
Innovation and Information Management

Lec 1: Linear Regression

Linear Regression

- Key idea
  - a linear approach for modelling the relationship between a scalar response and one or more explanatory variables
- Target
  - Find a set of coefficients to minimize the residual sum of squares (RSS).
- Assumption
  - Linearity; Homoscedasticity; Independence; Normality.
- Algorithm
  - Closed form solution: $\hat{\boldsymbol{\beta}} = (\mathrm{X}^\mathsf{T}\mathrm{X})^{-1}\mathrm{X}^\mathsf{T}\mathbf{y}.$
- Model diagnostic

The Least Squares Line



- For each observation in the data set, your *line* predicts where y should be.
- The *residual* from ith data point is how far the true y value is from where the line predicts.

$$e_i = y_i - \hat{y}_i$$

- Least-squares criterion: Find b_0, b_1 to minimize the residual sum of squares (RSS) or sum of squared errors (SSE)

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Summary

- Coefficients $b_1 = r \times \frac{s_y}{s_x}$ and $b_0 = \bar{y} - b_1\bar{x}$.
- Predicted value $\hat{y}_i = b_0 + b_1 x_i$.
- Actual value y_i.
- Residual $e_i = y_i - \hat{y}_i$.
- We choose the line to make SSE or RSS as small as possible.
- Both for linear relationship between two variables.
  - Same sign between b_1 and r.
- r does not depend on which is x and which is y.

RMSE $$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

R^2 $$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

```
> summary(lm(price~weight))

Call:
lm(formula = price ~ weight)

Residuals:
    Min      1Q  Median      3Q     Max
-85.159 -21.448  -0.869  18.972  79.370

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -259.63      17.32  -14.99   <2e-16 ***
weight       3721.02      81.79   45.50   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.84 on 46 degrees of freedom
Multiple R-squared:  0.9783,    Adjusted R-squared:  0.9778
F-statistic:  2070 on 1 and 46 DF,  p-value: < 2.2e-16
```

RMSE

```
> summary(lm(price~weight))

Call:
lm(formula = price ~ weight)

Residuals:
    Min      1Q  Median      3Q     Max
-85.159 -21.448  -0.869  18.972  79.370

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -259.63      17.32  -14.99   <2e-16 ***
weight       3721.02      81.79   45.50   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.84 on 46 degrees of freedom
Multiple R-squared:  0.9783      Adjusted R-squared:  0.9778
F-statistic:  2070 on 1 and 46 DF,  p-value: < 2.2e-16
```

R Square

Model Diagnostics

- Make sure there are no gross violations of the model
  - Is the relationship between x and y *linear*?
  - Do the residuals show iid normal behavior (i.e., *independent, equal variance, normality*)?
  - Are there *outliers* that may distort the model fit?
- Crucial steps in checking a model
  - A *y vs. x scatterplot* should reveal a linear pattern, linear dependence.
  - A *Residual vs. x scatterplot* should reveal no meaningful pattern.
  - A *Residual vs. Predicted scatterplot* should reveal no meaningful pattern.
  - A *histogram and normal quantile plot* of the residuals should be consistent with the assumption of normality of the errors.

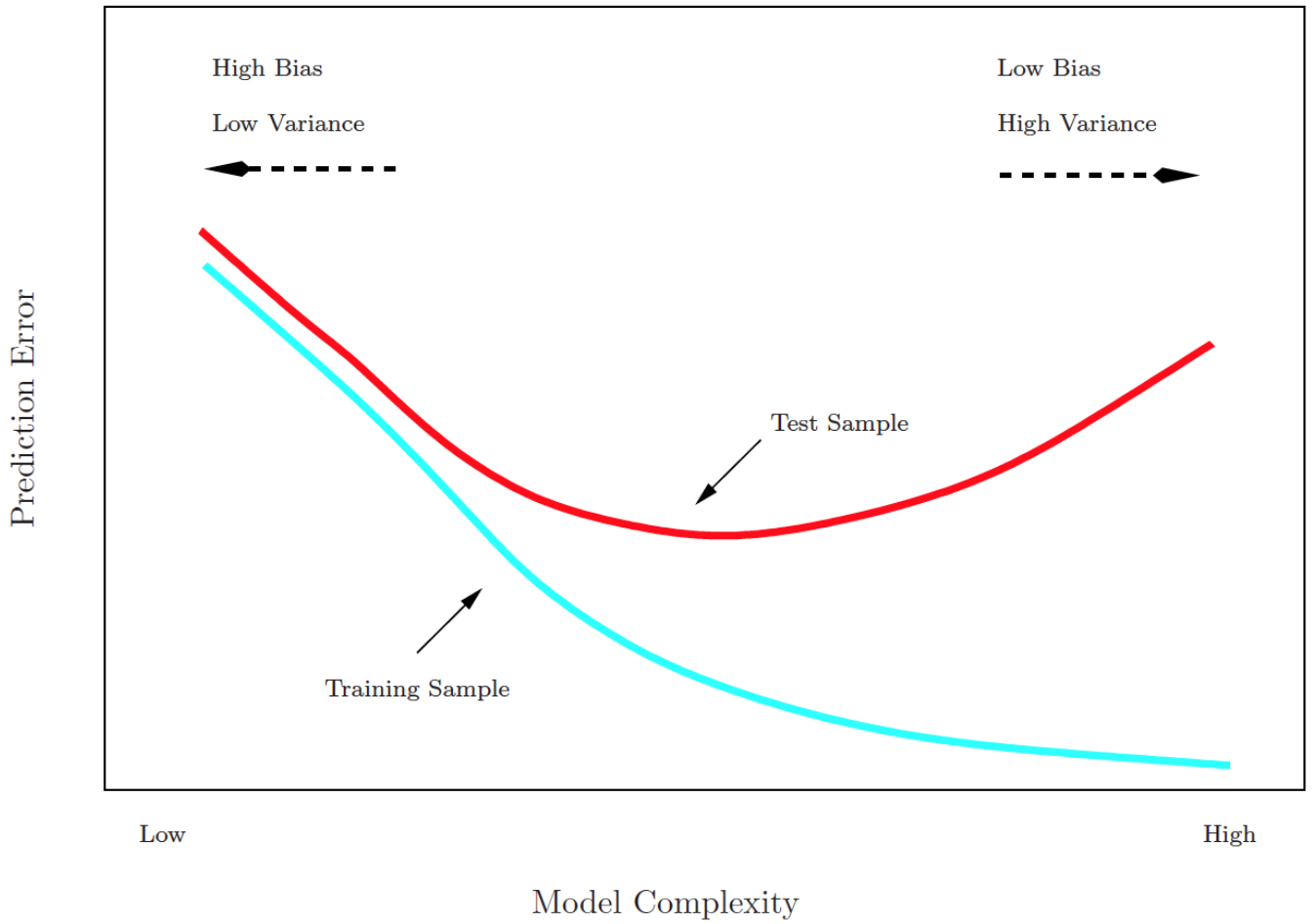
Inference

1. Is $\beta_1 = 0$? $\Longrightarrow$ t-statistics
2. Are all $\beta_j = 0$? $\Longrightarrow$ F-statistics
3. 95% Confidence interval of β_j? $\hat{\beta}_j \pm 2 * SE(\hat{\beta}_j)$
4. 95% Confidence interval of $\hat{y}_i$? $\hat{y}_i \pm 2 * RMSE$

Lec 2: Model Selection (Subset Selection)

Model Complexity & Prediction Error

- Key idea
  - Identify a subset of predictors and use them to fit model for prediction accuracy.



Estimating Test Error: Two Approaches

1. Indirectly estimate test error by making an adjustment to the training error to account for the bias due to overfitting.

2. Directly estimate the test error, using either a validation set approach or a cross-validation approach.

Model Comparison Criteria

Adjustment to the Training Error

- Adjusted R^2
- Mallow's C_p
- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)

Adjusted R^2

- For a least squares model with p variables, the Adjusted R^2 statistic is calculated as

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-p-1)}{\text{TSS}/(n-1)}$$

  where TSS is the total sum of squares.

- A large Adjusted R^2 indicates a model with a small test error.

- Maximizing the Adjusted R^2 is equivalent to minimizing $\frac{\text{RSS}}{n-p-1}$. While RSS always decreases as the number of variables in the model increases, $\frac{\text{RSS}}{n-p-1}$ may increase or decrease.

- Unlike R^2, Adjusted R^2 pays a price for the inclusion of unnecessary variables in the model.

Mallow's C_p and Bayesian Information Criterion

- Mallow's C_p:

$$C_p = \frac{1}{n}(\text{RSS} + 2p\hat{\sigma}^2).$$

- Select the model with the smallest C_p.
- The BIC is defined as

$$BIC = \frac{1}{n}(\text{RSS} + \log(n)p\hat{\sigma}^2).$$

- Select the model with the lowest BIC value.

Akaike Information Criterion

- The AIC is defined for a large class of models fit by maximum likelihood:

$$AIC = -2 \log L + 2 \cdot p$$

where L is the maximized value of the likelihood of the estimated model.

- Select the model with the smallest AIC

- For linear models with Gaussian errors
  - maximum likelihood = least squares
  - C_p and AIC: equivalent

Subset Selection

Best subset and stepwise model selection procedures

Best Subset Selection

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Stepwise Selection

- For computational reasons, best subset selection cannot be applied with very large *p*.
- Best subset selection may also suffer from statistical problems when *p* is large:
  - The larger the search space, the higher the chance of finding models that look good on the training data, even though they might not have any predictive power on future data.
- Thus an enormous search space can lead to overfitting and high variance of the coefficient estimates.
- Forward/Backward stepwise selection

Stepwise Selection

- Both forward and backward selection approach search through only $1 + p(p + 1)/2$ models, and so can be applied in settings where p is too large to apply best subset selection.

- Both forward and backward stepwise selection are not guaranteed to yield the best model containing a subset of the p predictors.

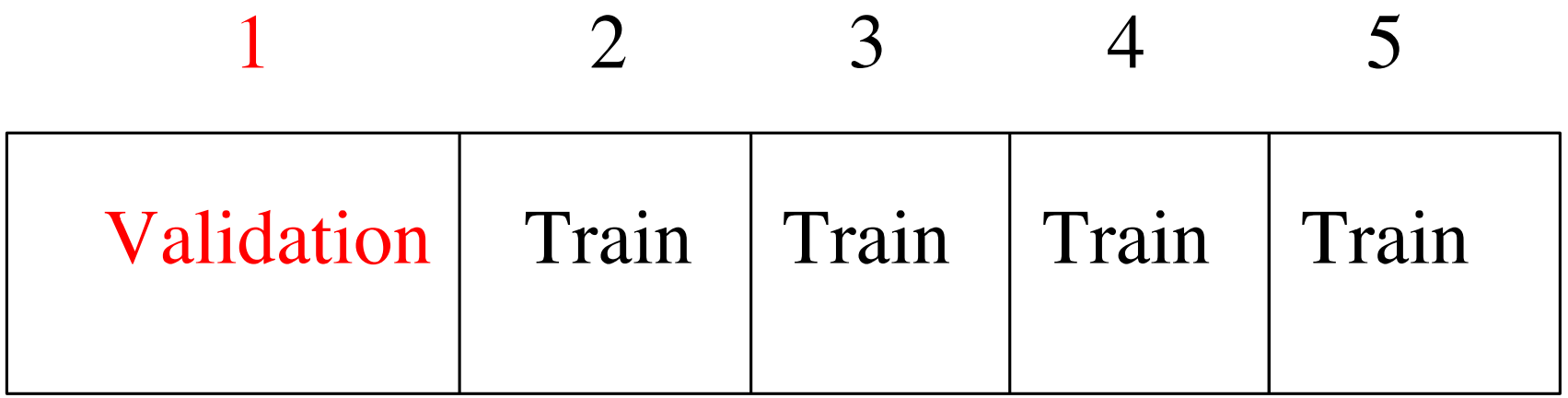
- Backward selection requires that the number of samples n is larger than the number of variables p (so that the full model can be fit). In contrast, forward stepwise can be used even when $n < p$, and so is the only viable subset method when p is very large.

