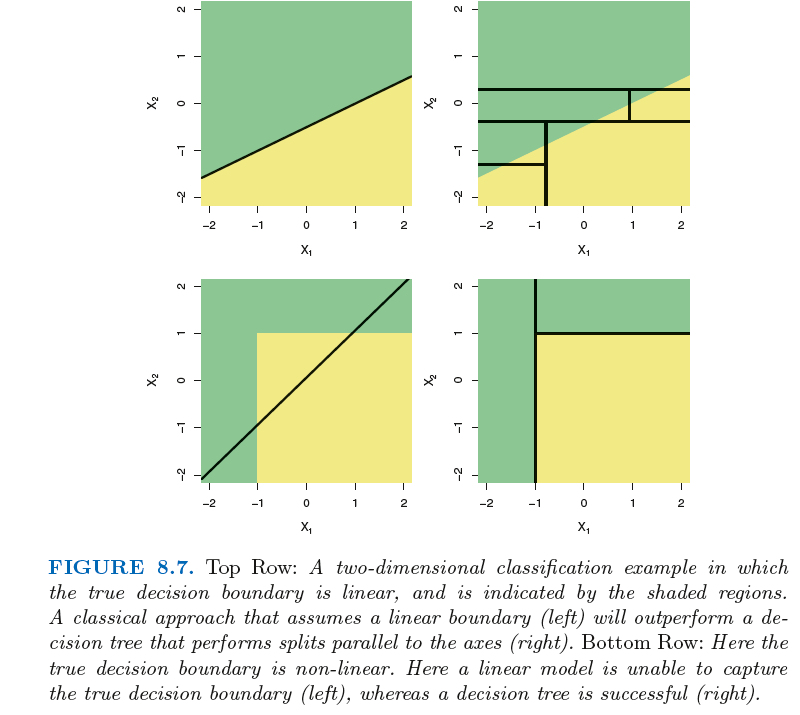
1. Decision tree.
2. Easy to explain
3. Without need to dummy transformation
4. However, not accurate as others



1. Bagging, randomForst, boosting
2. Bagging, sampling with replace[using B bootstrapped training set and average the predictions]
3. bagging improves prediction accuracy at the expense of interpretability
4. Bagging important variables. Gini index that decreased much

Variable importance: [After each tree is grown, the values of a given predictor are randomly permuted in the out-of-bag sample (the one third of unique observations that are not part of the bootstrap sample) and the prediction error of the tree on the modified OOB sample is compared with the prediction error of the tree on the untouched OOB sample.

This process is repeated for all input variables, and averaged over all trees. Finally, variables are given scores proportional to the overall decrease in accuracy that their permutation induced.]

1. rF
2. random forest is a special case of bagging, if the mtry you choose is equal to the total number of predictors, then there is no difference between the two models.
3. The number of observations used to build each tree in both bagging and RF is just the original data size. Just using bootstrapping to repeat sampling with replacement.
4. But when building these decision trees, each time a split in a tree is considered, *a random sample of m predictors* is chosen as split candidates from the full set of *p* predictors.The split is allowed to use only one of those *m* predictors. 🡪overcome overfitting compared with DT or bagging
5. Boosting
6. trees are grown *sequentially*
7. This highlights one difference between boosting and random forests: in boosting, because the growth of a particular tree takes into account the other trees that have already been grown, smaller trees are typically sufficient. Using smaller trees can aid in interpretability as well; for instance, using stumps leads to an additive model
8. Non-parametric approaches

by avoiding the

assumption of a particular functional form for *f*, they have the potential

to accurately fit a wider range of possible shapes for *f*.

But non-parametric approaches do suffer from a major

disadvantage: since they do not reduce the problem of estimating *f* to a

small number of parameters, a very large number of observations (far more

than is typically needed for a parametric approach) is required in order to

obtain an accurate estimate for *f*. 🡪overfitting

6

Trade-off between prediction accuracy and interpretability

Two goals: inference / prediction

In this setting, we might expect that it

will be best to use the most flexible model available. Surprisingly, this is

not always the case! We will often obtain more accurate predictions using

a less flexible method. This phenomenon, which may seem counterintuitive

at first glance, has to do with the potential for overfitting in highly flexible

methods.

1. *bias-variance trade-off*.
2. Classification setting
3. Bayes decision boundary (The Bayes classifier produces the lowest possible test error rate, calle the *Bayes error rate*) (need know the conditional distribution of *Y* given *X*)
4. K-Nearest Neighbors (Suppose that we choose *K* = 3. Then

KNN will first identify the three observations that are closest to the cross.

This neighborhood is shown as a circle. It consists of two blue points and

one orange point, resulting in estimated probabilities of 2*/*3 for the blue

class and 1*/*3 for the orange class. Hence KNN will predict that the black cross belongs to the blue class.)

1. Linear regression

*3.3.3 Potential Problems*

When we fit a linear regression model to a particular data set, many problems

may occur. Most common among these are the following:

1. *Non-linearity of the response-predictor relationships.*

2. *Correlation of error terms.[ Plots of residuals from simulated time series data sets generated*

*with differing levels of correlation ρ between error terms for adjacent time points.]*

3. *Non-constant variance of error terms.[* the magnitude of the residuals tends to increase with the fittedvalues. When faced with this problem, one possible solution is to transform

the response *Y* using a concave function such as log *Y* or*√Y* . Such

a transformation results in a greater amount of shrinkage of the larger responses,

leading to a reduction in heteroscedasticity.*]*

4. *Outliers.[ studentized*

*residuals*, computed by dividing each residual *ei* by its estimated standard

studentized

error. Observations whose studentized residuals are greater than 3 in abso- residual

lute value are possible outliers.*]*

5. *High-leverage points.[* In order to quantify an observation’s leverage, we compute the *leverage*

*statistic*.*]*

6. *Collinearity.[* In other words, since

limit and rating tend to increase or decrease together, it can be difficult to

determine how each one separately is associated with the response, balance*]*

*[Contour plots for the RSS values]*

*[*In other words, the importance

of the limit variable has been masked due to the presence of collinearity*]*

*[*it is possible for collinearity

to exist between three or more variables even if no pair of variables

has a particularly high correlation. We call this situation *multicollinearity*.multi-

Instead of inspecting the correlation matrix, a better way to assess multi- collinearity

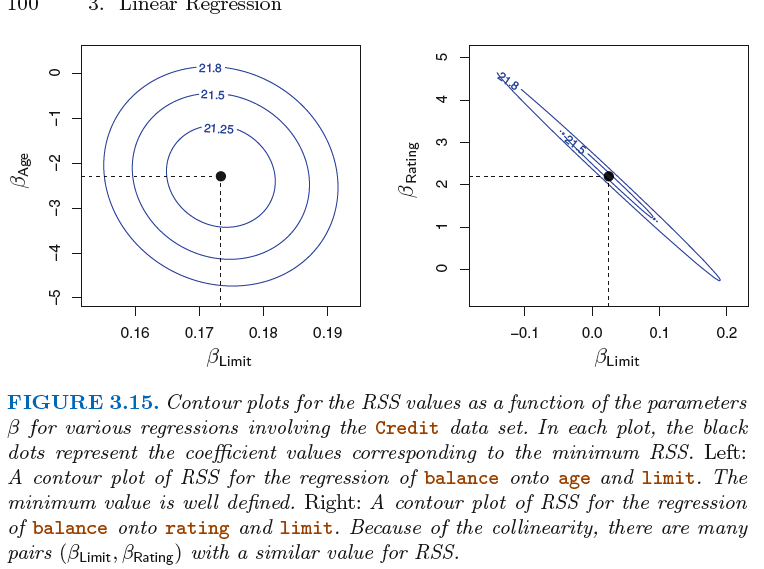
collinearity is to compute the *variance inflation factor* (VIF). The VIF is variance inflation factor

the ratio of the variance of ˆ*βj* when fitting the full model divided by the

variance of ˆ*βj* if fit on its own. The smallest possible value for VIF is 1,

which indicates the complete absence of collinearity.*]*

*[*where *R*2*Xj |X−j* is the *R*2 from a regression of *Xj* onto all of the otherpredictors.*]*



*In the case of collinearity, even a small change of data can cause an estimated coefficients which can lead to a small RSS. So that will involve an issue, least square will lead to a great deal of uncertainty of the coefficients estimation.*

1)

If the residual plot indicates that there are non-linear associations in the

data, then a simple approach is to use non-linear transformations of the

predictors, such as log*X*,

*√X*, and *X*2, in the regression model.

Linear regression & KNN regression

In fact, the increase in

dimension has only caused a small deterioration in the linear regression test

set MSE, but it has caused more than a ten-fold increase in the MSE for

KNN. This decrease in performance as the dimension increases is a common

problem for KNN, and results from the fact that in higher dimensions

there is effectively a reduction in sample size. In this data set there are

100 training observations; when *p* = 1, this provides enough information to

accurately estimate *f*(*X*). However, spreading 100 observations over *p* = 20

dimensions results in a phenomenon in which a given observation has no

*nearby neighbors*—this is the so-called *curse of dimensionality*.

As a general rule,

parametric methods will tend to outperform non-parametric approaches

when there is a small number of observations per predictor.

ClasSification

We are interested in predicting whether an

individual will default on his or her credit card payment, on the basis of

annual income and monthly credit card balance

**logistic regression**

Rather than modeling this response *Y*

directly, logistic regression models the *probability* that *Y* belongs to a particular

category.

Maximum likelihood is a very general approach that is used to fit many

of the non-linear models that we examine throughout this book. In the

linear regression setting, the least squares approach is in fact a special case

of maximum likelihood.

This simple example illustrates the dangers and subtleties associated

with performing regressions involving only a single predictor when other

predictors may also be relevant.-confounding

**linear discriminant analysis,** is popular for multiple-class classification

vs lr[LDA assumes that the observations within each

class are drawn from a multivariate Gaussian distribution with a classspecific

mean vector and a covariance matrix that is common to all *K*

classes.]

* When the classes are well-separated, the parameter estimates for the

logistic regression model are surprisingly unstable. Linear discriminant

analysis does not suffer from this problem.

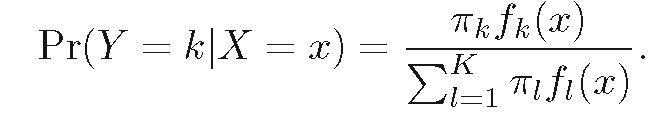
*•* If *n* is small and the distribution of the predictors *X* is approximately

normal in each of the classes, the linear discriminant model is again

more stable than the logistic regression model.

