Statistics 8330: Data Analysis III
Linear Regression: Model Selection and Regularization (Dimension Reduction)

Suggested Reading: James, Witten, Hastie, and Tibshirani (JWHT), 2013, Chapter 6

Supplemental Reading: Hastie, Tibshirani, and Friedman (HTF), 2009, Chapter 3

Christopher K. Wikle

University of Missouri
Department of Statistics

Dimension Reduction Methods

Now we consider a different way to deal with the bias-variance tradeoff in linear regression. In particular, we consider *dimension reduction* approaches that transform the original predictors into a new (and smaller) set of predictors.

The two most famous such methods are *principal component regression (PCR)* and *partial least squares (PLS)*. We considered Principal Components Analysis (PCA) in the context of multivariate analysis in DA II but have not focused on it in terms of regression.

Dimension Reduction Methods (cont.)

CKW

Let $Z_1, \ldots, Z_M$ represent $M < p$ linear combinations of our original p predictors:

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j, \quad m = 1, \ldots, M,$$

for constants $\phi_{1m}, \phi_{2m}, \ldots, \phi_{pm}$. We then consider the LS linear regression model:

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i, \quad i = 1, \ldots, n.$$

Clearly, the key is the choice of the constants (i.e., constraints) $\phi_{1m}, \ldots, \phi_{pm}$. Good choices of these can lead to better results than the LS solution on the original inputs.

Dimension Reduction Methods (cont.)

Note the special result obtained by substitution:

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{jm} x_{ij} = \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \phi_{jm} x_{ij} = \sum_{j=1}^{p} \beta_j x_{ij}.$$

Thus,

$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm},$$

which suggests that the reduced dimension regression is a special case of ordinary LS except that the parameters β_j must be constrained in this way. Note, if $M = p$ and the Z_m are all linearly independent, then this isn't a constraint at all, and the model fit is equivalent to the original LS fit.

As with the shrinkage methods, the constrained solution tends to be biased, but typically leads to lower variance in the fitted coefficients – particularly if $M << p$.

Principal Components Regression

CKW

Recall, in DA II we showed that PCA was useful for finding low-dimensional features in a large set of variables (i.e., it was an unsupervised learning method). The approach worked by finding the linear combination of the inputs that accounted for the largest fraction of variance of the inputs. Then, we found another linear combination of the inputs that accounted for the next highest amount of variation, subject to being orthogonal to the first, etc.

Mathematically, we can obtain the principal components in the following way using the **singular value decomposition** (SVD) of the centered data matrix. Let $\mathbf{X}$ denote the centered data matrix, which is of dimension $n \times p$. A famous result from matrix algebra, gives the SVD:

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T,$$

where $\mathbf{U}$ and $\mathbf{V}$ are $n \times p$ and $p \times p$ orthogonal matrices, respectively, and $\mathbf{D}$ is a $p \times p$ diagonal matrix with diagonal entries $d_1 \geq d_2 \geq \ldots d_p \geq 0$, called *singular values*.

PC Regression (cont.)

CKW

Note that one can then show that

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T,$$

which is the *eigen-decomposition* of $\mathbf{X}^T\mathbf{X}$ (and of the sample covariance matrix, $\mathbf{S} = \mathbf{X}^T\mathbf{X}/n$ up to the factor n).

Critically, the columns of $\mathbf{V}$ are called *eigenvectors* and correspond to the principal component weights. Thus, one can construct new variables as:

$$\mathbf{z}_1 = \mathbf{X}\mathbf{v}_1, \ldots, \mathbf{z}_M = \mathbf{X}\mathbf{v}_M,$$

where $\mathbf{v}_m$ is the mth column of $\mathbf{V}$. These $\mathbf{z}$s are the principal components. Thus, the elements of $\mathbf{v}_m$ play the role of the ϕs given previously.

In this case,

$$\text{var}(\mathbf{z}_m) = \frac{d_m^2}{n}.$$

Thus, large (small) singular values correspond to directions in the column space of $\mathbf{X}$ that have large (small) variance.

PC Regression (cont.)

CKW

PC regression (PCR) then consists of constructing the first M principal components (i.e., $Z_1, \ldots, Z_M$; note, we switched back from the vector notation in the previous slide, $Z_1 \equiv \mathbf{z}_1$, etc.) and uses these as predictors in the linear regression model fit by least squares.

There is a critical assumption in PCR that the directions in which $X_1, \ldots, X_p$ have the most variation are the directions that are most associated with Y. This is not always the case, but it is typically effective. Then, we still capture most of the information in the inputs (in terms of variance) but if $M << p$ we protect against over fitting.