Algorithm Comparison

Stage	Best Subset Selection	Forward Stepwise Selection	Backward Stepwise Selection
Start	Any model	Null model $\mathcal{M}_0$	Full model $\mathcal{M}_p$
Selection	$\mathcal{M}_0, \ldots, \mathcal{M}_p$ sequence doesn't matter Largest R^2	$\mathcal{M}_0 \to \cdots \to \mathcal{M}_p$ sequence matters Largest R^2	$\mathcal{M}_p \to \cdots \to \mathcal{M}_0$ sequence matters Largest R^2
Decision	Chose the best model among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ according to CV, Adjusted R^2, C_p, AIC, BIC.		

K-fold CV Illustration

Divide data into K roughly equal-sized parts ($K = 5$ here)



The Computing Details

- Let the K parts be $C_1, C_2, \ldots C_K$, where C_k denotes the indices of the observations in part k. There are n_k observations in part k: if N is a multiple of K, then $n_k = n/K$.
- Compute

$$\text{CV}_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} \text{MSE}_k$$

  where $\text{MSE}_k = \sum_{i \in C_k} (y_i - \hat{y}_i)^2 / n_k$, and $\hat{y}_i$ is the fit for observation i, obtained from the data with part k removed.
- Setting $K = n$ yields n-fold or *leave-one out cross-validation* (LOOCV).

Lec 3: Shrinkage Methods

Ridge Regression

- Recall that the least squares fitting procedure estimates $\beta_0, \beta_1, \ldots, \beta_p$ using the values that minimize

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

- In contrast, the ridge regression coefficient estimates $\hat{\beta}^R$ are the values that minimize

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2,$$

where $\lambda \geq 0$ is a *tuning parameter*, to be determined separately.

The Bias-Variance Decomposition

- Assume that

$$Y = f(X) + \varepsilon$$

where E(ε)=0 and Var(ε)=σ_ε^2.

- At an input point $X = x_0$, the expected squared prediction error is

$$\begin{aligned}
\mathrm{Err}(x_0) &= E[(Y - \hat{f}(x_0))^2 | X = x_0] \\
&= \sigma_\varepsilon^2 + [\mathrm{E}\hat{f}(x_0) - f(x_0)]^2 + E[\hat{f}(x_0) - \mathrm{E}\hat{f}(x_0)]^2 \\
&= \sigma_\varepsilon^2 + \mathrm{Bias}^2(\hat{f}(x_0)) + \mathrm{Var}(\hat{f}(x_0)) \\
&= \mathrm{Irreducible\ Error} + \mathrm{Bias}^2 + \mathrm{Variance}.
\end{aligned}$$

- The more complex the model, the lower the bias but the higher the variance.

The Lasso

- Ridge regression does have one obvious disadvantage: unlike subset selection, which will generally select models that involve just a subset of the variables, ridge regression will include all p predictors in the final model
- The *Lasso* is a relatively recent alternative to ridge regression that overcomes this disadvantage. The lasso coefficients, $\hat{\beta}^L_\lambda$, minimize the quantity

How do we select best λ?

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \text{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|$$

CV.

- In statistical parlance, the lasso uses an ℓ_1 (pronounced "ell 1") penalty instead of an ℓ_2 penalty. The ℓ_1 norm of a coefficient vector β is given by $\|\beta\|_1 = \sum|\beta_j|$.

Best Subset, Lasso, and Ridge (Another Formulation: Constraint)

- Best Subset

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} I(\beta_j \neq 0) \leq s.$$

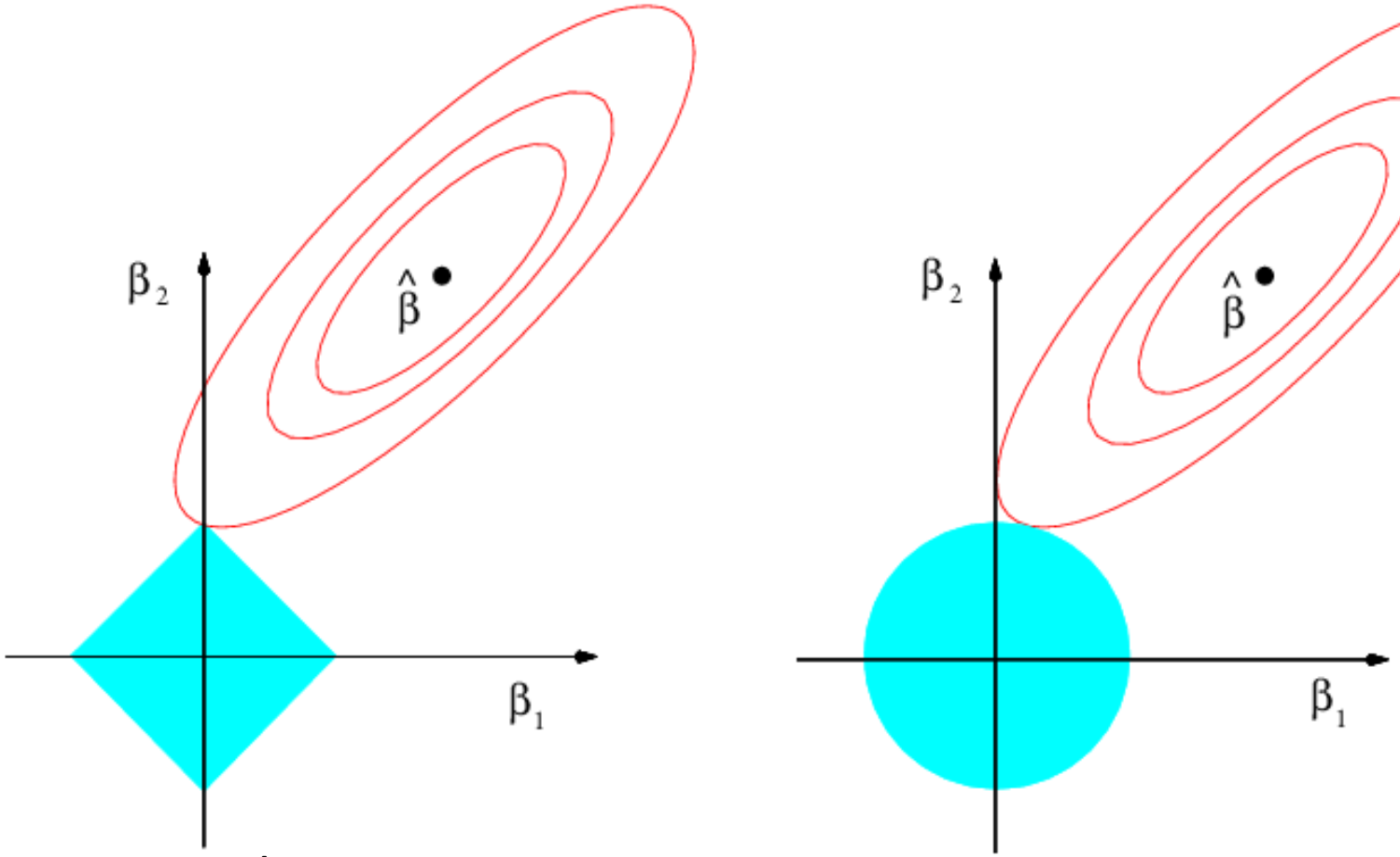
- Lasso

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

- Ridge

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s$$

The Geometry of Ridge and Lasso



$\hat{\beta}$: the least square estimate
red ellipse: region of constant RSS (increasing)
blue region: lasso/ridge constraint (driven by s)

Tuning Parameter Selection

- Similar to subset selection, for ridge regression and lasso, need a method to determine which of the models under consideration is best.
  - Method to select a value for the tuning parameter λ or equivalently, the value of the constraint s.
- Cross-validation provides a simple way to tackle this problem.
  - Choose a grid of λ values, and compute the CV error rate for each value of λ.
  - Select the value of λ for which the CV error is the smallest.
  - Refit the model using all available observations and the selected value of λ.

Dimension Reduction Methods

- The methods that we have discussed so far have involved fitting linear regression models, via least squares or a shrunken approach, using the original predictors, $X_1, X_2, \ldots, X_p$.

- There exist a class of approaches that transform the predictors and then fit a least squares model using the transformed variables.

- We will refer to these techniques as dimension reduction-based regression methods.
  - Principal Component Regression
  - Partial Least Squares

Lec 4: Logistic Regression

Logistic Regression

- Key idea
  - Use a logistic function to model a binary dependent variable.
  - The logistic function

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

- Target
  - Maximize the likelihood function

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1-p(x_{i'}))$$

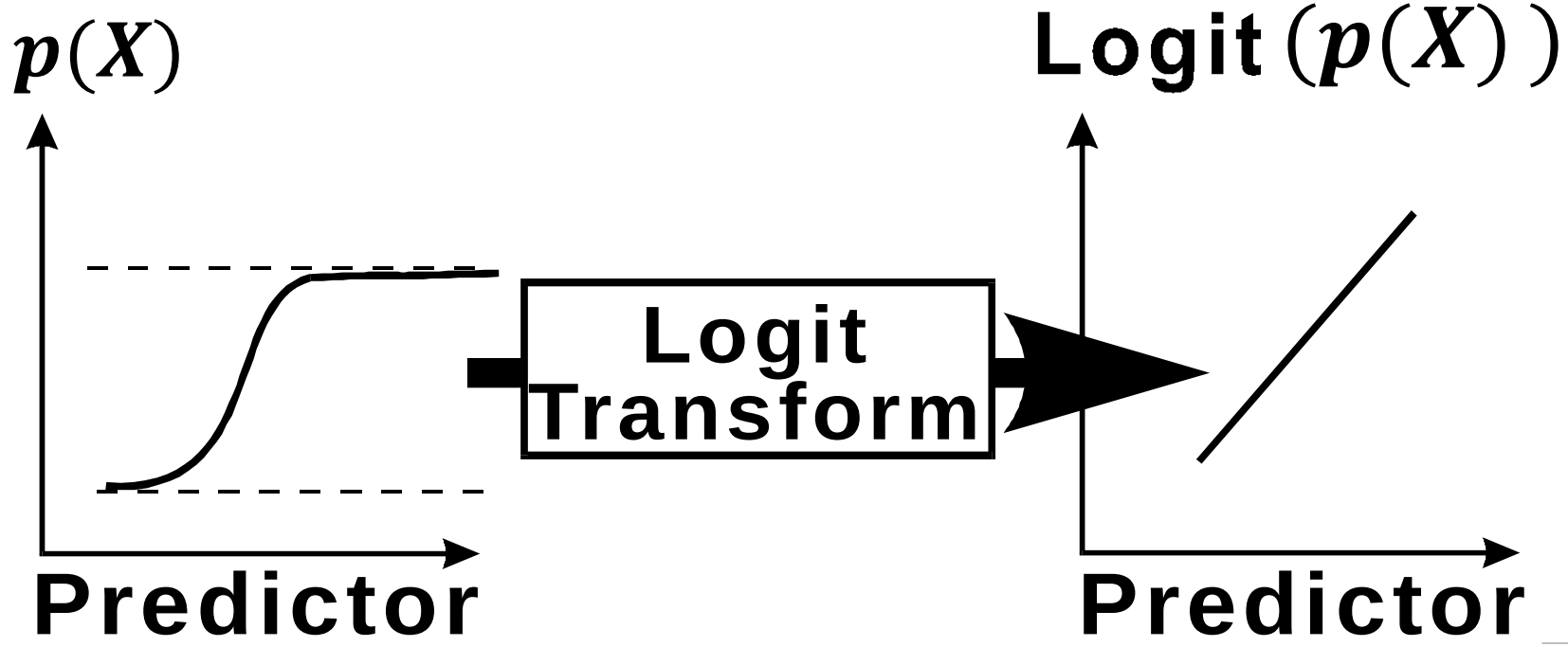
- Algorithm
  - Maximum Likelihood Estimation

Logistic Regression

A bit of rearrangement gives

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X.$$

This monotone transformation is called the *log odds* or *logit* transformation of $p(X)$.