*•* As mentioned in Section 4.3.5, linear discriminant analysis is popular

when we have more than two response classes. Based on bayese’s theory

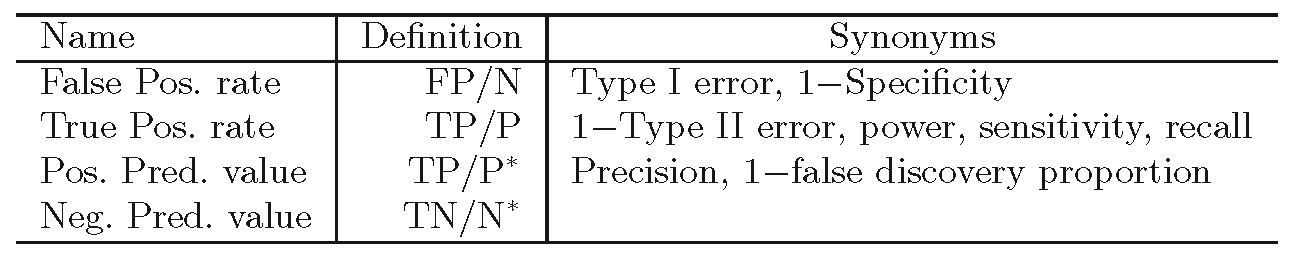


However, if we are concerned about incorrectly predicting

the default status for individuals who default, then we can consider

lowering this threshold. For instance, we might label any customer with a

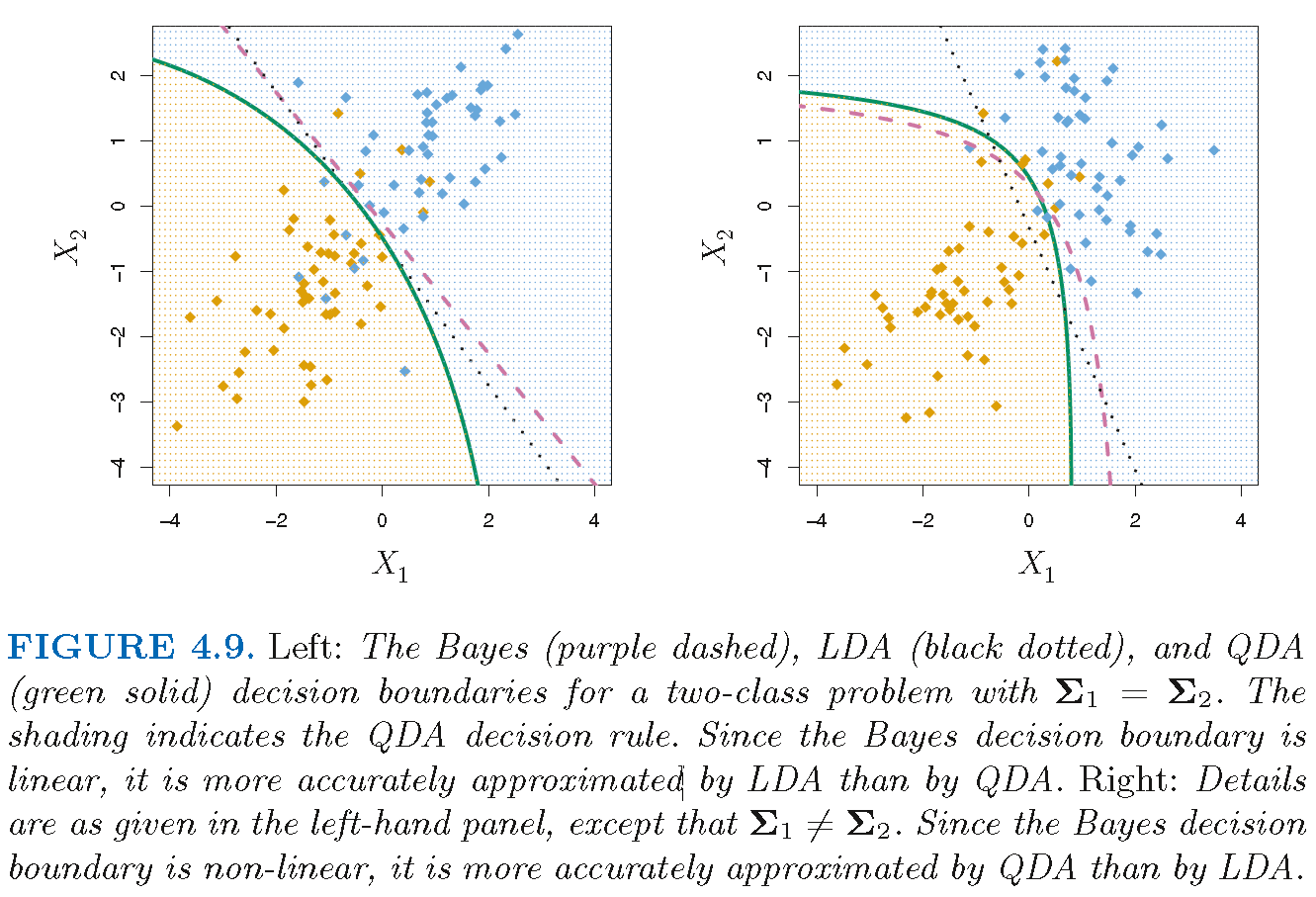
posterior probability of default above 20% to the *default* class.



**Quadratic Discriminant Analysis**

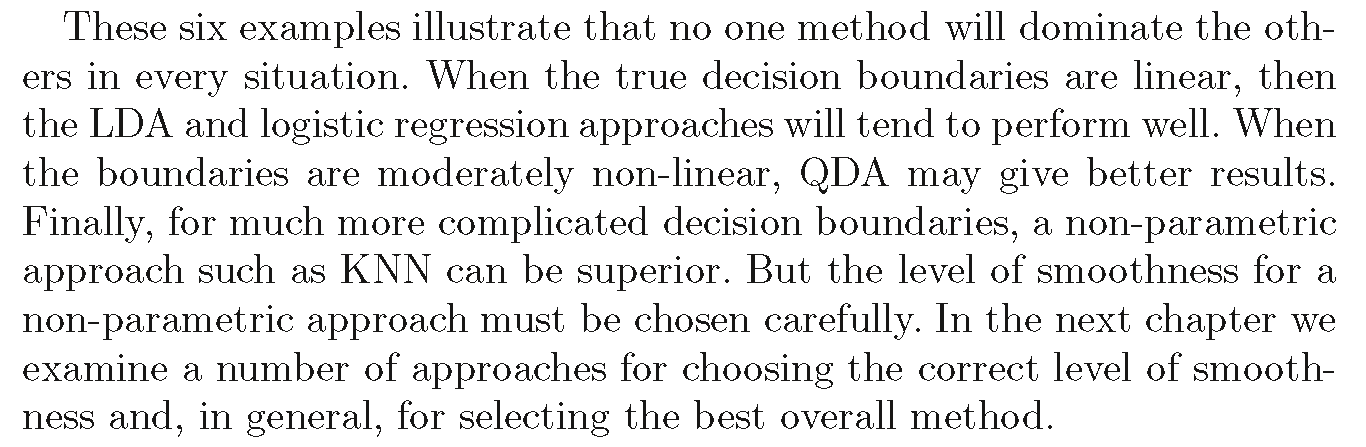
However, unlike LDA, QDA assumes that each class has

its own covariance matrix.