We also get the added benefit of the new predictors being orthogonal (i.e., $\mathbf{z}_j^T \mathbf{z}_k = 0, \text{for } j \neq k$), which eliminates collinearity effects in the LS estimation.

PC Regression (cont.) [JWHT, 2013, Fig. 6.18]

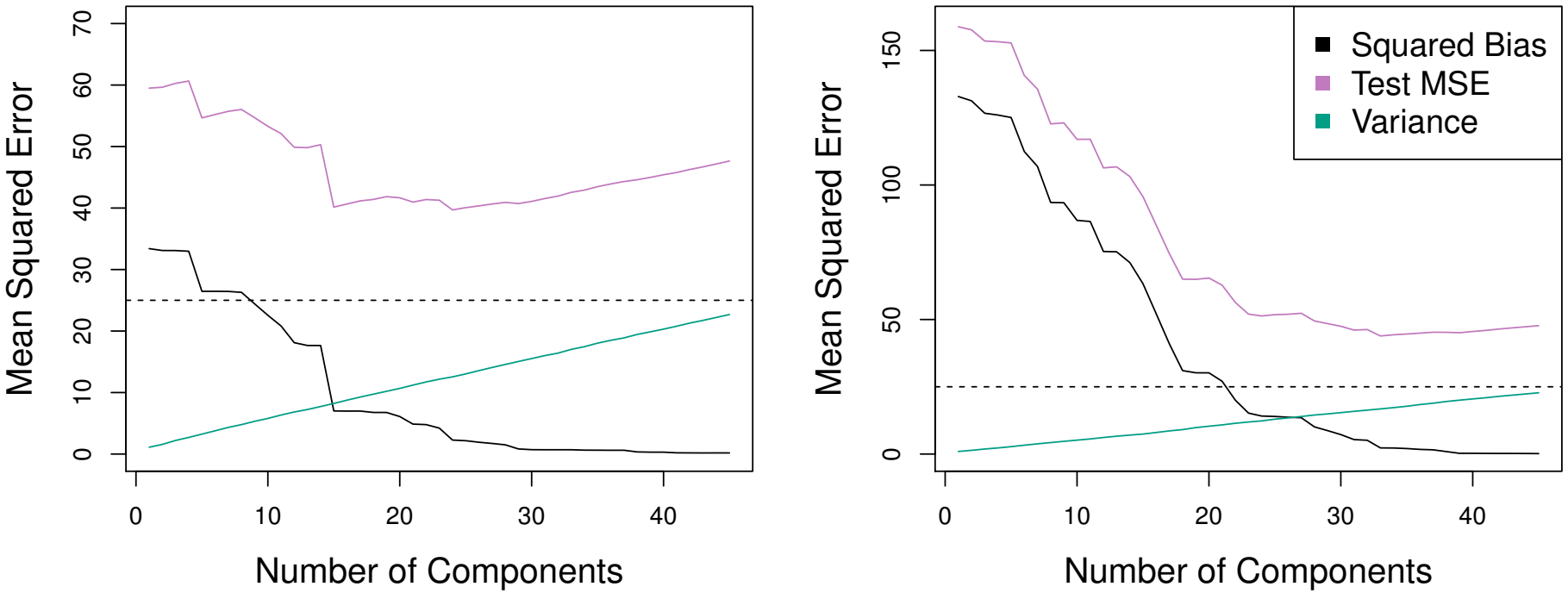

FIGURE 6.18. *PCR was applied to two simulated data sets.* Left: *Simulated data from Figure 6.8.* Right: *Simulated data from Figure 6.9.*

PC Regression (cont.)

CKW

It is important to note that although PCR uses a low-dimensional set of predictors,it does *not* perform model selection because all p variables go into the linear combination that makes the new PC inputs.

In fact, PCR is more closely tied to ridge regression than either best subset regression or lasso regression. One can show from the SVD decomposition given above and the formula for the ridge regression estimate (see HTF, 2009, Ch. 3.4.1) that:

$$\mathbf{X}\hat{\beta}^R = \sum_{j=1}^{p} \mathbf{u}_j \frac{d_j^2}{d_j^2 + \lambda} \mathbf{u}_j^T \mathbf{y},$$

where $\mathbf{u}_j$ corresponds to the jth column of $\mathbf{U}$ from the SVD given earlier. One can show that the LS estimate gives $\mathbf{X}\hat{\beta} = \mathbf{U}\mathbf{U}^T\mathbf{y}$, so that the ridge solution goes to the LS solution when $\lambda = 0$. Otherwise, the singular vectors that have smaller variation (i.e., smaller d_j^2) are shrunk more than those with larger variation.