Odds Ratio

- From $p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$, we have $\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x} = e^{\beta_0} e^{\beta_1 x}$

- The odds ratio increases **multiplicatively** by e^{β_1} for every 1-unit increase in x
  - The odds at $X = x + 1$ are e^{β_1} times the odds at $X = x$
  - $\frac{odds(x+1)}{odds(x)} = e^{\beta_1}$
- Therefore, e^{β_1} **is an odds ratio!**

- e^{β_1} represents the change in the odds of the outcome (multiplicatively) by increasing x by 1 unit
  - If $\beta_1 > 0$, the odds and probability increase as x increases ($e^{\beta_1} > 1$)
  - If $\beta_1 < 0$, the odds and probability decrease as x increases ($e^{\beta_1} < 1$)
  - If $\beta_1 = 0$, the odds and probability are the same at all x levels ($e^{\beta_1}=1$)

Log Odds or Logit

- The sign ($\pm$) of β_1 determines whether the **log odds** of y is increasing or decreasing *for every 1-unit increase* in x.
  - If $\beta_1 > 0$, there is an increase in the **log odds** of y for every 1-unit increase in x.
  - If $\beta_1 < 0$, there is a decrease in the **log odds** of y for every 1-unit increase in x.
  - If $\beta_1 = 0$ there is *no linear relationship* between the **log odds** and x.

Maximum Likelihood Estimation

We use maximum likelihood to estimate the parameters.

$$\ell(\beta_0, \beta) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i)).$$

This *likelihood* gives the probability of the observed zeros and ones in the data. We pick β_0 and β_1 to maximize the likelihood of the observed data.

Most statistical packages can fit linear logistic regression models by maximum likelihood. In `R` we use the `glm` function.

	Coefficient	Std. Error	Z-statistic	P-value
`Intercept`	-10.6513	0.3612	-29.5	< 0.0001
`balance`	0.0055	0.0002	24.9	< 0.0001

Multiple Logistic Regression

- Extension to more than one predictor variable (either numeric or dummy variables).
- With p predictors, the model is written:

$$p(x) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}$$

- Adjusted Odds ratio for raising x_i by 1 unit, holding all other predictors constant:

$$OR_i = e^{\beta_i}$$

- odds = (base odds) OR_1 OR_2 ... OR_k
- Many models have nominal/ordinal predictors, and widely make use of dummy variables

Multivariate Logistic Regression

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.8690	0.4923	-22.08	< 0.0001
balance	0.0057	0.0002	24.74	< 0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	-0.6468	0.2362	-2.74	0.0062

Lec 5: Multinomial Regression

Nominal and Ordinal Responses

- Nominal response
  - Red, green, blue
  - Yes, no
  - Sick, healthy
- Ordinal response
  - Young, middle aged, old
  - Dislike very much, dislike, no opinion, like, like very much

Nominal Logistic Regression

- Choose one category as the reference category, say the 1st category
- Define the logits for the other categories as

$$\text{logit}(\pi_j) \equiv \log(\frac{\pi_j}{\pi_1}) = x^T\beta_j, \quad \text{for } j = 2, \ldots, J.$$

- The joint density is

$$f(\mathbf{y}|n) = (\pi_1)^{y_1} \ldots (\pi_J)^{y_J} \frac{n!}{y_1! \ldots y_J!},$$

which leads to the following likelihood function,

$$l(\beta|\mathbf{y}, n) \propto \prod_{j=2}^{J} (\frac{\pi_j}{\pi_1})^{y_j} = \exp(\sum_j y_j x^T \beta_j).$$

Nominal Logistic Regression Estimate

- Given MLE, we have

$$\hat{\pi}_j = \hat{\pi}_1 \exp(x^T \hat{\beta}_j), \quad \text{for } j = 2, \ldots, J.$$

- Since the probabilities add up to 1, we have

$$\hat{\pi}_1 = \frac{1}{1 + \sum_{j=2}^{J} \exp(x^T \hat{\beta}_j)}$$

$$\hat{\pi}_j = \frac{\exp(x_j^T \hat{\beta}_j)}{1 + \sum_{j=2}^{J} \exp(x^T \hat{\beta}_j)}, \quad \text{for } j = 2, \ldots, J.$$

- Changing the reference category won't change the above probabilities.

Example: Car Preference

Define the following three dummy variables,

- x_1: the indicator of men;
- x_2: the indicator of age 24-40 years;
- x_3: the indicator of age > 40 years.

The model is then

$$\log(\frac{\pi_j}{\pi_1}) = \beta_{0j} + \beta_{1j}x_1 + \beta_{2j}x_2 + \beta_{3j}x_3, \quad j = 2, 3.$$

R Command

```
> car <- data.frame(res.unim=c(26, 9, 5, 40, 17, 8),
          res.im=c(12, 21, 14, 17, 15, 15),
          res.veim=c(7, 15, 41, 8, 12, 18),
          sex=c(rep("F", 3), rep("M",3)),
          age=rep(c("18-23", "24-40", ">40"), 2))
> car
  res.unim res.im res.veim sex   age
1       26     12        7   F 18-23
2        9     21       15   F 24-40
```

```
> library(nnet)  ### special library containing ‘‘multinom’’
> options(contrasts=c("contr.treatment", "contr.poly"))
> car.mult <- multinom(cbind(res.unim, res.im, res.veim)~sex+age,
                                  data=car)
> summary(car.mult)
Coefficients:
  (Intercept)       sexM age24-40   age>40
2  -0.5907992 -0.3881301 1.128268 1.587709
3  -1.0390726 -0.8130202 1.478104 2.916757

Std. Errors:
  (Intercept)      sexM  age24-40    age>40
2   0.2839756 0.3005115 0.3416449 0.4028997
3   0.3305014 0.3210382 0.4009256 0.4229276
```

Ordinal Logistic Regression

- Ordinal responses are common in marketing research, opinion polls, and so on where soft measures are common.

- Cumulative logit model

$$\log \frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_J} = x^T \beta_j.$$

- Special case: Proportional odds model

$$\log \frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_J} = \beta_{0j} + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}.$$

Example: Car Preference

The following proportional odds model was fitted to the data:

$$\log\left(\frac{\pi_1}{\pi_2+\pi_3}\right) = \beta_{01} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

$$\log\left(\frac{\pi_1+\pi_2}{\pi_3}\right) = \beta_{02} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

which leads to the estimates below,

$$\beta_{01} = 0.044, \beta_{02} = 1.655, \beta_1 = 0.576,$$

$$\beta_2 = -1.147, \beta_3 = -2.232.$$

R Command

```
> library(MASS)
> freq <- c(car$res.unim, car$res.im, car$res.veim)
> res <- c(rep(c("unim", "im", "veim"), c(6,6,6)))
> res <- factor(res, levels=c("unim", "im", "veim"), ordered=T)
> car.ord <- data.frame(res=res, sex=rep(car$sex, 3),
                        age=rep(car$age, 3), freq=freq)
> car.polr <- polr(res~sex+age, data=car.ord, weights=freq)
```

```
> car.polr
Coefficients:
      sexM   age24-40     age>40
-0.5762219  1.1470976  2.2324560

Intercepts:
   unim|im    im|veim
0.04353746 1.65497620

Residual Deviance: 581.2956
AIC: 591.2956
```

Lec 6: Poisson Regression

Poisson Regression

- Poisson regression is used for modeling count data (neither quantitative nor qualitative).
- Three ingredients:
  - Likelihood of response = Poisson density

$$\ell(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{n} \frac{e^{-\lambda(x_i)} \lambda(x_i)^{y_i}}{y_i!},$$

  - Find expectation of y condition on x

$$\mathrm{E}(Y|X_1, \ldots, X_p) = \lambda(X_1, \ldots, X_p)$$

  - Link expectation with the linear function of predictors

$$\log(\lambda(X_1, \ldots, X_p)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Estimate β by maximizing the likelihood.

Poisson vs Linear Regression

- An increase in X_j by one unit is associated with a change in $E(Y) = \lambda = e^{X\beta}$ by a factor of $\exp(\beta_j)$.
  - By contrast, in linear regression, an increase in X_j by one unit is associated with an increase of β_j in $E(Y) = \mu = X\beta$.

- Poisson regression implicitly assumes that mean equals variance.
  - By contrast, linear regression assumes the variance takes on a constant value.

- Poisson regression gives non-negative predictions.
  - By contrast, linear regression predictions can be negative.

R implementation

- Recall logistic R command:

  *fit <- **glm**(Y~X, data, family=**binomial**(logit))*

- "glm" here means generalized linear models.

```
> mod.pois <- glm(
    bikers ~ mnth + hr + workingday + temp + weathersit,
    data = Bikeshare, family = poisson
  )
> summary(mod.pois)

Call:
glm(formula = bikers ~ mnth + hr + workingday + temp + weathersit,
    family = poisson, data = Bikeshare)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-20.7574   -3.3441   -0.6549    2.6999   21.9628

Coefficients:
                          Estimate Std. Error  z value Pr(>|z|)
(Intercept)               2.693688   0.009720  277.124  < 2e-16 ***
mnthFeb                   0.226046   0.006951   32.521  < 2e-16 ***
mnthMarch                 0.376437   0.006691   56.263  < 2e-16 ***
mnthApril                 0.691693   0.006987   98.996  < 2e-16 ***
mnthMay                   0.910641   0.007436  122.469  < 2e-16 ***
...
hr20                      1.370588   0.008973  152.737  < 2e-16 ***
hr21                      1.118568   0.009215  121.383  < 2e-16 ***
hr22                      0.871879   0.009536   91.429  < 2e-16 ***
hr23                      0.481387   0.010207   47.164  < 2e-16 ***
workingday                0.014665   0.001955    7.502 6.27e-14 ***
temp                      0.785292   0.011475   68.434  < 2e-16 ***
weathersitcloudy/misty   -0.075231   0.002179  -34.528  < 2e-16 ***
weathersitlight rain/snow -0.575800   0.004058 -141.905  < 2e-16 ***
weathersitheavy rain/snow -0.926287   0.166782   -5.554 2.79e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1052921  on 8644  degrees of freedom
Residual deviance:  228041  on 8605  degrees of freedom
AIC: 281159

Number of Fisher Scoring iterations: 5
```

Lec 7: Discriminant Analysis

Bayes Theorem for Classification

Thomas Bayes was a famous mathematician whose name represents a big subfield of statistical and probablistic modeling. Here we focus on a simple result, known as Bayes theorem:

$$\Pr(Y = k|X = x) = \frac{\Pr(X = x|Y = k) \cdot \Pr(Y = k)}{\Pr(X = x)}$$

One writes this slightly differently for discriminant analysis:

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}, \quad \text{where}$$

- $f_k(x) = \Pr(X = x|Y = k)$ is the *density* for X in class k. Here we will use normal densities for these, separately in each class.
- $\pi_k = \Pr(Y = k)$ is the marginal or *prior* probability for class k.

Linear Discriminant Analysis when $p = 1$

The Gaussian density has the form

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}$$

Here μ_k is the mean, and σ_k^2 the variance (in class k). We will assume that all the $\sigma_k = \sigma$ are the same.