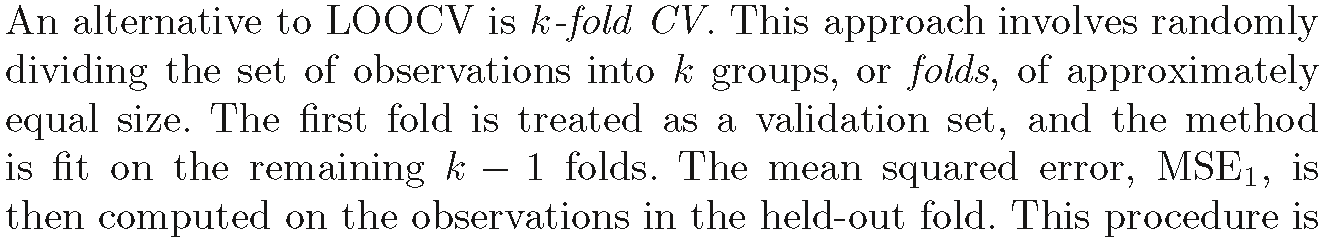


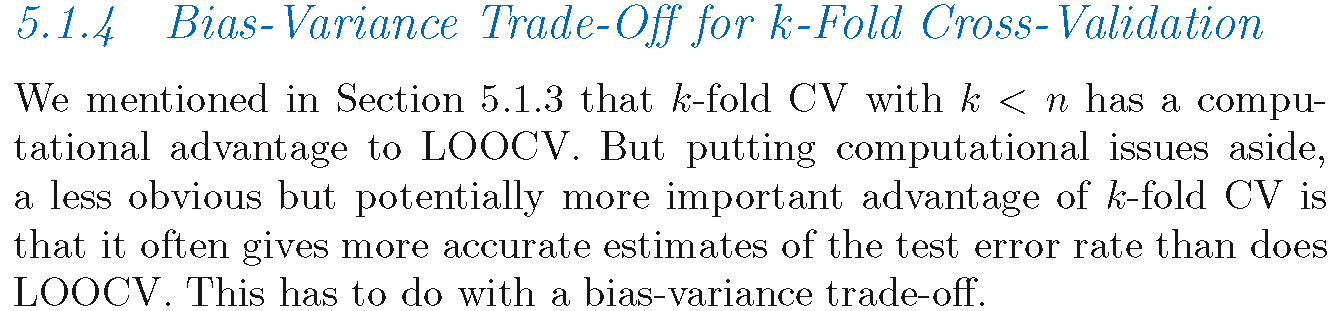
*Comparison of classification methods*

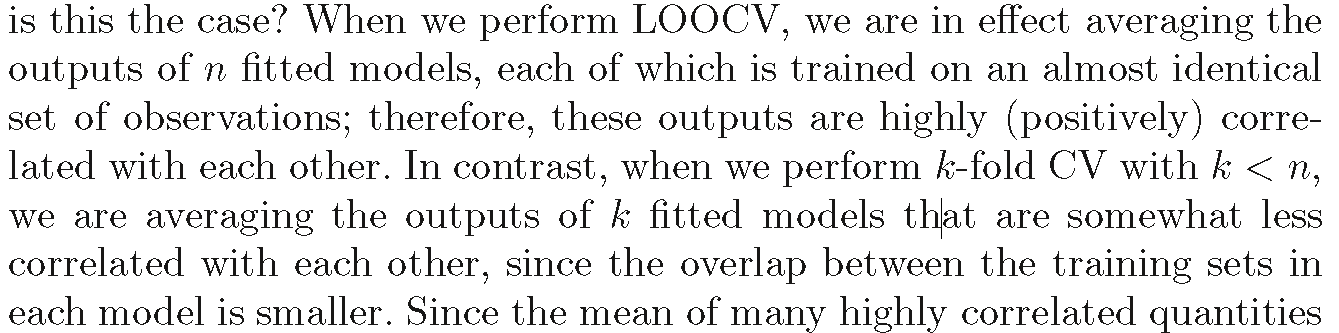
*Logistic regression LDA QDA KNN*

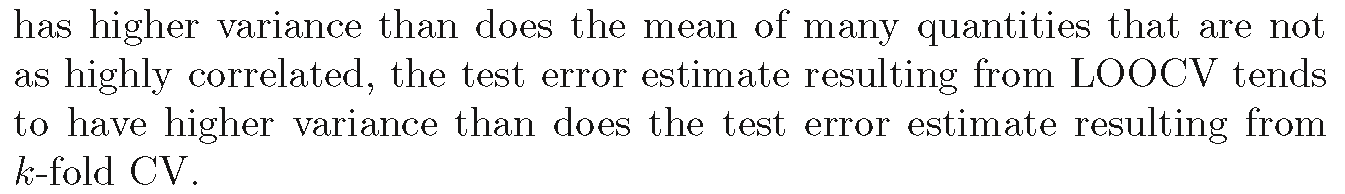


*Resampling method*



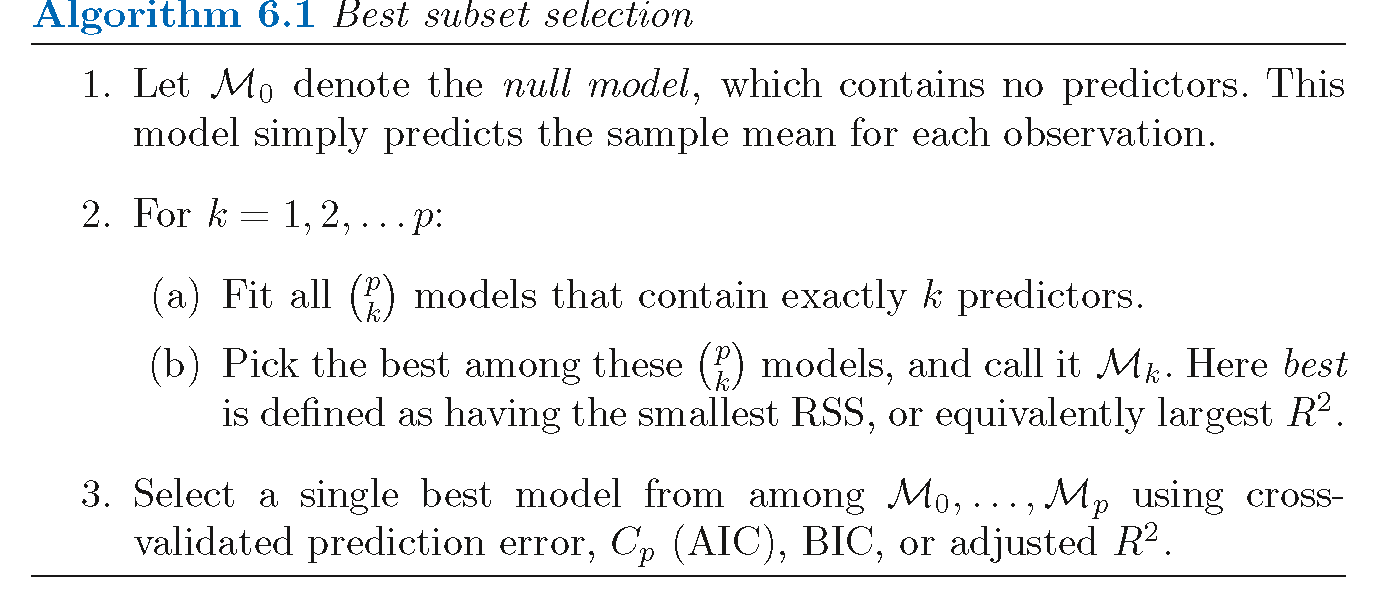


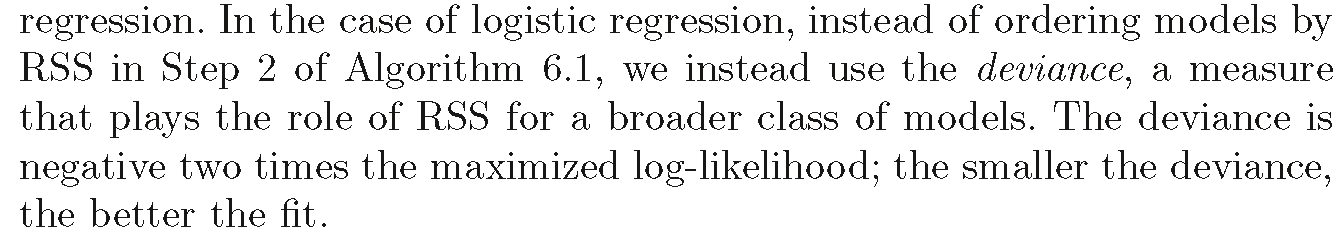




Linear Model Selection

and Regularization

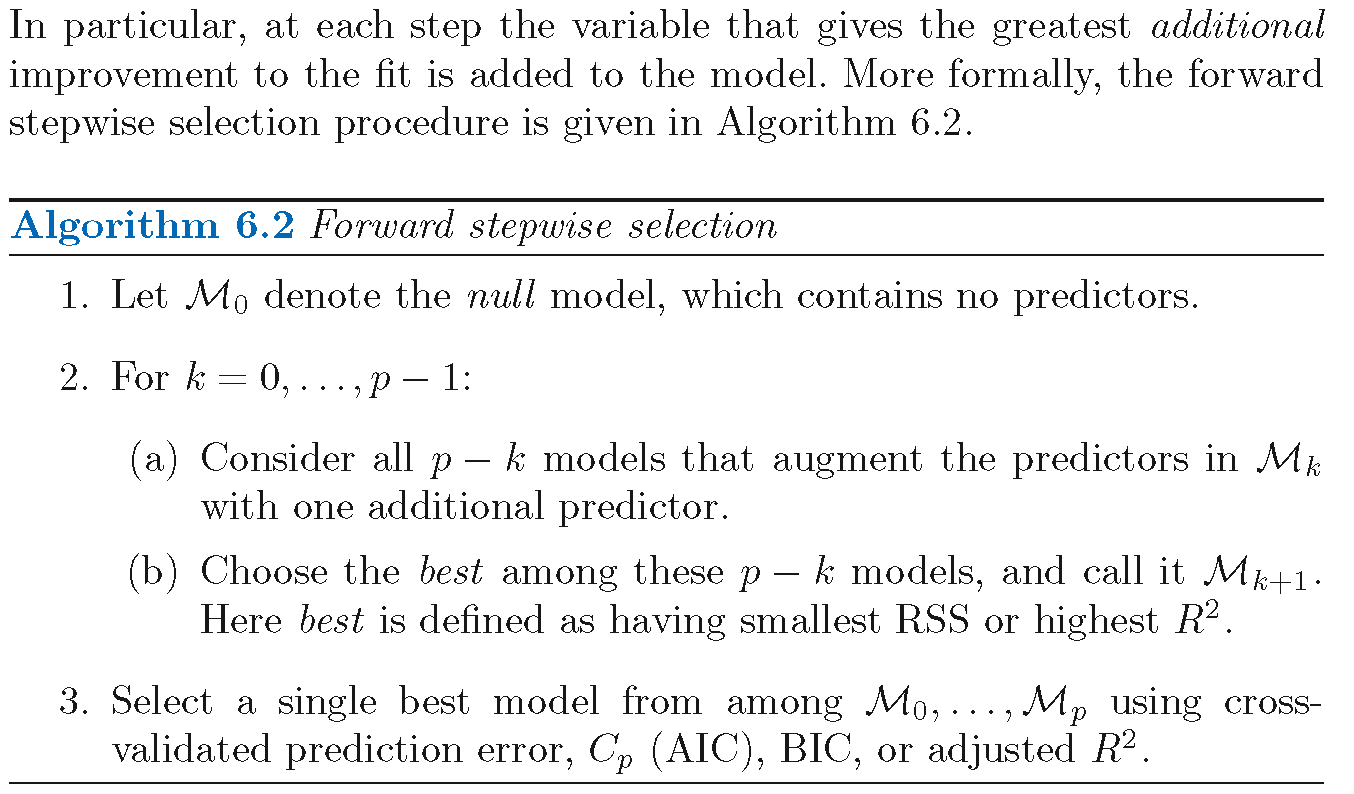


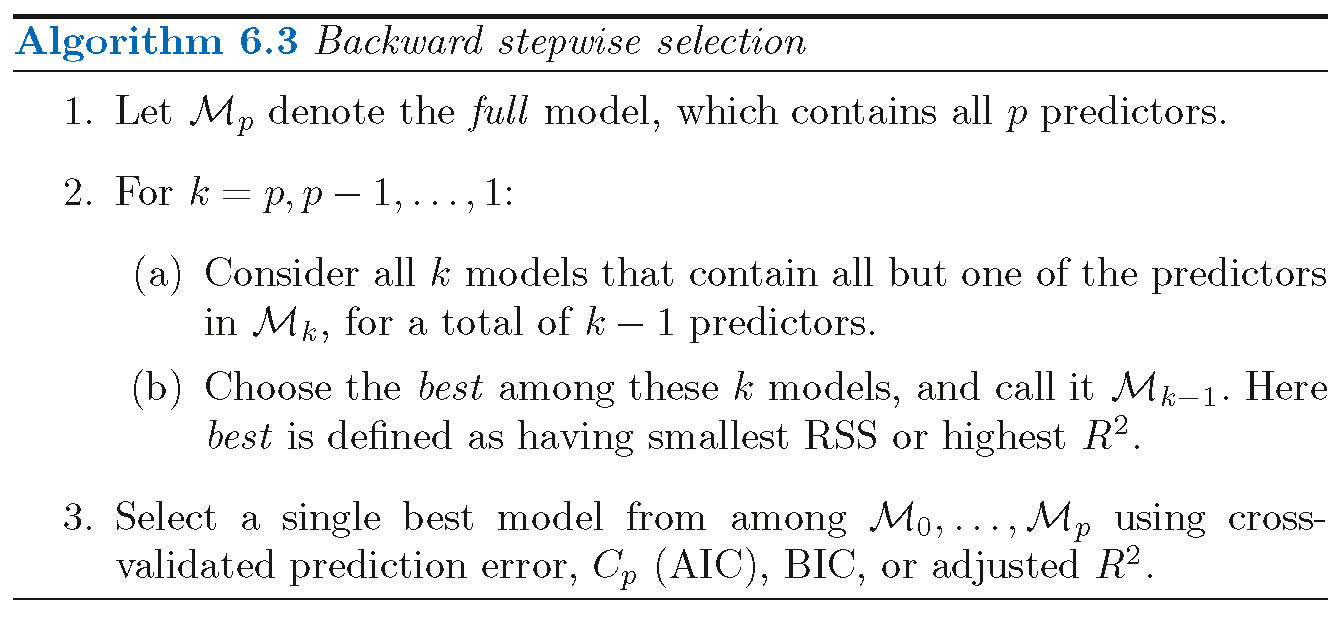


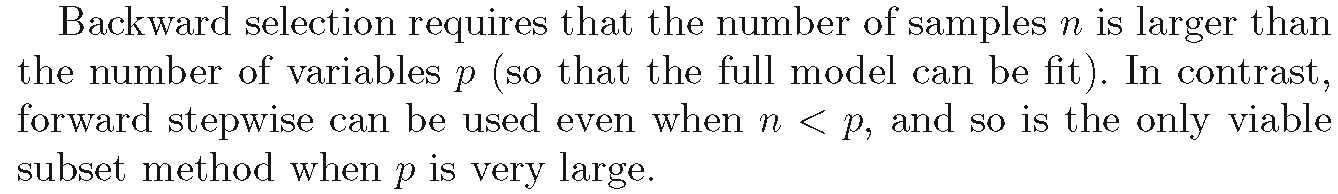
computationally infeasible for values of large *p,* We present computationally

efficient alternatives to best subset selection next.