PC Regression (cont.)

CKW

Note that PC regression is typically fairly efficient computationally because the SVD is efficient to compute and, with $M << p$, the LS fit on the PCs is very efficient.

We typically standardize the predictors before performing the SVD to generate the PCs. Otherwise, the predictors with the largest scale typically account for the most variance, regardless of their importance relative to the output. If the variables are all measured in the same units, then it is sometimes reasonable to use the unstandardized inputs to calculate the PCs.

Partial Least Squares (PLS)

CKW

PCR works by finding linear combinations of the data (or directions in the $\mathbf{X}$ matrix) that maximize variance and are orthogonal. It is very much an *unsupervised* procedure in that the choice of these directions is not at all dependent on Y, and thus may not actually provide good predictions of Y.

One could imagine that the regression relationship could be improved if we select these new variables in a way that accounts for the output Y – i.e., a *supervised* procedure.

The **partial least squares (PLS)** procedure provides such a supervised alternative to PCR. PLS is also a dimension reduction method that finds new features $Z_1, \ldots, Z_M$, that are linear combinations of the original inputs, but it selects these in a way to account for the covariability with the response Y.

Partial Least Squares (cont.)

CKW

To calculate the PLS directions (recall, $Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$):

- Standardize the p predictors
- For Z_1, set each ϕ_{j1} equal to the coefficient from the simple linear regression of Y on X_j (recall, this is proportional to the correlation between Y and X_j)
- To calculate Z_2, calculate the residuals of a regression of each variable (separately) on Z_1 (call these e_j); this corresponds to the information remaining that has not been explained by Z_1
- We then set ϕ_{j2} equal to the coefficient from the simple linear regression of Y on each e_j; this then allows us to get Z_2
- This is repeated M times to get $Z_1, \ldots, Z_M$

Partial Least Squares (cont.)

CKW

Comparison to PCR: (HTF, 2009, Ch 3.5)

One can show that the mth PC direction coefficients $\mathbf{v}_m$ are the $\boldsymbol{\alpha}$ that solve:

$$\max_{\alpha} [\text{var}(\mathbf{X}\boldsymbol{\alpha})],$$

subject to $||\boldsymbol{\alpha}|| = 1$, $\boldsymbol{\alpha}^T \mathbf{S}\mathbf{v}_\ell = 0$, $\ell = 1, \ldots, m-1$, where $\mathbf{S}$ is the sample covariance matrix associated with $\mathbf{X}$. (Note, the last condition ensures that each $\mathbf{z}_m = \mathbf{X}\mathbf{v}_m$ is uncorrelated with all of the previous $\mathbf{z}_\ell$).

The mth PLS direction $\hat{\phi}_m$ can be shown to be the $\boldsymbol{\alpha}$ that solves:

$$\max_{\alpha} [\text{corr}^2(\mathbf{y}, \mathbf{X}\boldsymbol{\alpha})\text{var}(\mathbf{X}\boldsymbol{\alpha})],$$

subject to $||\boldsymbol{\alpha}|| = 1$, $\boldsymbol{\alpha}^T \mathbf{S}\hat{\phi}_\ell = 0$, $\ell = 1, \ldots, m-1$.

Partial Least Squares (cont.)

Consider the first PCR and PLS direction for an example shown in JWHT (2013, 6.3.2):