Plugging this into Bayes formula, we get a rather complex expression for $p_k(x) = \Pr(Y = k|X = x)$:

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2}}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_l}{\sigma}\right)^2}}$$

Happily, there are simplifications and cancellations.

Discriminant Functions

To classify at the value $X = x$, we need to see which of the $p_k(x)$ is largest. Taking logs, and discarding terms that do not depend on k, we see that this is equivalent to assigning x to the class with the largest *discriminant score*:

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

Note that $\delta_k(x)$ is a *linear* function of x.

If there are $K = 2$ classes and $\pi_1 = \pi_2 = 0.5$, then one can see that the *decision boundary* is at

$$x = \frac{\mu_1 + \mu_2}{2}.$$

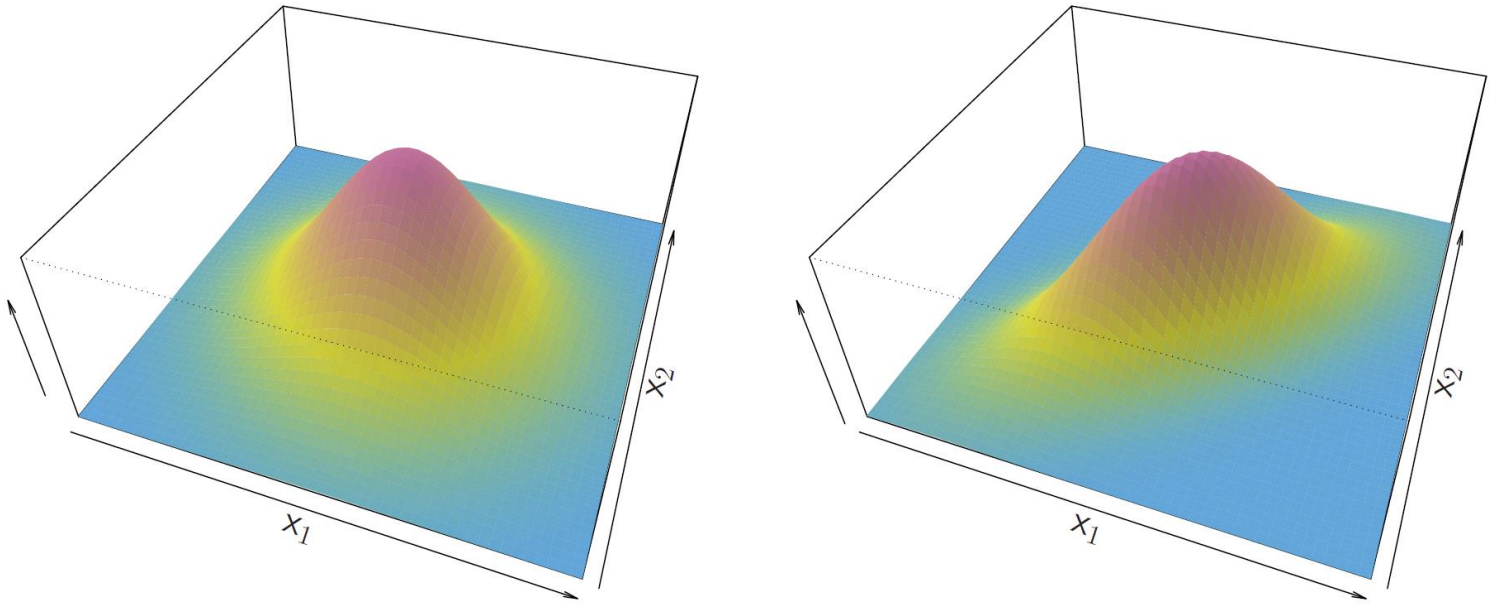
(See if you can show this)

Estimating the Parameters

$$
\begin{aligned}
\hat{\pi}_k &= \frac{n_k}{n} \\
\hat{\mu}_k &= \frac{1}{n_k} \sum_{i:\, y_i = k} x_i \\
\hat{\sigma}^2 &= \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:\, y_i = k} (x_i - \hat{\mu}_k)^2 \\
&= \sum_{k=1}^{K} \frac{n_k - 1}{n-K} \cdot \hat{\sigma}_k^2
\end{aligned}
$$

where $\hat{\sigma}_k^2 = \frac{1}{n_k - 1} \sum_{i:\, y_i = k} (x_i - \hat{\mu}_k)^2$ is the usual formula for the estimated variance in the kth class.

Linear Discriminant Analysis when $p > 1$



$$\text{Density: } f(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \mathbf{\Sigma}^{-1}(x-\mu)}$$

$$\text{Discriminant function: } \delta_k(x) = x^T \mathbf{\Sigma}^{-1}\mu_k - \frac{1}{2}\mu_k^T \mathbf{\Sigma}^{-1}\mu_k + \log \pi_k$$

Despite its complex form,

$\delta_k(x) = c_{k0} + c_{k1}x_1 + c_{k2}x_2 + \ldots + c_{kp}x_p$ — a linear function.

From $\delta_k(x)$ to Probabilities

Once we have estimates $\hat{\delta}_k(x)$, we can turn these into estimates for class probabilities:

$$\widehat{\Pr}(Y = k|X = x) = \frac{e^{\hat{\delta}_k(x)}}{\sum_{l=1}^{K} e^{\hat{\delta}_l(x)}}.$$

So classifying to the largest $\hat{\delta}_k(x)$ amounts to classifying to the class for which $\widehat{\Pr}(Y = k|X = x)$ is largest.

When $K = 2$, we classify to class 2 if $\widehat{\Pr}(Y = 2|X = x) \geq 0.5$, else to class 1.

Classification Terminology

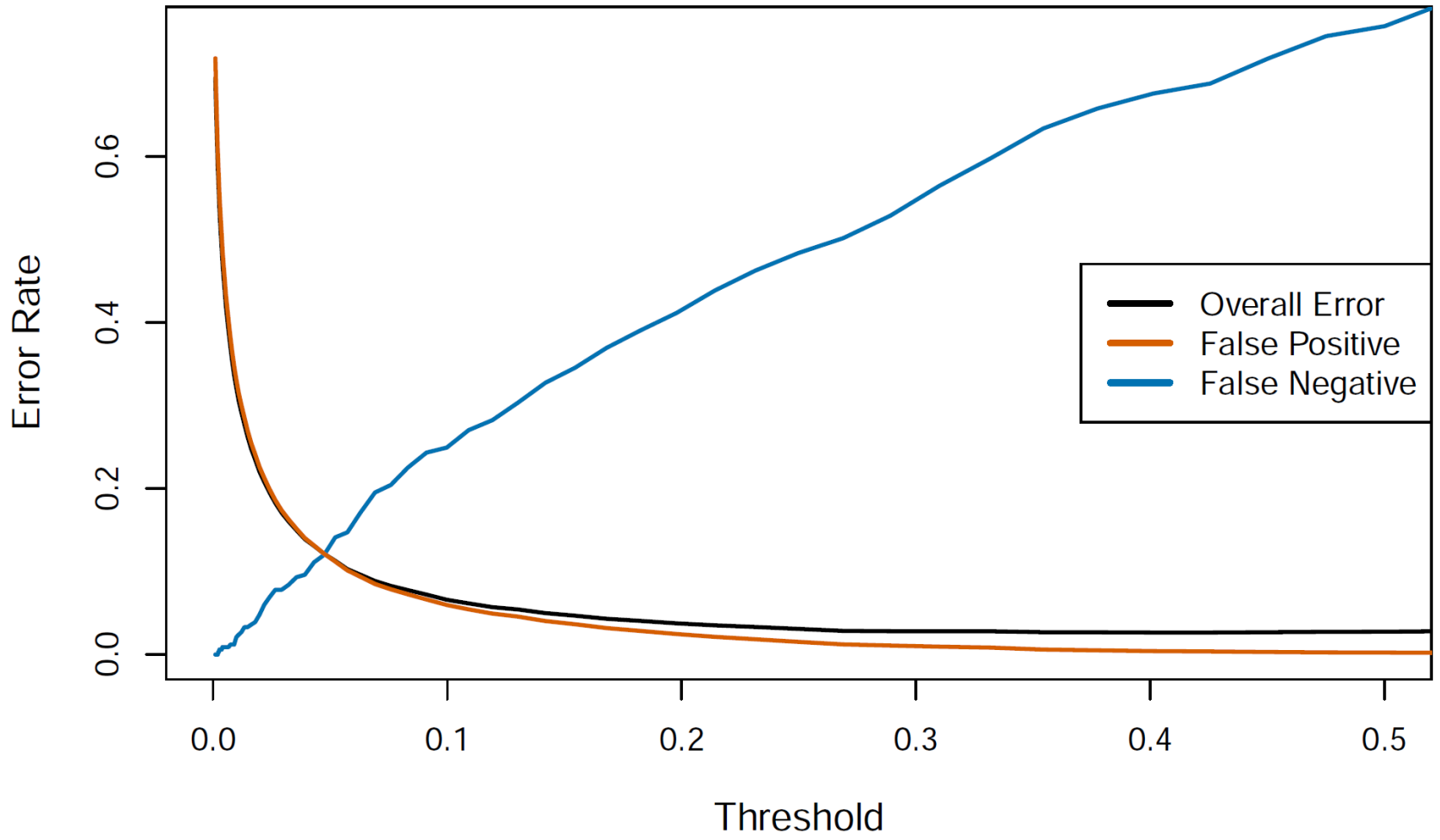
		Predicted class		
		− or Null	+ or Non-null	Total
True class	− or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N*	P*	

TABLE 4.6. *Possible results when applying a classifier or diagnostic test to a population.*

Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P*	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N*	

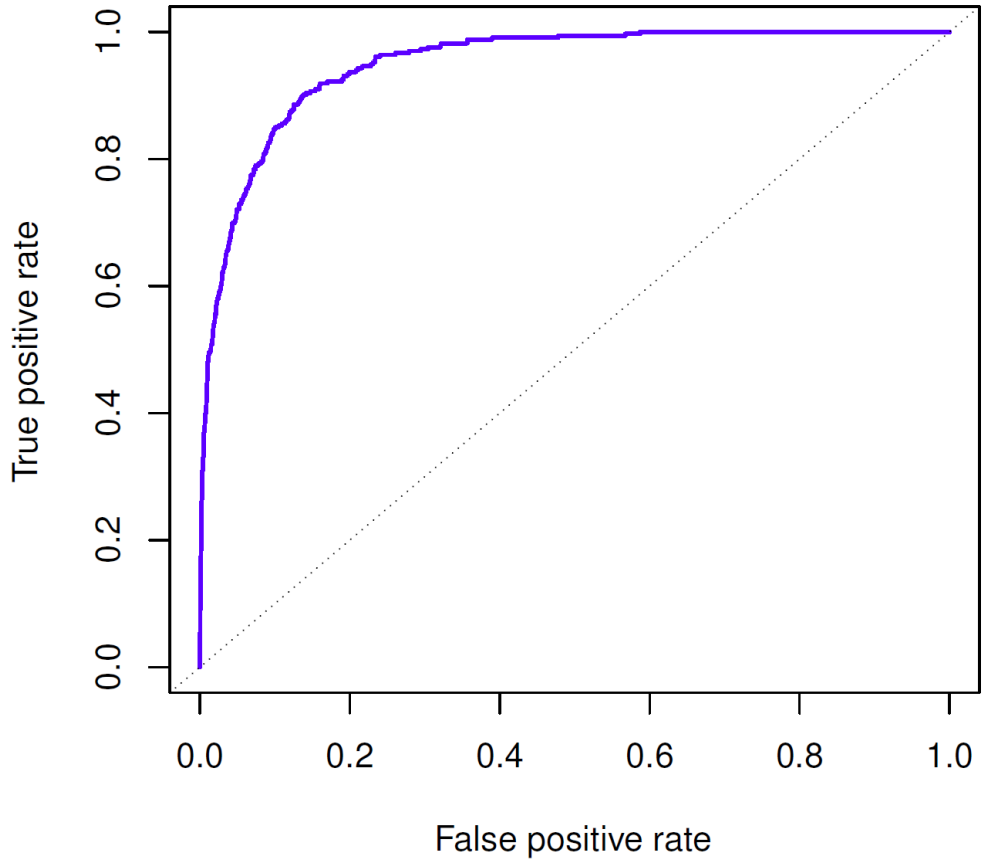
TABLE 4.7. *Important measures for classification and diagnostic testing, derived from quantities in Table 4.6.*

Varying the Threshold



In order to reduce the false negative rate, we may want to reduce the threshold to 0.1 or less

The Receiver Operating Characteristics (ROC) Curve



- The ROC plot
  - False positive rate: 1-Specificity; healthy identified as not
  - True positive rate: Sensitivity; sick identified as so
- Higher area under the curve (AUC) is better

Other Forms of Discriminant Analysis

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

When $f_k(x)$ are Gaussian densities, with the same covariance matrix $\mathbf{\Sigma}$ in each class, this leads to linear discriminant analysis. By altering the forms for $f_k(x)$, we get different classifiers.