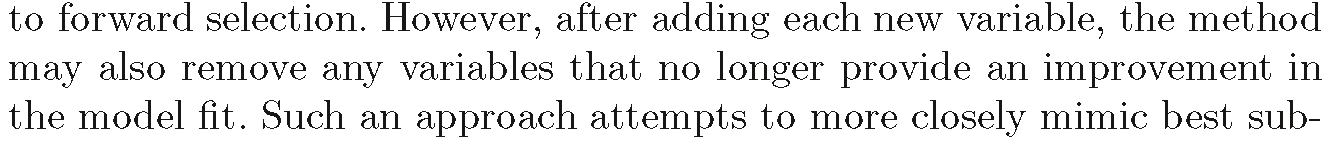
*stepwise* methods



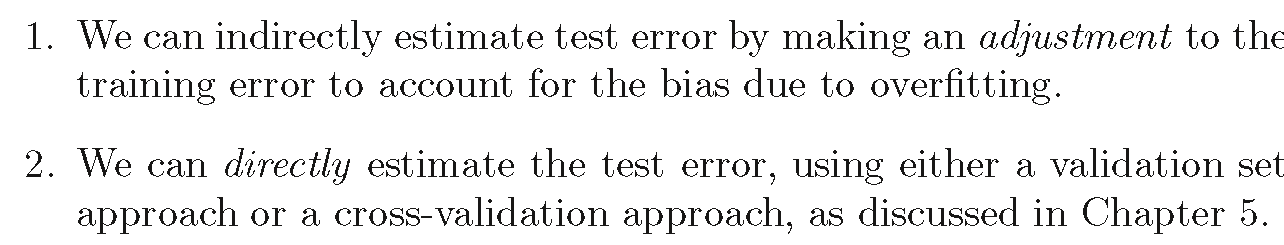


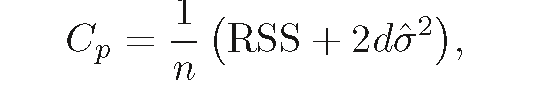


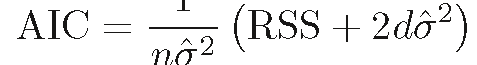
Hybrid Approaches

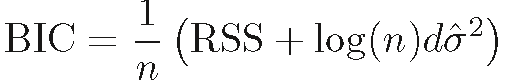


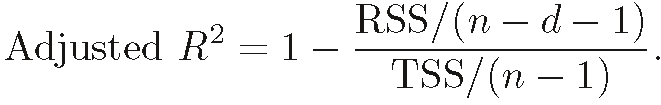
*6.1.3 Choosing the Optimal Model*

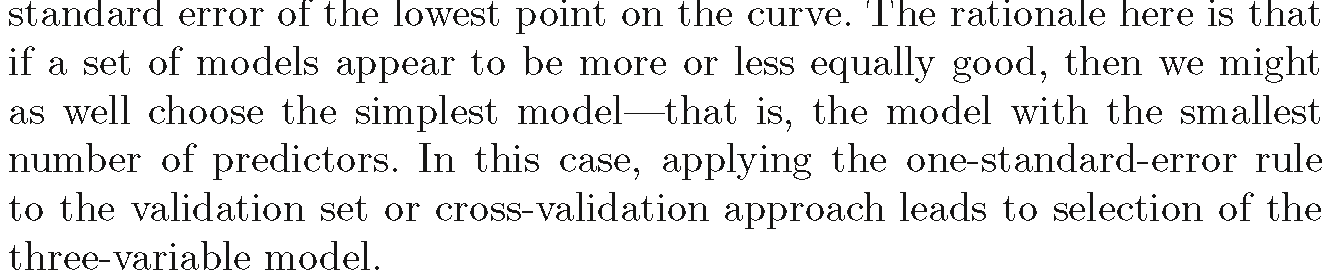






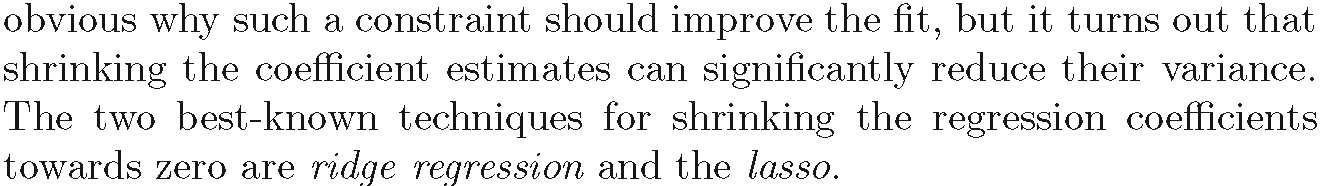