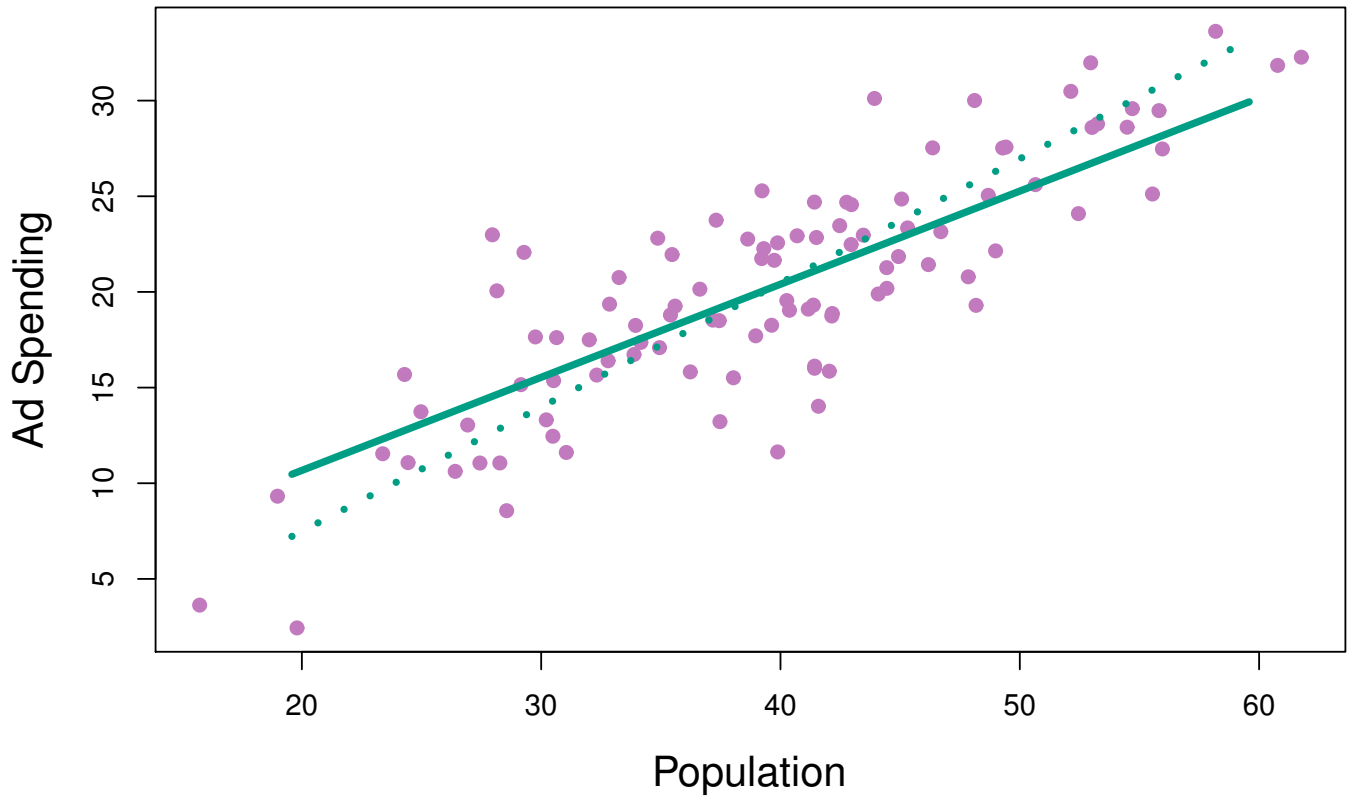

FIGURE 6.21. *For the advertising data, the first PLS direction (solid line) and first PCR direction (dotted line) are shown.*

Partial Least Squares (cont.)

CKW

Comments on PLS:

- The predictors and response are typically standardized before performing PLS
- The choice of M is typically made by cross-validation
- Although PLS seeks linear combinations that maximize both the correlation to the output as well as the variance in the inputs, in many cases the variance tends to dominate and PLS behaves similarly to ridge regression and PCR
- Although the use of the output to supervise the selection of the reduced dimensions can reduce bias, it also has the potential to increase variance, reducing the added value of the approach relative to PCR
- PLS is closely related to the notion of *canonical correlation analysis (CCA)* in multivariate statistics, which seeks to find a reduced set of linear combinations of inputs and linear combination of outputs that maximize the correlation; e.g., finding the coefficients $\boldsymbol{\alpha}, \boldsymbol{\beta}$ such that

$$\max_{\alpha,\beta} [\text{corr}(\mathbf{Y}\boldsymbol{\beta}, \mathbf{X}\boldsymbol{\alpha})],$$

where again, one finds a set of M of these with orthogonality constraints.

Regularization and Dimension Reduction Example (HTF, 2009, Table 3.3)

TABLE 3.3. *Estimated coefficients and test error results, for different subset and shrinkage methods applied to the prostate data. The blank entries correspond to variables omitted.*

Term	LS	Best Subset	Ridge	Lasso	PCR	PLS
`Intercept`	2.465	2.477	2.452	2.468	2.497	2.452
`lcavol`	0.680	0.740	0.420	0.533	0.543	0.419
`lweight`	0.263	0.316	0.238	0.169	0.289	0.344
`age`	−0.141		−0.046		−0.152	−0.026
`lbph`	0.210		0.162	0.002	0.214	0.220
`svi`	0.305		0.227	0.094	0.315	0.243
`lcp`	−0.288		0.000		−0.051	0.079
`gleason`	−0.021		0.040		0.232	0.011
`pgg45`	0.267		0.133		−0.056	0.084
Test Error	0.521	0.492	0.492	0.479	0.449	0.528
Std Error	0.179	0.143	0.165	0.164	0.105	0.152

Considerations for High Dimensional Data

As we have mentioned, traditional linear regression works quite well in low-dimensional settings (i.e., when $p << n$). Increasingly, we are seeing an explosion in the number of potential predictors (p) that are available for fitting models. Many of these may be related to the response, many of them may not be.