- With Gaussians but different $\mathbf{\Sigma}_k$ in each class, we get *quadratic discriminant analysis*.
- With $f_k(x) = \prod_{j=1}^{p} f_{jk}(x_j)$ (conditional independence model) in each class we get *naive Bayes*. For Gaussian this means the $\mathbf{\Sigma}_k$ are diagonal.
- Many other forms, by proposing specific density models for $f_k(x)$, including nonparametric approaches.

Quadratic Discriminant Analysis

With different covariance matrix Σ_k, the Bayes classifier assigns an observation $X = x$ to the class with the largest

$$\begin{aligned}\delta_k(x) &= -\frac{1}{2}(x-\mu_k)^T\mathbf{\Sigma}_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\mathbf{\Sigma}_k| + \log\pi_k \\ &= -\frac{1}{2}x^T\mathbf{\Sigma}_k^{-1}x + x^T\mathbf{\Sigma}_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\mathbf{\Sigma}_k^{-1}\mu_k - \frac{1}{2}\log|\mathbf{\Sigma}_k| + \log\pi_k\end{aligned}$$

The decision boundary is a quadratic function of x.

Naive Bayes

- Recall that Bayes Theorem gives

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}.$$

- QDA assumes each $f_k(x)$ is p-dim multivariate Gaussian(μ_k, Σ_k).
- LDA further assume $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_K = \Sigma$.
- Naive Bayes assumes features are independent

$$f_k(x) = f_{k1}(x_1) \times f_{k2}(x_2) \times \cdots \times f_{kp}(x_p),$$

where $f_{kj}(x_j)$ can be either Gaussian or a general unknown distribution.

Comparison of LDA, QLA, NB, Logistic

- LDA is a special case of QDA.
- Any linear classifier (e.g. LDA) is a special case of NB.

	LDA	QDA	NB
Gaussian $f_{kj}(x_j)$?	Yes	Yes	Not necessarily
Diagonal Σ_k?	No	No	Yes
Shared $\Sigma_k = \Sigma$?	Yes	No	No

Lec 8: Support Vector Machine

Hyperplane

$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = 0$ defines a p-dimensional hyperplane.

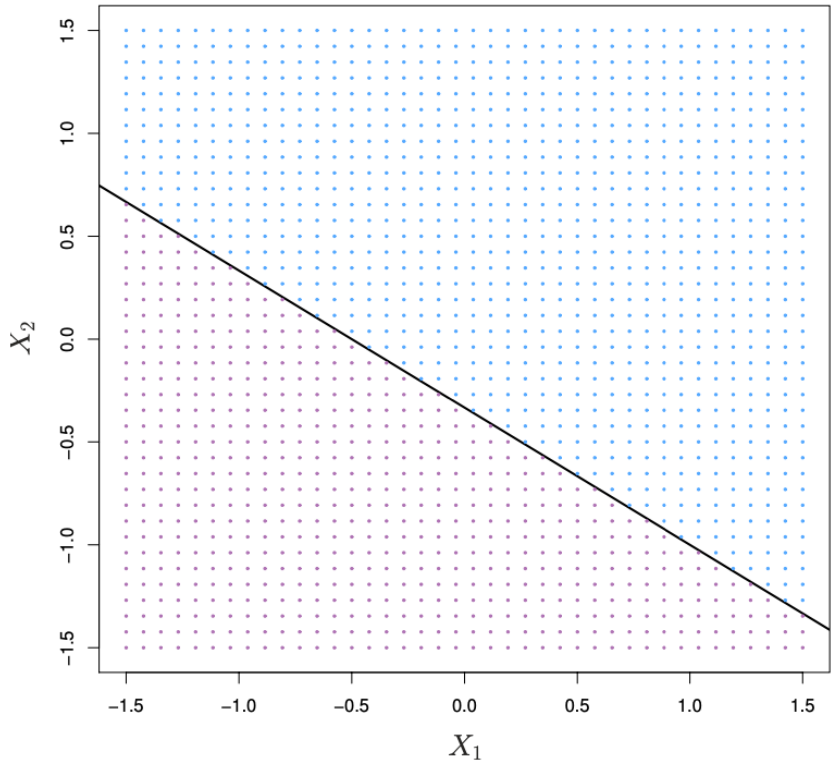


FIGURE 9.1. *The hyperplane* $1 + 2X_1 + 3X_2 = 0$ *is shown. The blue region is the set of points for which* $1 + 2X_1 + 3X_2 > 0$, *and the purple region is the set of points for which* $1 + 2X_1 + 3X_2 < 0$.

Construction of Maximal Margin Classifier

- Collect n training observations $X_1, X_2, \ldots, X_n$ of p dimensions, and associated class labels $y_1, y_2, \ldots, y_n \in \{-1,1\}$.

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{maximize}}\, M$$

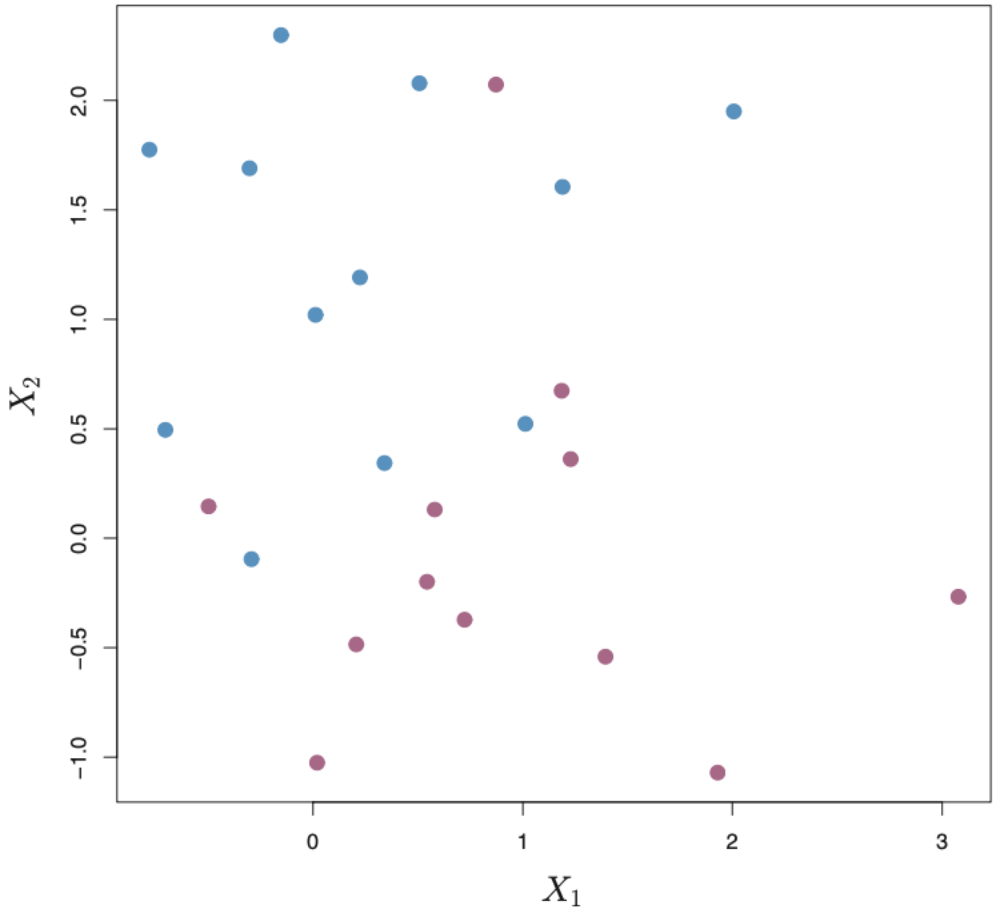
$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M \;\; \forall\, i = 1, \ldots, n.$$

- The inequality constraint: each observation is on the correct side of the hyperplane, provided $M > 0$.
- The equality constraint: just a normalization, since rescaling β does not change the hyperplane.
- Maximal M: each observation is on the correct side and at least a distance M from the hyperplane.

The Non-separable Case

- If no separating hyperplane exists, that means no maximal margin classifier, or no solution with $M > 0$.



The Non-separable Case: Soft Margin

- The non-separable case pursues:
  - Greater robustness to individual observations,
  - Better classification for most of the training observations.

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n}{\text{maximize}} \quad M$$

$$\text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1,$$

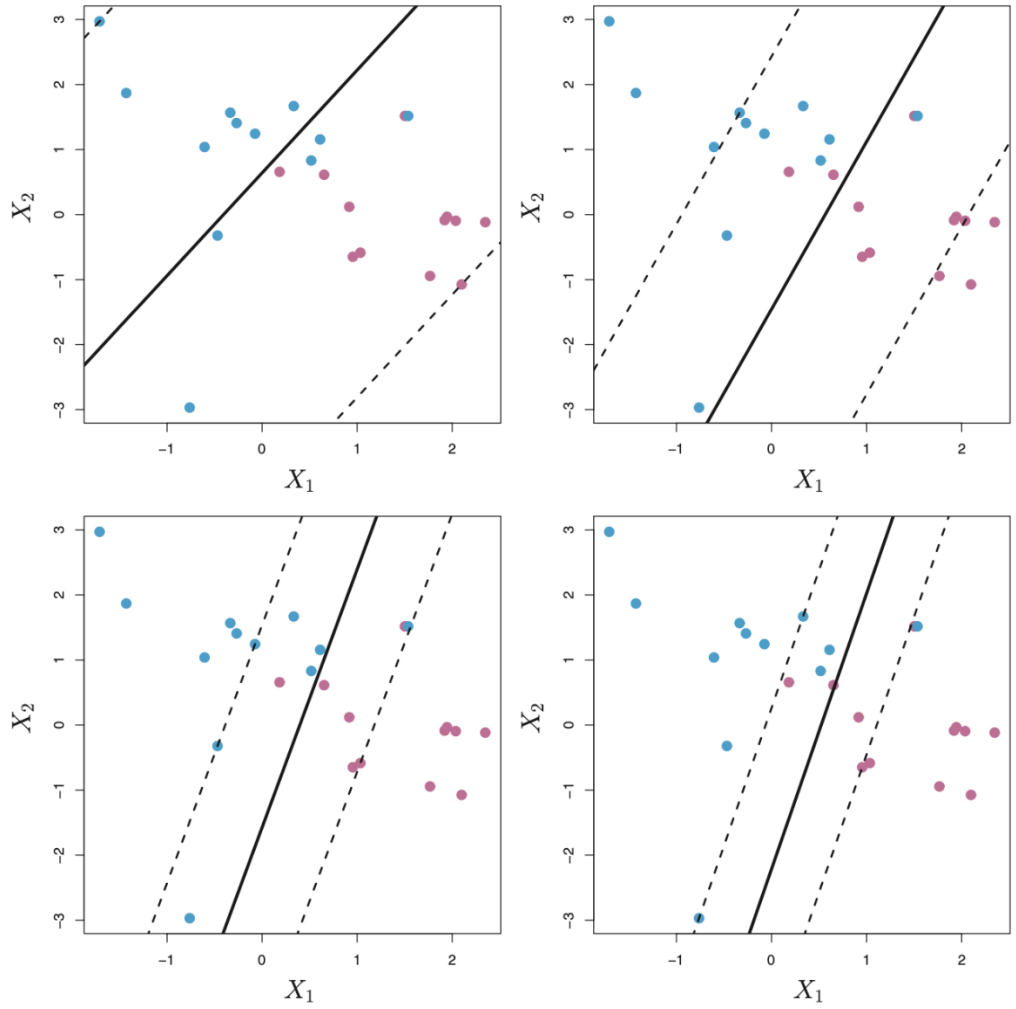
$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

$$\epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq C,$$

- Slack variables ϵ_i allow observations to be on the wrong side.
- Cost C is a budget for the amount that the margin can be violated.
- "Support Vectors": observations that lie on the margin, or on the wrong side of the margin.