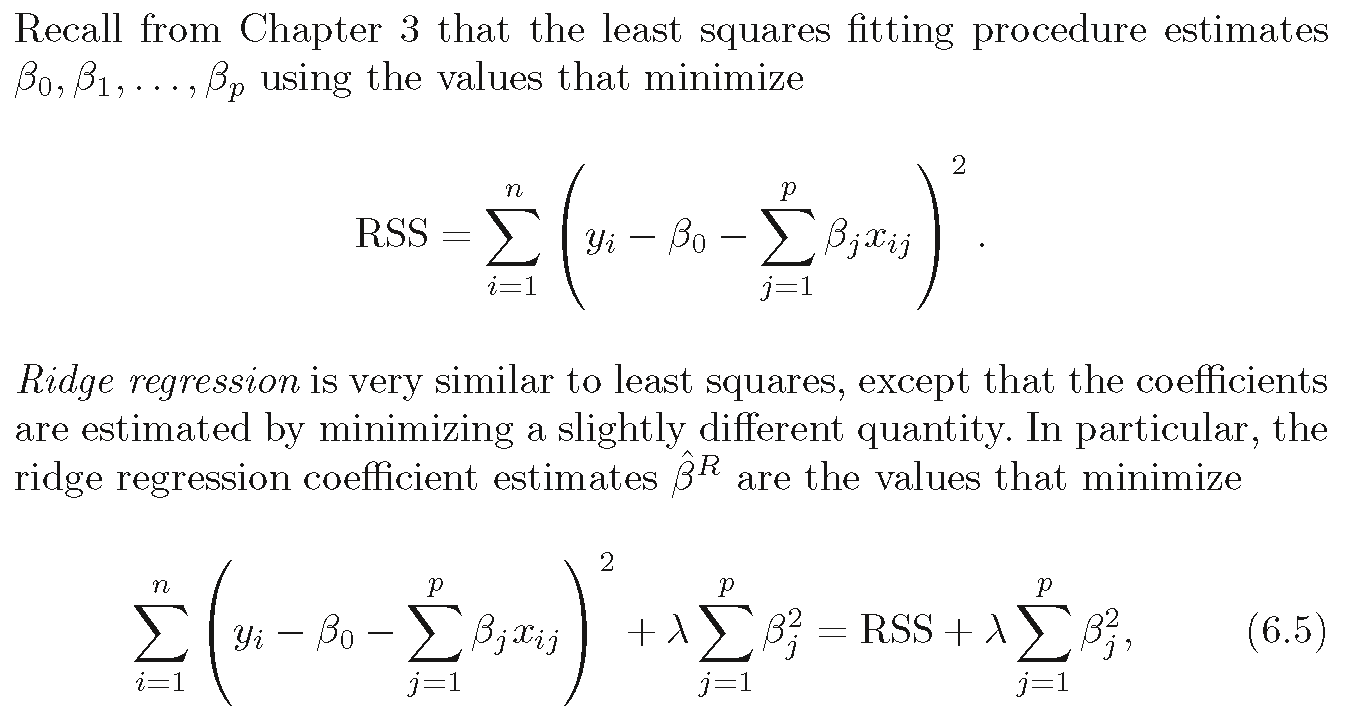


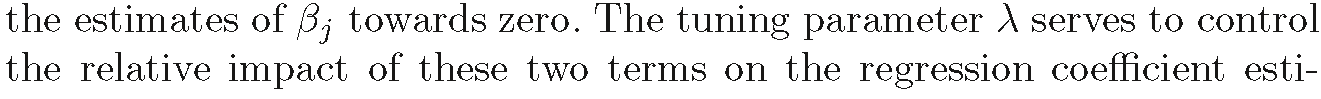


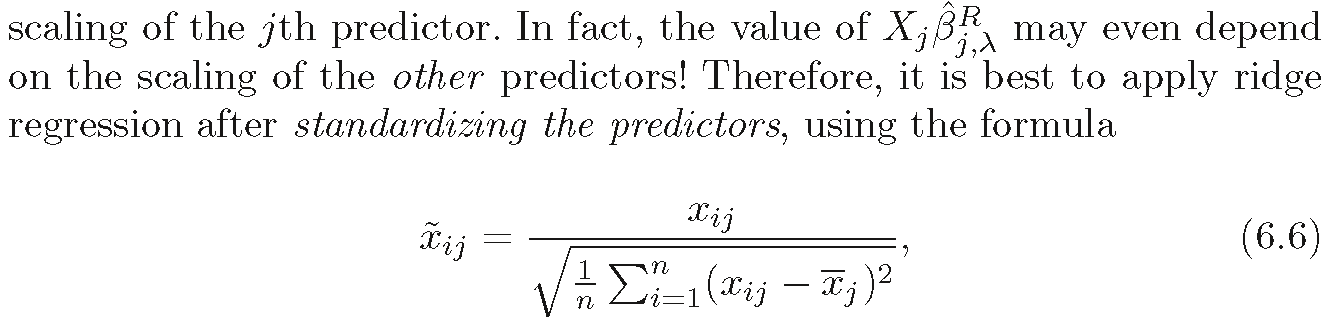
6.2 Shrinkage Methods





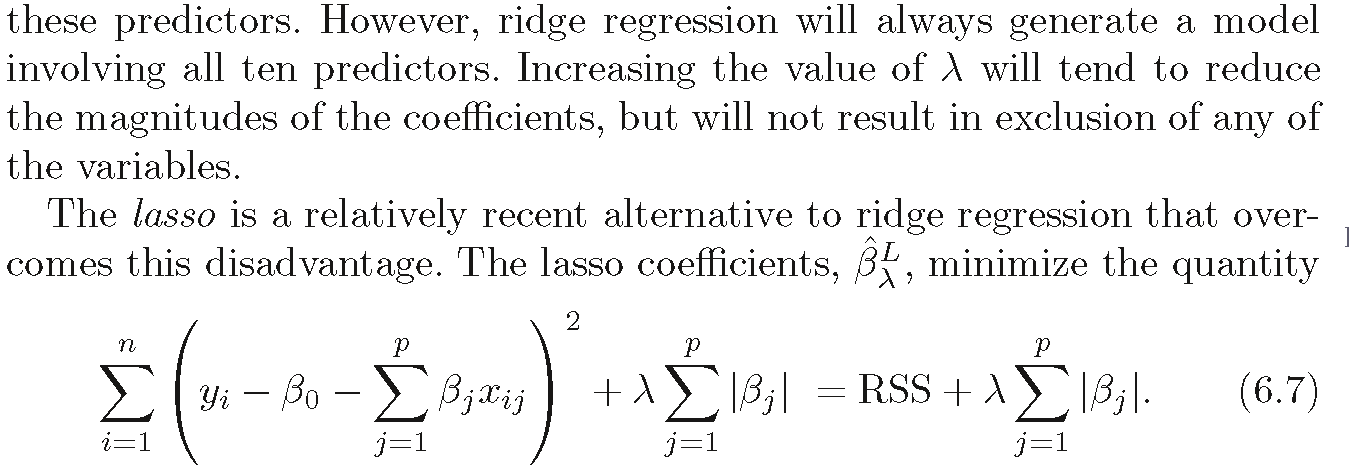


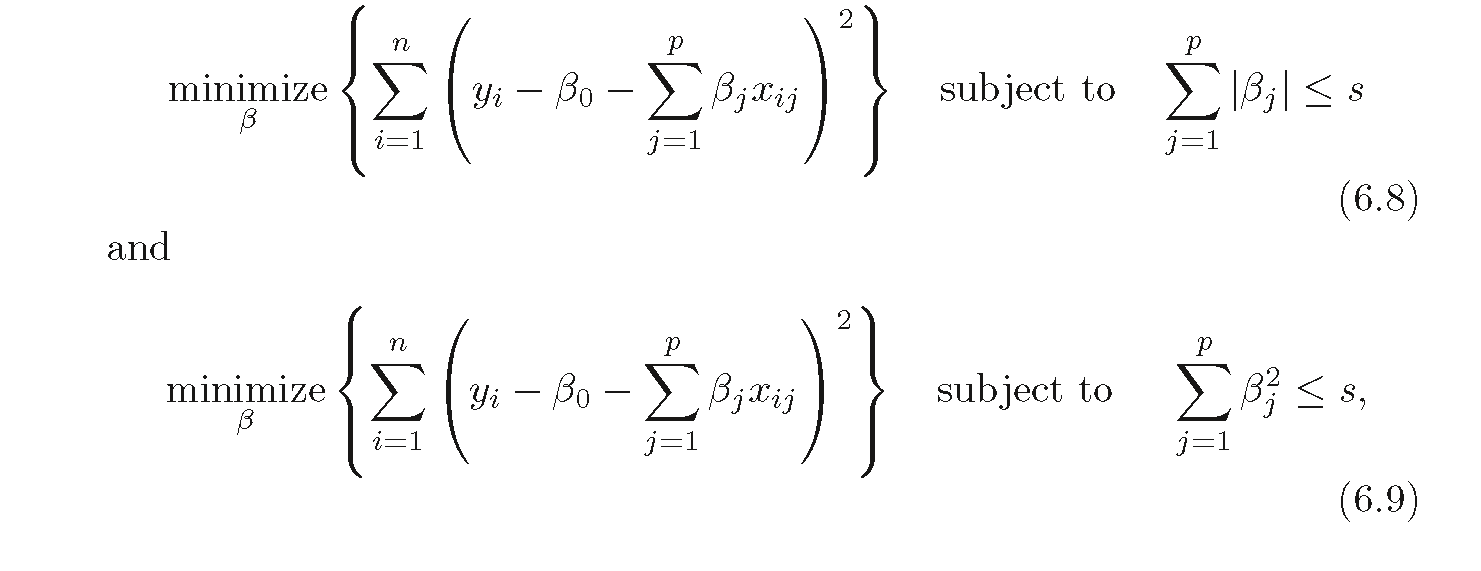


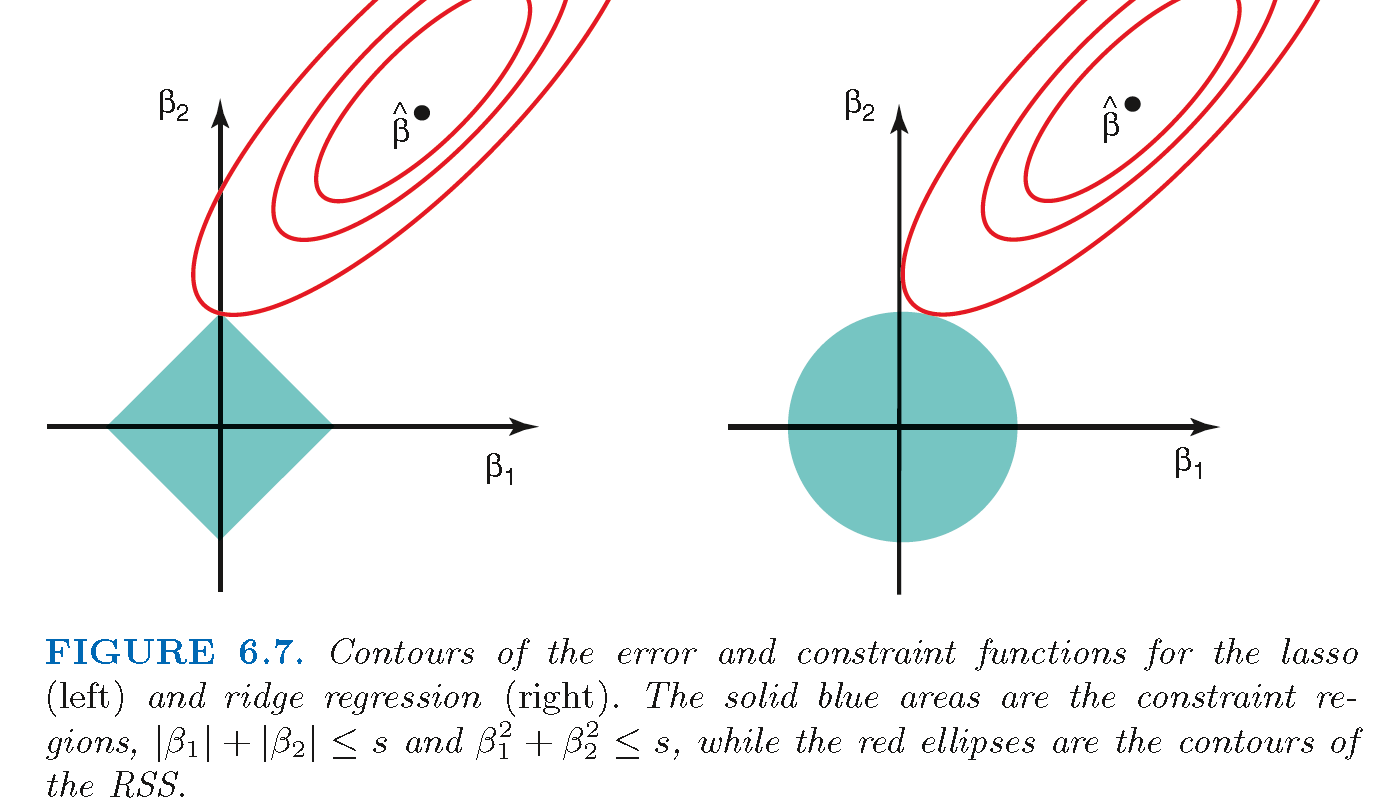


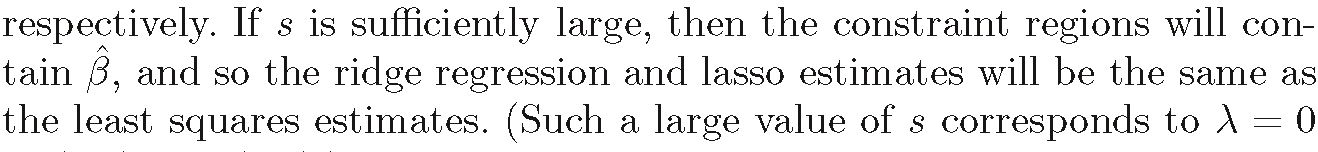


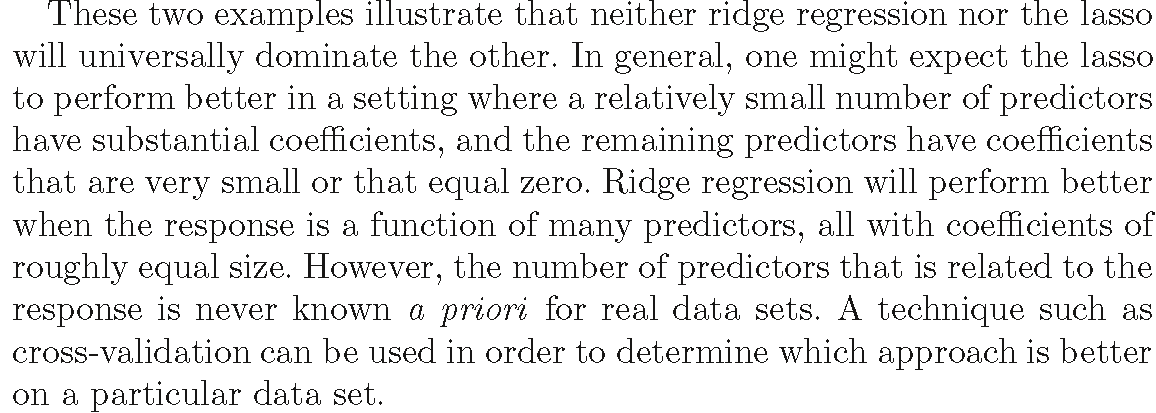


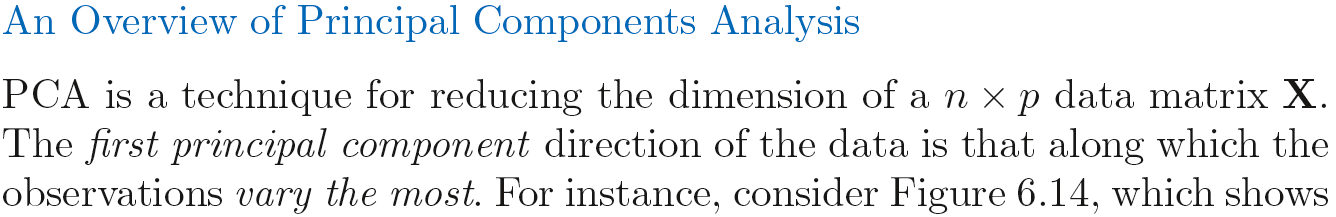