The main problem with situations where $p \approx n$ or $p > n$ is that it is easy to over fit the responses. That is, the model is TOO flexible. This leads invariably to models that appear better than they really are. Consider the examples from JWHT (2013, 6.4.2):

Considerations for High Dimensional Data (cont.)

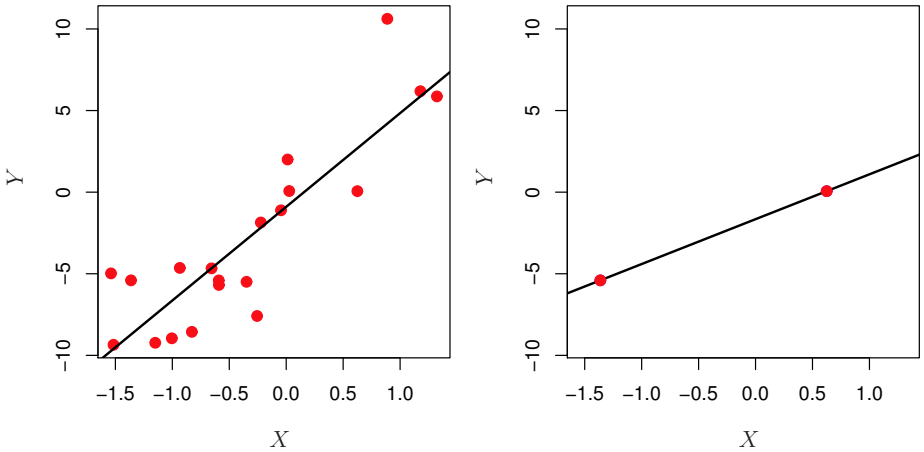

FIGURE 6.22. Left: *Least squares regression in the low-dimensional setting.* Right: *Least squares regression with $n = 2$ observations and two parameters to be estimated (an intercept and a coefficient).*

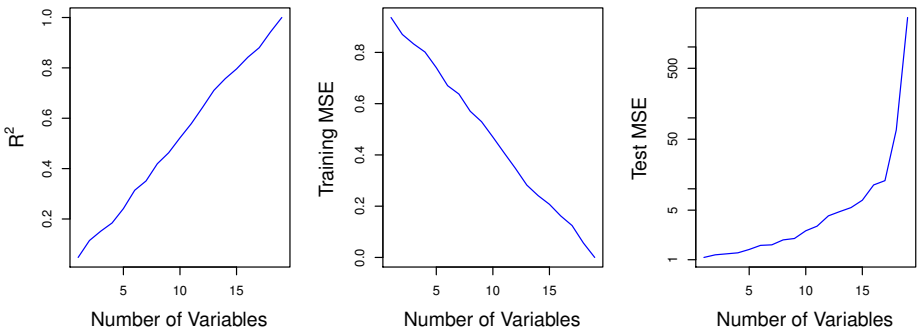

FIGURE 6.23. *On a simulated example with $n = 20$ training observations, features that are completely unrelated to the outcome are added to the model.* Left: *The R^2 increases to 1 as more features are included.* Center: *The training set MSE decreases to 0 as more features are included.* Right: *The test set MSE increases as more features are included.*

Considerations for High Dimensional Data (cont.)

As we have seen, one of the best ways to deal with the "large p, small n" problem is through the use of less flexible models (i.e., those that are regularized or constrained) such as subset selection, ridge regression, the lasso, PCR, or PLS. Although these play a crucial role in dealing with high-dimensional problems, there are still issues that we must consider.

- When p is large there is always an issue with collinearity; methods that select subsets (including the lasso) are not guaranteed to give the most scientific model
- p-values are not typically useful in high-dimensional regressions
- Appropriate tuning parameter selection is important for good predictive performance
- The test error tends to increase as the dimensionality of the problem increases, unless the additional features are truly associated with the response (i.e., the curse of dimensionality!)

Considerations for High Dimensional Data (cont.)

The last point on the previous page is very important: as technology allows for even larger sets of input variables, they may lead to better predictive models only if those variables are really related to the response, but if they are not strongly related or not related at all, they can lead to worse results. That is, the variance associated with their fitting may outweigh the reduction in bias they bring (see JWHT, 2013, 6.4.3).

This is something to think about when building models or advising people on building models!