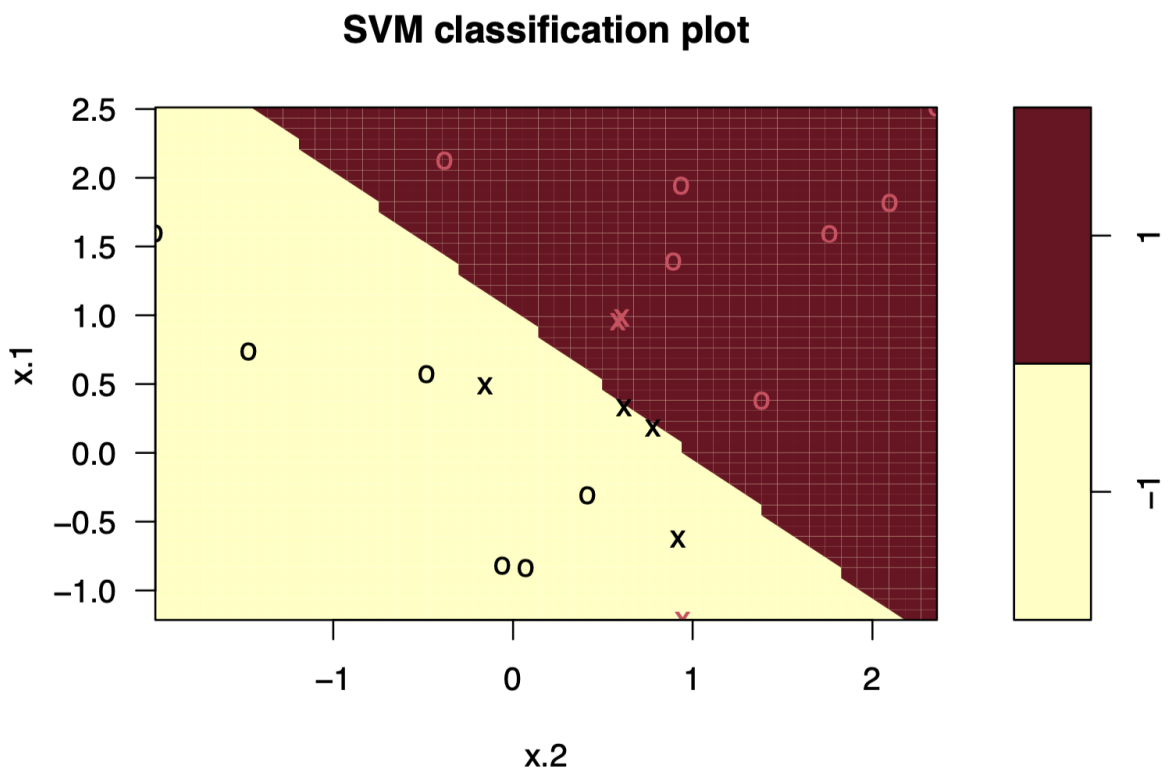
The Non-separable Case: Soft Margin

- Larger C leads to wider margin with more violations, thus a classifier that is more biased but has lower variance.



Implementation

```
plot(svmfit, dat)
```




```
dat <- data.frame(x = x, y = as.factor(y))
# response must be a factor to perform classification

library(e1071)
svmfit <- svm(y ~ ., data = dat, kernel = "linear",
              cost = 10, scale = FALSE)
```

```
tune.out <- tune(svm, y ~ ., data = dat, kernel = "linear",
                 ranges = list(cost = c(0.001, 0.01, 0.1, 1, 5, 10, 100)))
summary(tune.out)
```

```
##
## Parameter tuning of 'svm':
##
## - sampling method: 10-fold cross validation
##
## - best parameters:
##  cost
##   0.1
##
## - best performance: 0.05
##
## - Detailed performance results:
##    cost error dispersion
## 1 1e-03  0.55  0.4377975
## 2 1e-02  0.55  0.4377975
## 3 1e-01  0.05  0.1581139
## 4 1e+00  0.15  0.2415229
## 5 5e+00  0.15  0.2415229
## 6 1e+01  0.15  0.2415229
## 7 1e+02  0.15  0.2415229
```

Compare with LDA and Logistic Regression

- Support vector classifier is based on a small subset of the training samples (the support vectors). It is robust to samples far away from the hyperplane.
- LDA classifier depends on the mean and covariance matrix of all of the observations within each class. It is not robust to any observations.
- Logistic regression, similar to SVM, also has low sensitivity to observations far from the decision boundary.

SVMs with More than Two Classes

Two popular ideas:

- One-Versus-One Approach:
  - Construct $\binom{K}{2}$ SVMs for each pair of classes.
  - Assign a test observation to the class to which it was most frequently assigned in these $\binom{K}{2}$ pairwise classifications.
- One-Versus-All Approach:
  - Construct K SVMs, each time comparing one of the K classes to the remaining $K-1$ classes.
  - Assign a test observation x to the class for which $\beta_{0k} + \beta_{1k}x_1 + \cdots + \beta_{pk}x_p$ is largest.

Support Vector Regression

- Classification:

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{minimize}} \left\{ \sum_{i=1}^{n} \max\left[0, 1 - y_i f(x_i)\right] + \lambda \sum_{j=1}^{p} \beta_j^2 \right\},$$

- Regression:

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{minimize}} \left\{ \sum_{i=1}^{n} \max[0, |y_i - f(x_i)| - \epsilon] + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

- ϵ-insensitive loss + ridge penalty

Lec 9: Principal Component Analysis

The First Principal Component

- Consider a set of features: $X_1, X_2, \ldots, X_p$ on n individuals, and their normalized linear combination:

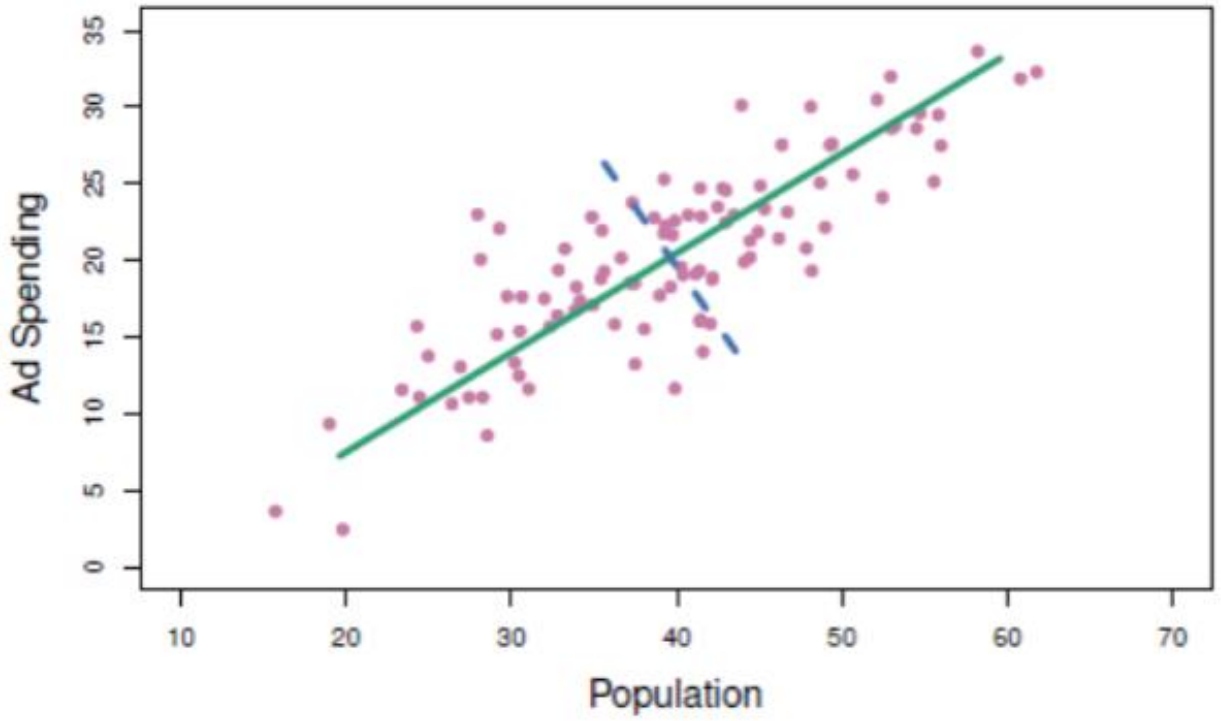
$$Z = \phi_1 X_1 + \phi_2 X_2 + \cdots + \phi_p X_p \text{ with } \sum_{j=1}^{p} \phi_j^2 = 1$$

- The first principal component (PC) is one such linear combination $Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + \cdots + \phi_{p1} X_p$ that has the largest variance

$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximize}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} \phi_{j1}^2 = 1.$$

- The entries $z_{11}, z_{21}, \ldots, z_{n1}$ are the PC scores.
- The PC loading vector: $\phi_1 = (\phi_{11}, \phi_{21}, \ldots, \phi_{p1})^T$

Geometry of PCA



- The 1st PC loading vector ϕ_1 defines a direction in the feature space along which the data vary the most – the green line
- If we project the data points onto this direction, the project values are the PC scores in Z_1.