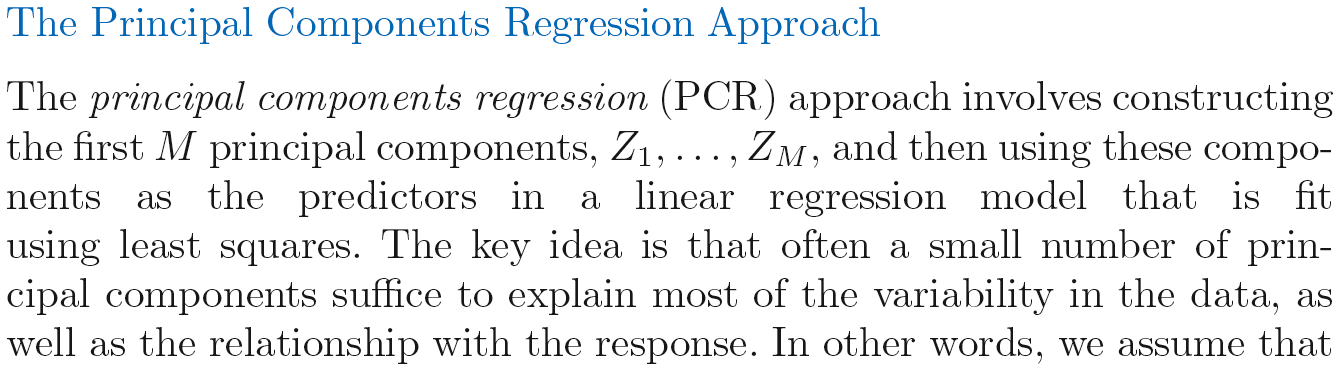


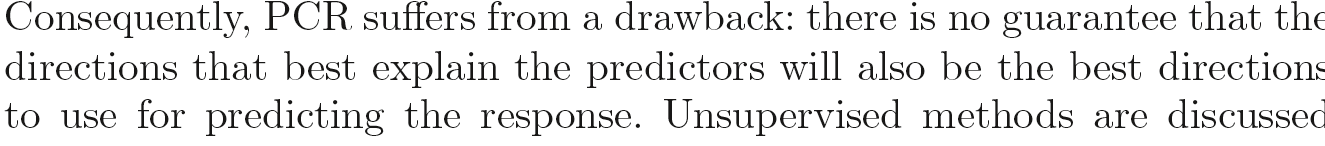


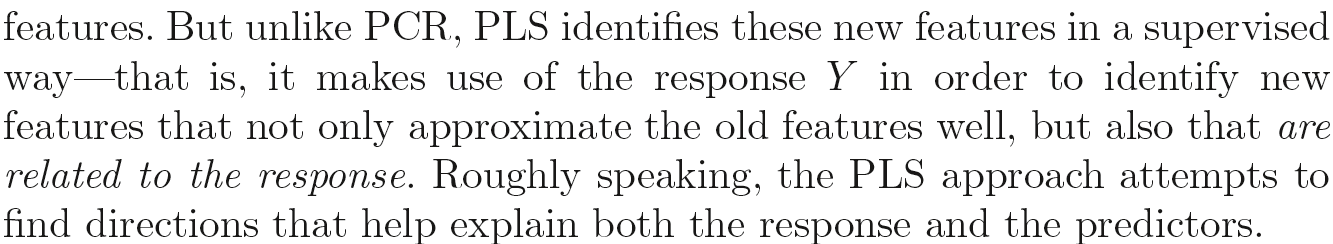


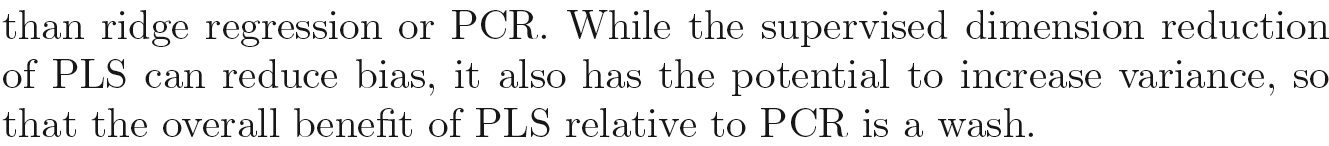








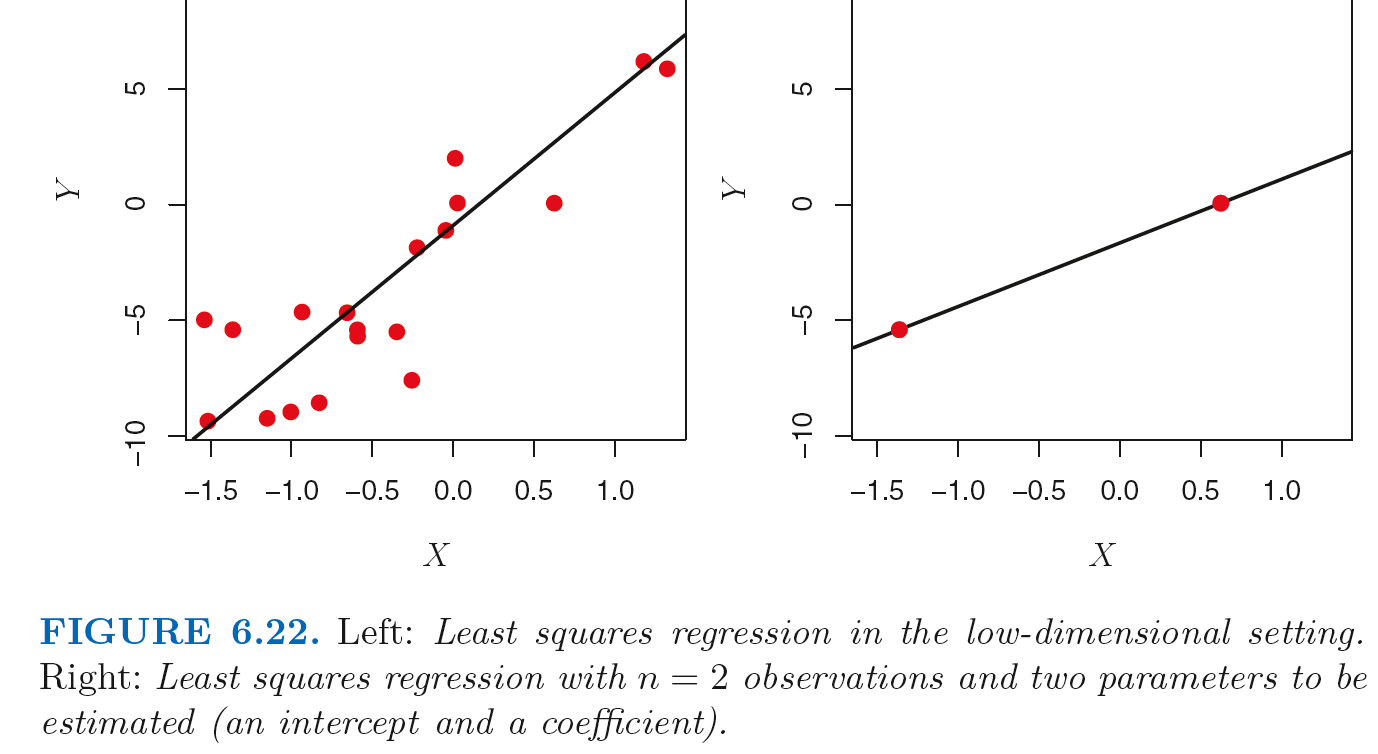


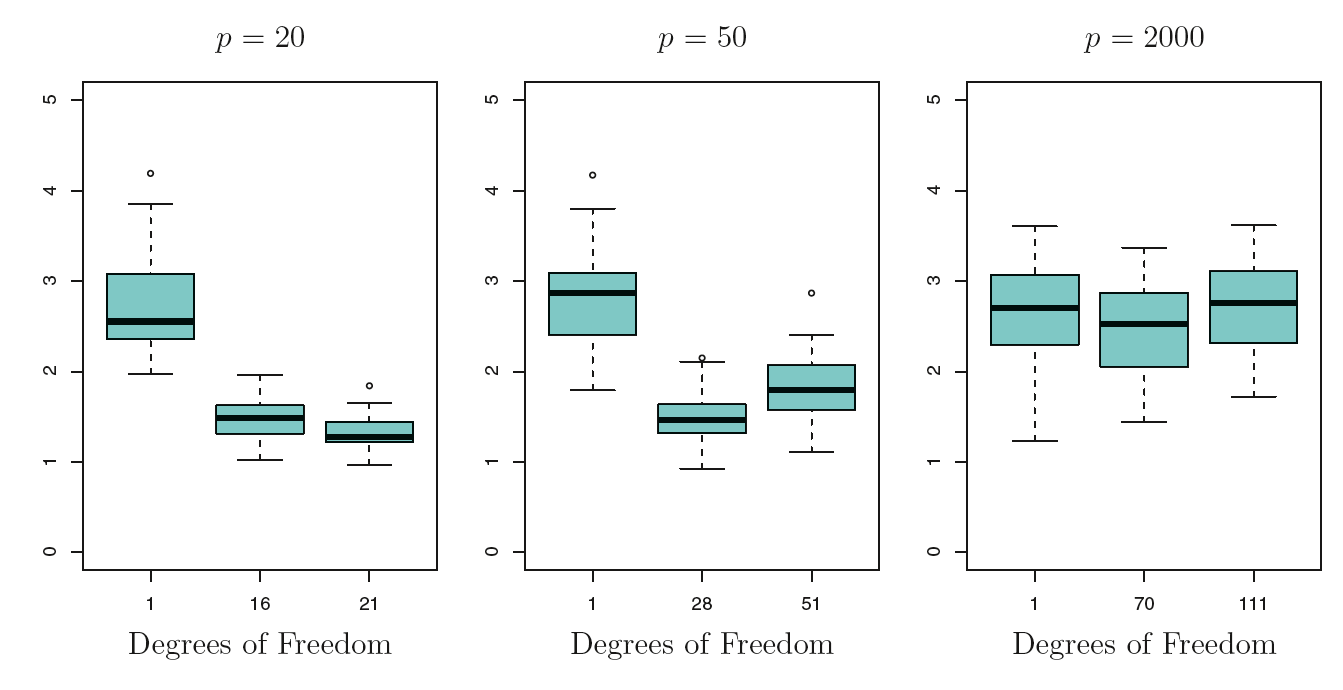


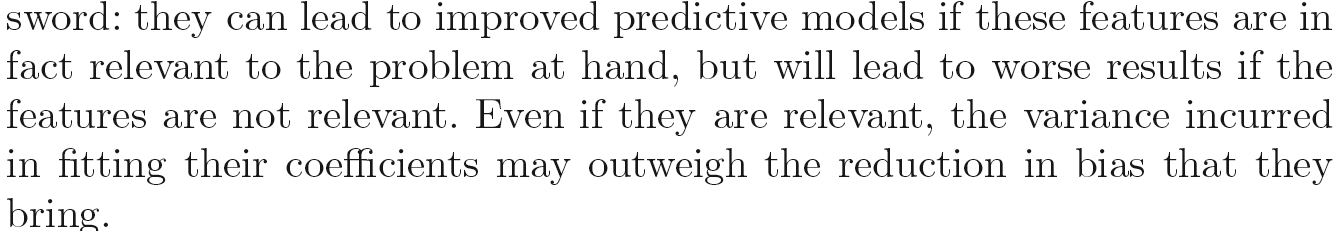
*PLS-partial least square*

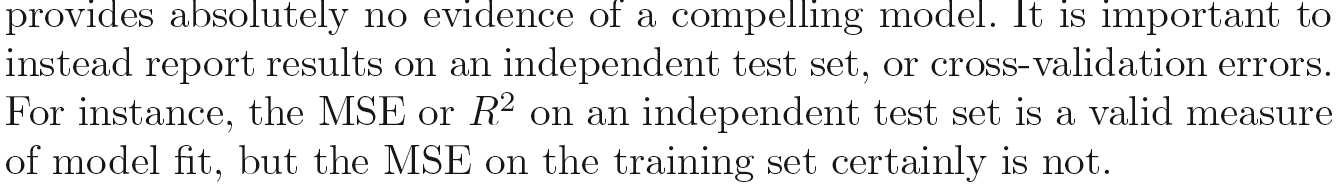