Geometry of PCA

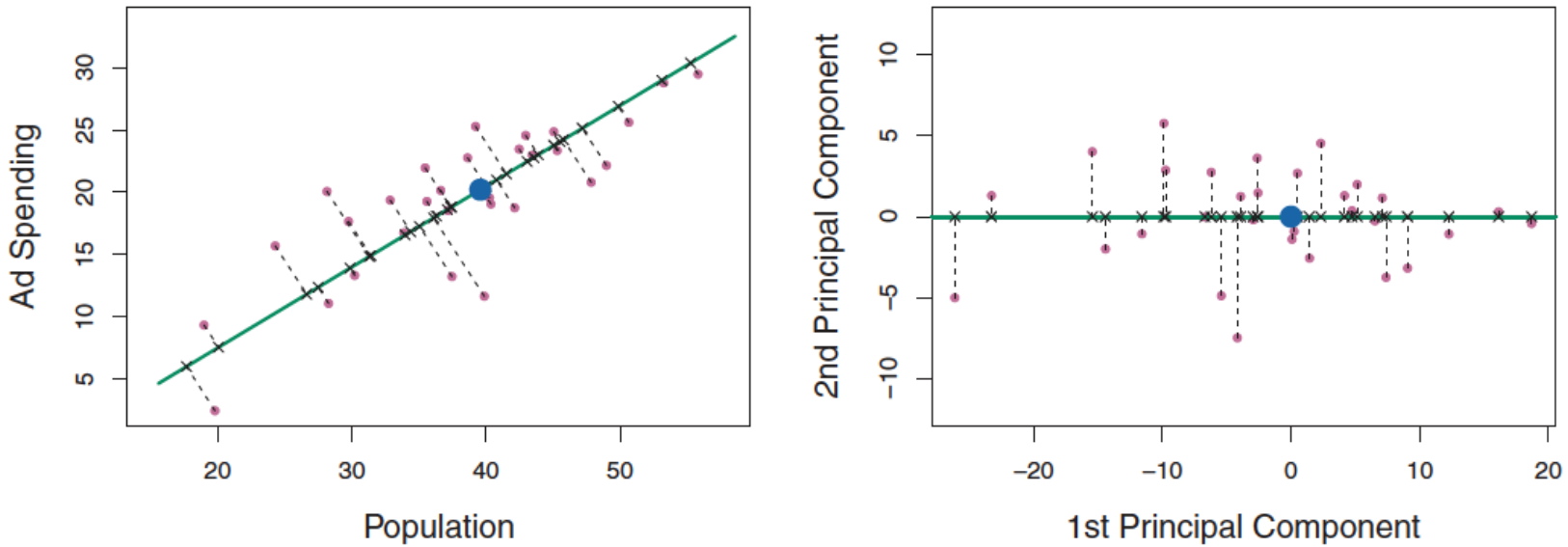


FIGURE 6.15. *A subset of the advertising data. The mean* pop *and* ad *budgets are indicated with a blue circle.* Left: *The first principal component direction is shown in green. It is the dimension along which the data vary the most, and it also defines the line that is closest to all* n *of the observations. The distances from each observation to the principal component are represented using the black dashed line segments. The blue dot represents* $(\overline{\texttt{pop}}, \overline{\texttt{ad}})$. Right: *The left-hand panel has been rotated so that the first principal component direction coincides with the* x*-axis.*

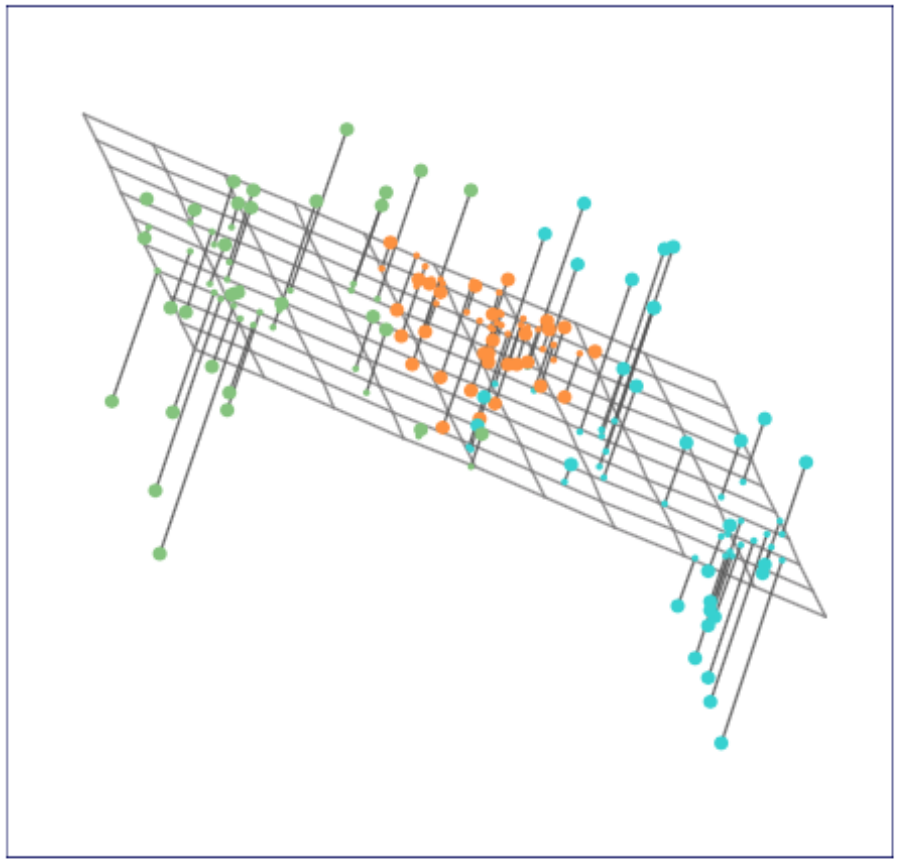
Higher Order Principal Components

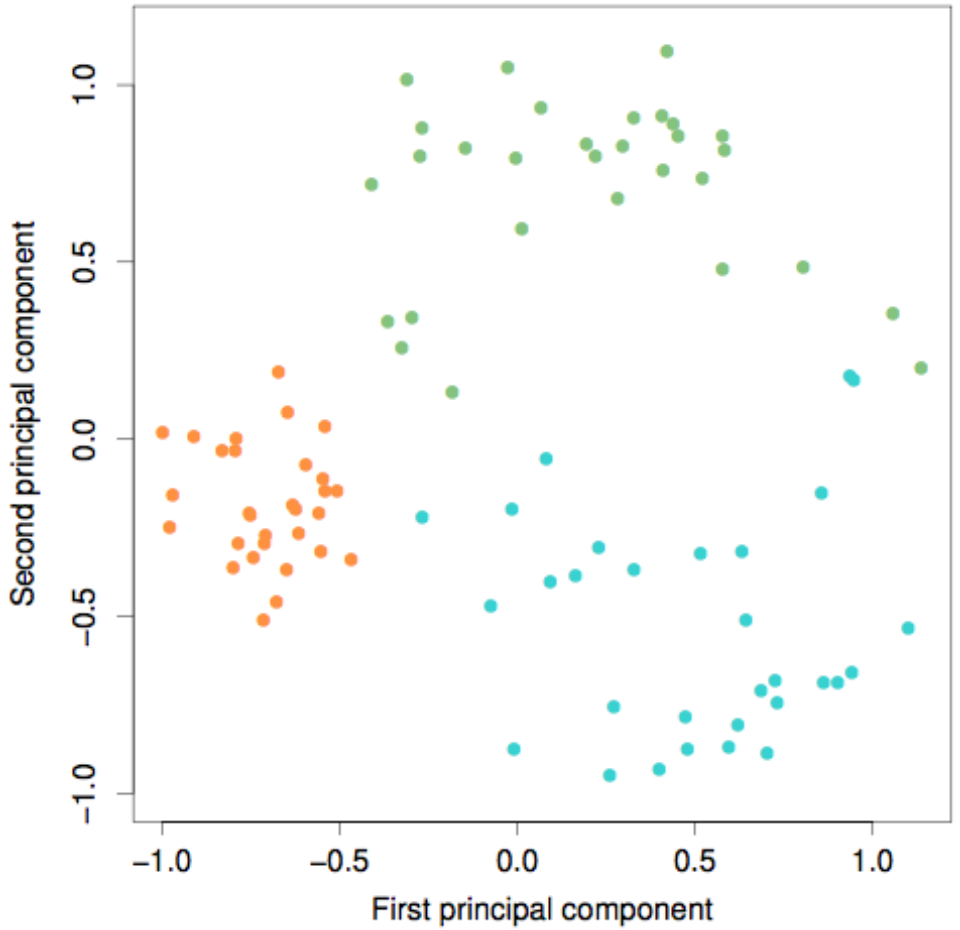
- The second principal component Z_2 is the normalized linear combination that has the maximal variance among all linear combinations that are *uncorrelated* with Z_1. (The blue dash line in page 11).

- The un-correlatedness is equivalent to the orthogonality between ϕ_1 and ϕ_2.

- Similarly,
  - The third principal component Z_3 has the maximum variance that are *uncorrelated* with both Z_1 and Z_2.
  - …
  - The coefficients in the linear combinations are the corresponding loadings.

Another Interpretation of Principal Components





- PCA finds hyperplanes closest to the observations!

Closest Hyperplanes

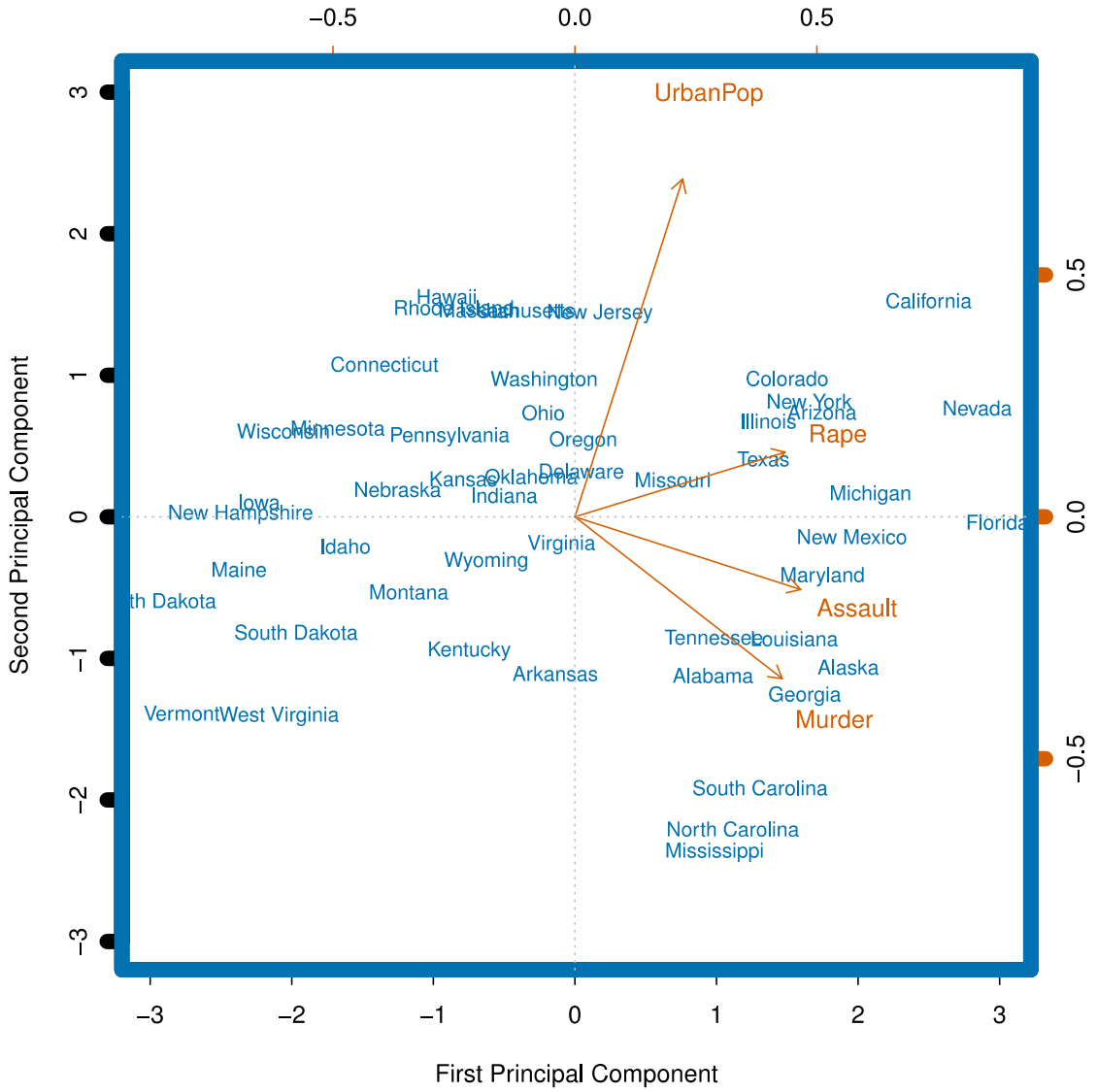
- The first M PC vectors give the following approximation.

$$x_{ij} \approx \sum_{m=1}^{M} z_{im}\phi_{jm}.$$

- Actually, $\hat{a}_{im} = z_{im}, \hat{b}_{jm} = \phi_{jm}$ does minimize the MSE.

$$\underset{\mathbf{A}\in\mathbb{R}^{n\times M},\mathbf{B}\in\mathbb{R}^{p\times M}}{\text{minimize}} \left\{ \sum_{j=1}^{p}\sum_{i=1}^{n} \left(x_{ij} - \sum_{m=1}^{M} a_{im}b_{jm} \right)^2 \right\}.$$

USAarrests: Biplot



Biplot: Details

- Biplot shows both the PC scores and loadings.
- PC scores: the blue state names
- PC loading vectors: the orange arrows (with axes on the top and right)

	PC1	PC2
Murder	0.5358995	-0.4181809
Assault	0.5831836	-0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

- For example, the loading for Rape on PC1 is 0.54, and PC2 is 0.17 (center of the word)

R Commands (Section 12.5)

PCA

```
> pr.out=prcomp(USArrests, scale=TRUE)
> names(pr.out)
[1] "sdev"     "rotation" "center"   "scale"    "x"
```

PC Loadings

```
> pr.out$rotation
             PC1    PC2    PC3    PC4
Murder   -0.536  0.418 -0.341  0.649
Assault  -0.583  0.188 -0.268 -0.743
UrbanPop -0.278 -0.873 -0.378  0.134
Rape     -0.543 -0.167  0.818  0.089
```

Principal components are only unique up to a sign change.