6.4 Considerations in High Dimensions

*6.4.1 High-Dimensional Data*



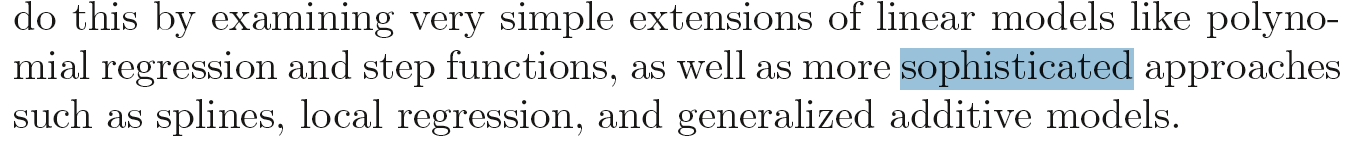


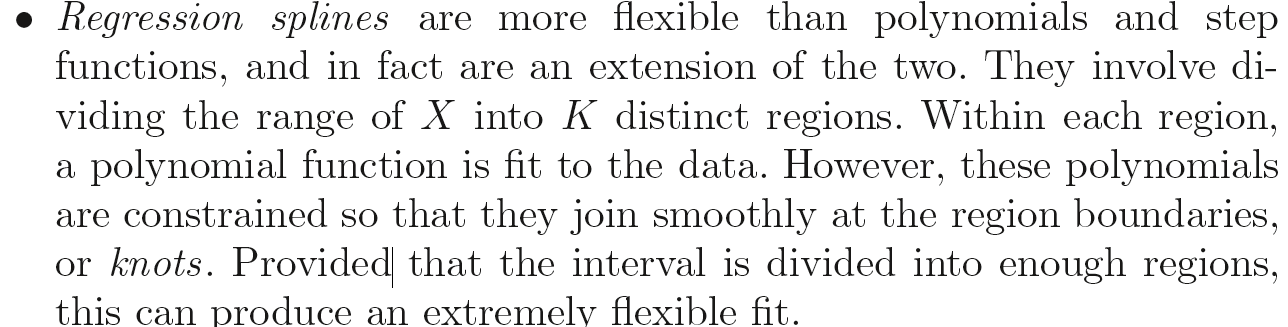


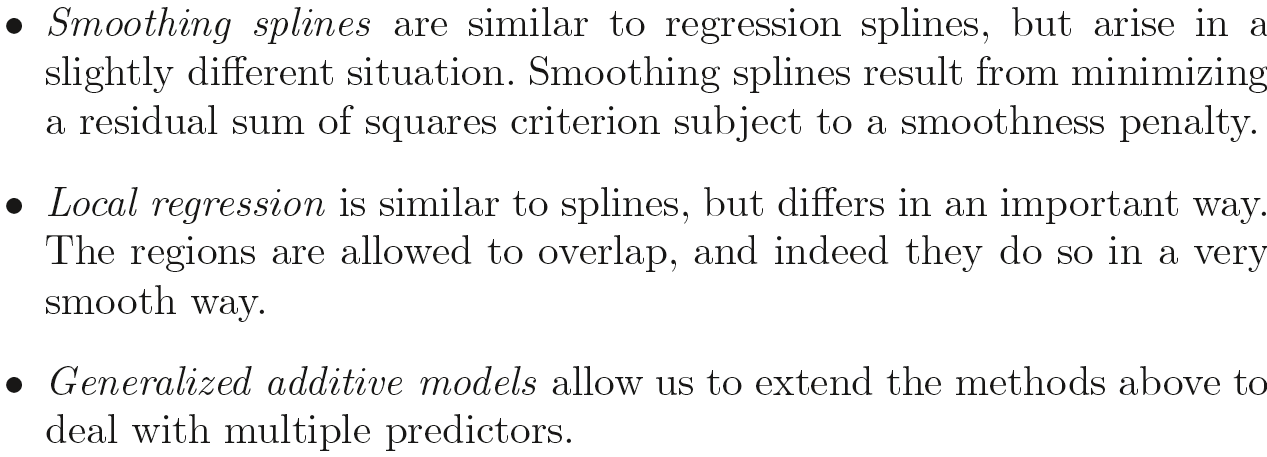


7

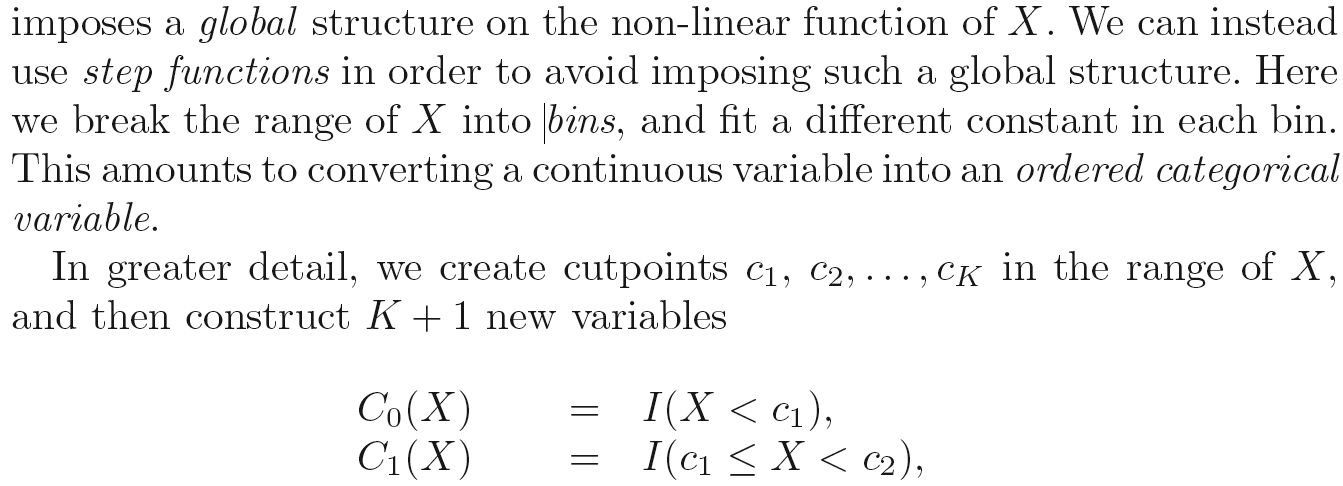
Moving Beyond Linearity





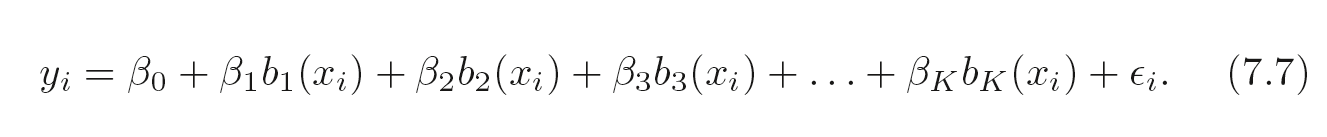


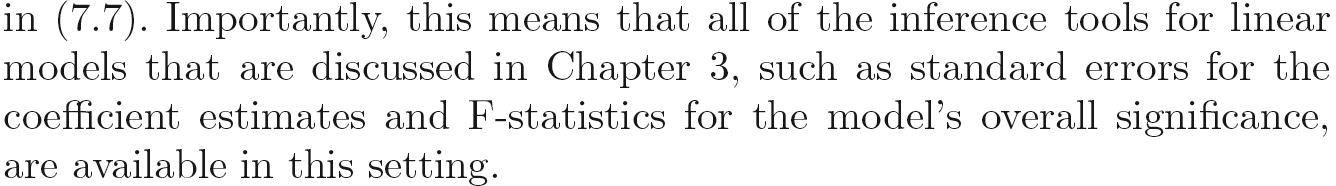
7.2 Step Functions



These are sometimes called *dummy* variables.