PC -> -PC; Loadings -> -Loadings

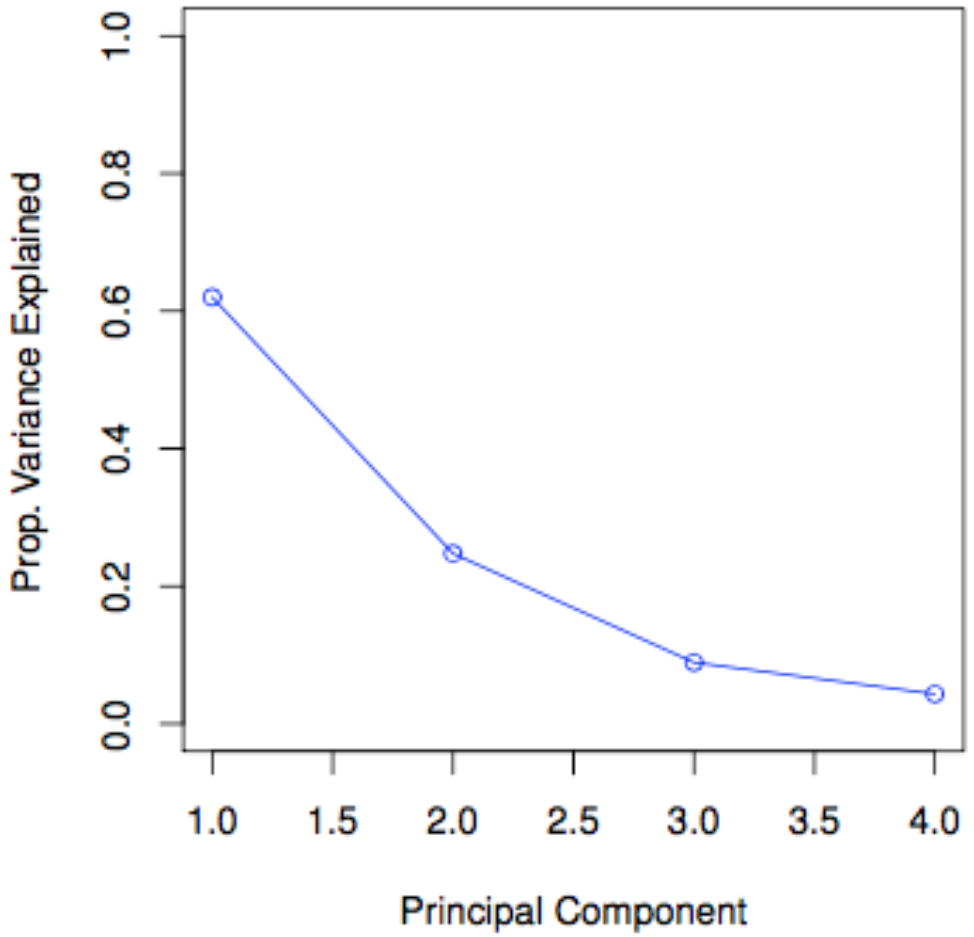
PC Scores

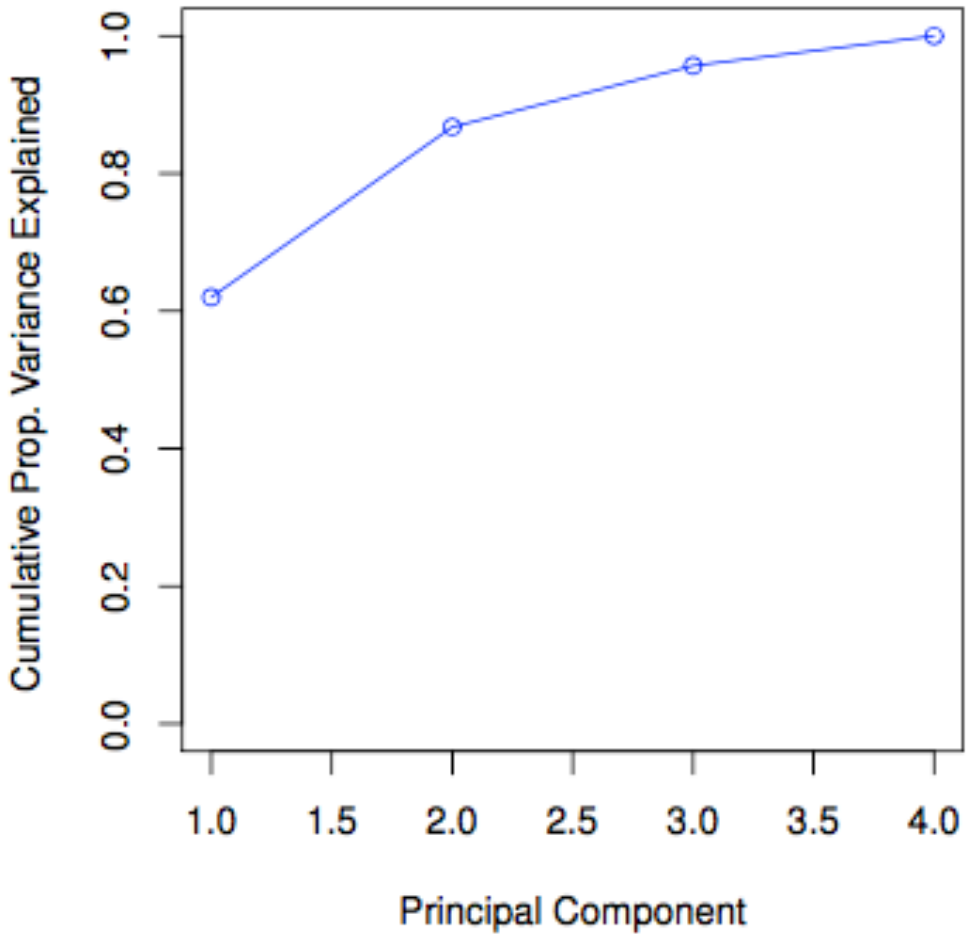
```
> dim(pr.out$x)
[1] 50  4
```

Biplot

```
> biplot(pr.out, scale=0)
```

Scree Plot to Determine # of PCs





Elbow Hunting!

R Commands (Section 12.5)

```
> pr.var <- pr.out$sdev^2
> pr.var
[1] 2.480 0.990 0.357 0.173
```

```
> pve <- pr.var / sum(pr.var)
> pve
[1] 0.6201 0.2474 0.0891 0.0434
```

```
> plot(pve, xlab = "Principal Component",
    ylab = "Proportion of Variance Explained", ylim = c(0, 1),
    type = "b")
> plot(cumsum(pve), xlab = "Principal Component",
    ylab = "Cumulative Proportion of Variance Explained",
    ylim = c(0, 1), type = "b")
```

Principal Components Regression

$$y_i = \beta_0 + \sum \beta_j x_{ij} + \epsilon_i$$

Instead, consider $\mathbf{u}_j$ the jth principal component of $\mathbf{x}$.

Then model

$$y_i = \beta'_0 + \sum_{j=1}^{p'} \beta'_j u_{ij} + \epsilon_i$$

for some $p' < p$.

Principal Components Regression

$$y_i = \beta'_0 + \sum_{j=1}^{p'} \beta'_j u_{ij} + \epsilon_i$$

Advantages:

- u_{ij} are uncorrelated – stability of estimates
- dimension reduction
- stable variable selection

Choosing p'

- Use enough PCs to caputure 90% of variation
- Maximize adjusted R^2
- Do variable selection (ie, not necessarily leading PCs)

Partial Least Squares

- In PCR, we regress Y on directions that best represent the predictors $X_1, \ldots, X_p$. These directions are chosen to be PCs $Z_1, \ldots, Z_M$ in an unsupervised way.

- In PLS, we hope to choose the directions using Y.
  - $Z_1 = \psi_{11}X_1 + \psi_{21}X_2 + \cdots + \psi_{p1}X_p$ where $\psi_{j1} \propto corr(Y, X_j)$
  - $Z_2 = \cdots$ un-correlated with Z_1

- While the supervised dimension reduction of PLS can reduce bias, it also has the potential to increase variance.

Netflix Problem / Recommender System: Missing Data and Matrix Completion

	Jerry Maguire	Oceans	Road to Perdition	A Fortunate Man	Catch Me If You Can	Driving Miss Daisy	The Two Popes	The Laundromat	Code 8	The Social Network	⋯
Customer 1	●	●	●	●	4	●	●	●	●	●	⋯
Customer 2	●	●	3	●	●	●	3	●	●	3	⋯
Customer 3	●	2	●	4	●	●	●	●	2	●	⋯
Customer 4	3	●	●	●	●	●	●	●	●	●	⋯
Customer 5	5	1	●	●	4	●	●	●	●	●	⋯
Customer 6	●	●	●	●	●	2	4	●	●	●	⋯
Customer 7	●	●	5	●	●	●	●	3	●	●	⋯
Customer 8	●	●	●	●	●	●	●	●	●	●	⋯
Customer 9	3	●	●	●	5	●	●	1	●	●	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

- Can we impute the missing values / complete the data matrix with only a small proportion of observations?

Principal Components with Missing Values

- PCA finds the closest M-dim hyperplane

$$\underset{\mathbf{A}\in\mathbb{R}^{n\times M},\mathbf{B}\in\mathbb{R}^{p\times M}}{\text{minimize}} \left\{ \sum_{j=1}^{p}\sum_{i=1}^{n}\left(x_{ij}-\sum_{m=1}^{M}a_{im}b_{jm}\right)^2 \right\}.$$

- With missing values, we solve

$$\underset{\mathbf{A}\in\mathbb{R}^{n\times M},\mathbf{B}\in\mathbb{R}^{p\times M}}{\text{minimize}} \left\{ \sum_{(i,j)\in\mathcal{O}}\left(x_{ij}-\sum_{m=1}^{M}a_{im}b_{jm}\right)^2 \right\},$$

- We can estimate a missing observation by $\hat{x}_{ij}=\sum_{m=1}^{M}\hat{a}_{im}\hat{b}_{jm}$

Usage of PCA

- Data visualization
- EDA
- Clustering
- Dimension reduction for regression / classification
- Missing data imputation