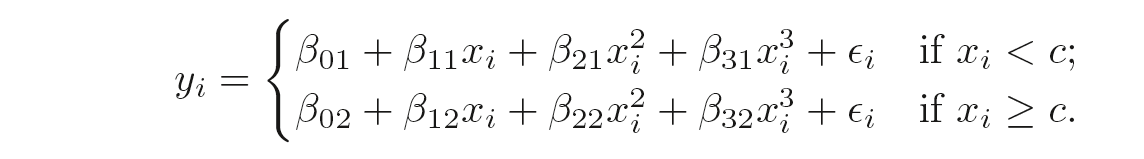
7.3 Basis Functions





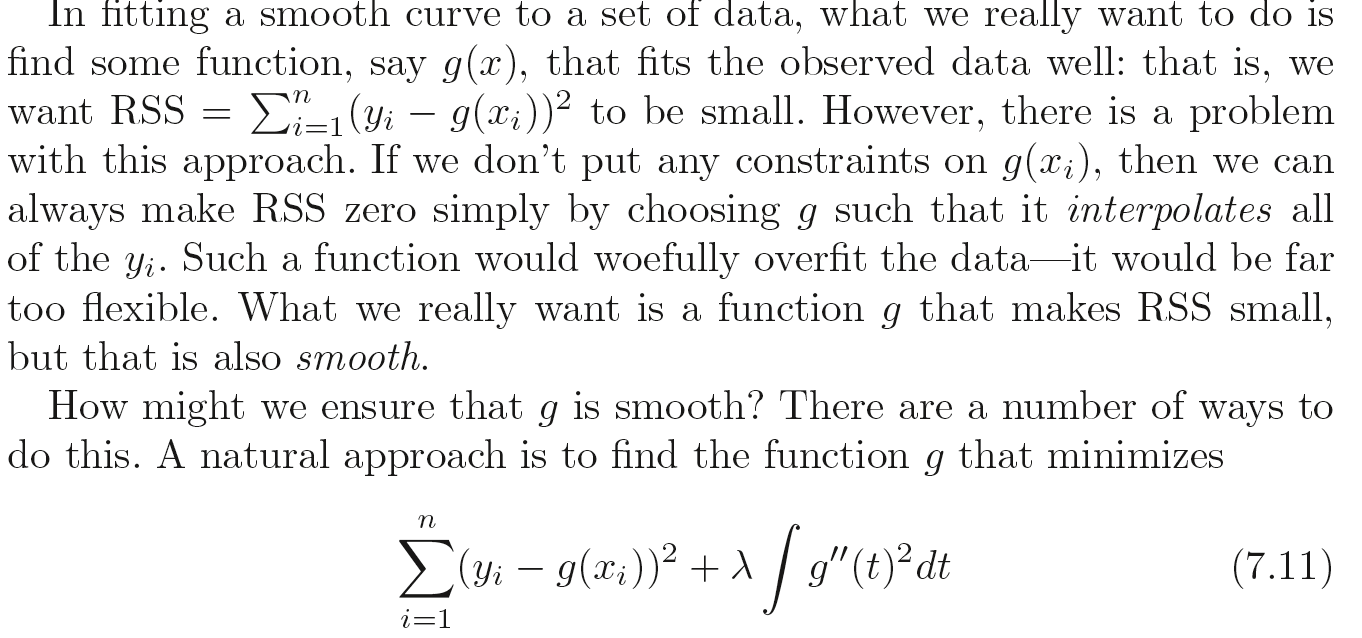
7.4 Regression Splines

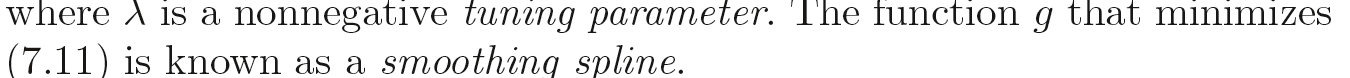
*7.4.1 Piecewise Polynomials*

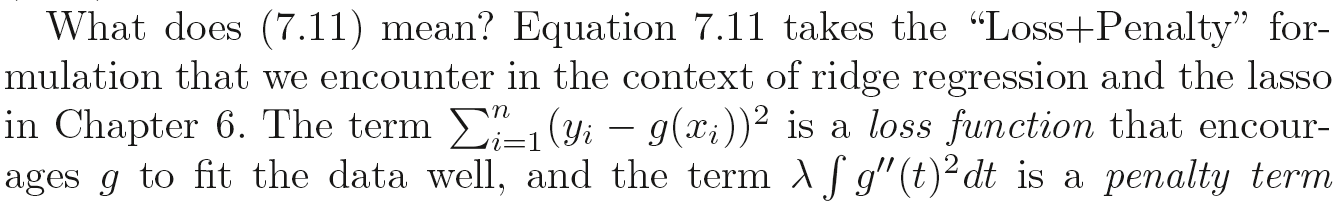


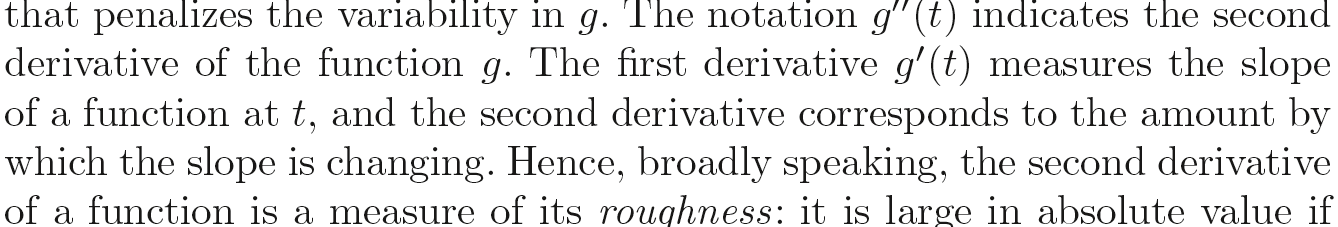
7.5 Smoothing Splines

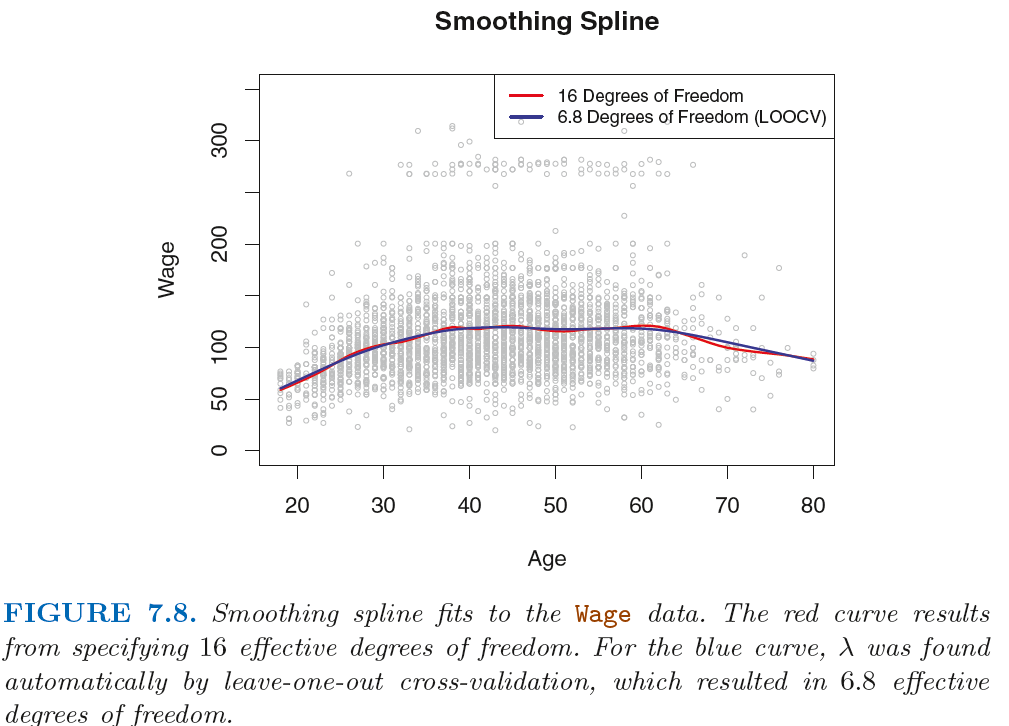
*7.5.1 An Overview of Smoothing Splines*



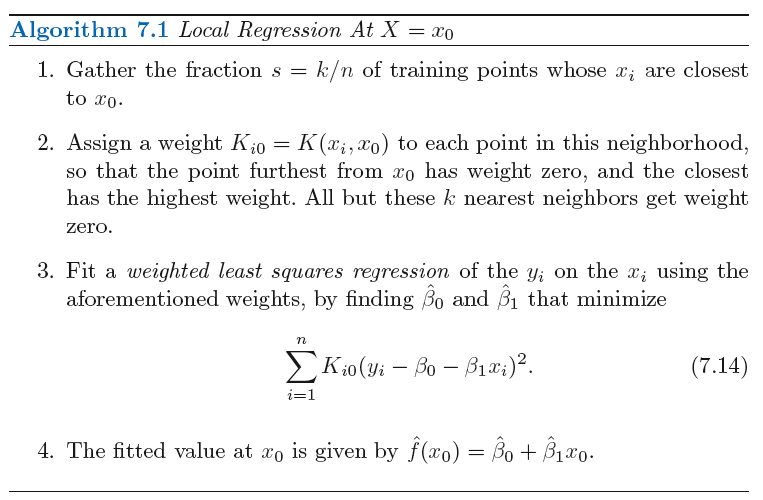








7.6 Local Regression



*However, local regression will have poor performance when p is as large as 3 or more, because in such case, there will be few observations close to x0. The nearest-neighbors regression suffers from this issue in high dimensions as well.*