BUAD 5082

Machine Learning II

Dimension Reduction

(ISLR, Chapter 6)

This Week’s Agenda

- The basic idea behind dimension reduction techniques
- Today:
  - Introduction to Principal Component Analysis
  - Principal Component Regression
- Wednesday:
  - Partial Least Squares Regression
  - Considerations in High Dimensions

Quick Review

- Major question being addressed in last few sessions:
  - How to control the increased variance associated with models that include a large number of predictors *p*?
- Have discussed two approaches so far:
  - Subset Selection: Best, Forward, Backward
    - Eliminate relatively unimportant predictors altogether
  - Shrinkage methods: Ridge, Lasso
    - Keep all *p* predictors, but shrink the coefficients (and in the case of the Lasso, eliminate some (called Regularization))
- Current topic: Dimension Reduction
  - Approaches that *transform* the predictors and then fit a least squares model using these transformed variables.

DIMENSION REDUCTION WITH PCA

Dimension Reduction

- Let $Z_1, Z_2 \ldots, Z_M$ represent M < p *linear combinations* of our original p predictors. That is,

$$Z_m = \sum_{j=1}^{p} \varphi_{jm} X_j$$

for some constants $\varphi_{1m}, \varphi_{2m} \ldots, \varphi_{pm}$ where $m = 1, 2 \ldots, M$.

- We can then fit the linear regression model

$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \varepsilon_i$ instead of $y_i = \beta_0 + \sum_{j=1}^{p} \beta_i x_{ip} + \varepsilon_i$

using least squares. Note that in this expression the regression coefficients are given by $\theta_1, \theta_2 \ldots, \theta_M$ - the z_{im} have already been specified.

- Note also that in the foregoing,

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

where $\beta_j = \sum_{m=1}^{M} \theta_m \varphi_{jm}$

Dimension Reduction

- The term *dimension reduction* comes from the fact that this approach reduces the problem of estimating the p+1 coefficients $\beta_0, \beta_1, \ldots, \beta_p$ to the simpler problem of estimating the $M + 1$ coefficients $\theta_0, \theta_1, \ldots, \theta_M$, where $M < p$.
- Dimension reduction serves to constrain the estimated β_j coefficients, since now they must take the form $\sum_{m=1}^{M} \theta_m \varphi_{jm}$.
- This is a bias/variance story:
  - Constraining the form of the coefficients has the potential to bias the coefficient estimates, but…
  - …in situations where p is large relative to n, selecting a value of $M << p$ can significantly reduce the variance of the fitted coefficients.

Dimension Reduction

- All dimension reduction methods work in two steps:
  1. The transformed predictors $Z_1, Z_2, \ldots, Z_M$ are obtained.
  2. The model is fit using these M predictors.
- However, the choice of $Z_1, Z_2, \ldots, Z_M$, or equivalently, the selection of the φ_{jm}'s, can be achieved in different ways.
- In the next session, we will consider two approaches for this task:
  - Principal Components
  - Partial Least Squares.

PRINCIPAL COMPONENT ANALYSIS (BRIEFLY)

We will discuss this topic more thoroughly when we discuss Unsupervised Learning later in the course.

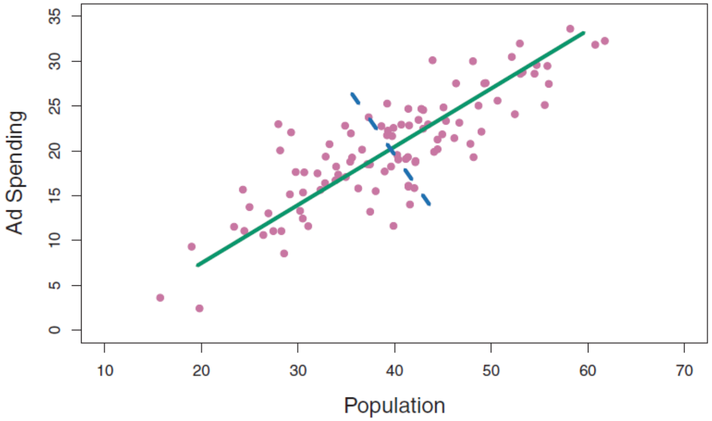
Principal Component Analysis

- PCA is a technique for reducing the dimension of a data set X from $n \times p$ to $n \times (p - d)$, where $d > 0$.
  - When faced with a large set of (possibly highly correlated) variables, principal components allow us to summarize this set with a smaller number of representative variables that collectively explain most of the variability in the original set and are uncorrelated.
- Applications of Principal Component Analysis (PCA)
  - Machine Learning: To reduce the dimensionality of data in predictive analytics (our current purpose)
  - Visualization
  - Anomaly Detection
  - Matching/Distance Calculations
    - Facial Recognition – eigenfaces
  - Compression
  - Latent Semantic Indexing
    - A text analytics technique we’ll return to later in the semester

Principal Component Analysis

- The *first principal component* **direction** of a dataset is that direction along which the observations *vary the most*.
  - The implicit assumption is that the response will tend to vary most in the directions of high variance of the inputs.
  - This is often a reasonable assumption, since predictors are often chosen for study because they vary with the response variable.
- The second and subsequent principal component directions are the directions along which the projected observations vary the most, subject to the constraint that each new direction is orthogonal to all the foregoing principal component directions
- There can be at most $\min(n - 1, p)$ principal components.

Principal Component Analysis



- Example: The panel above plots population size in tens of thousands of people, against ad spending for a particular company in thousands of dollars, for 100 cities (only a subset of the points are shown).
- The green solid line represents the first principal component **direction** of the data.
- That is, if we *projected* the 100 observations onto this line then the resulting projected observations would have the largest possible variance; projecting the observations onto any other line would yield projected observations with lower variance.
  - Projecting a point onto a line simply involves finding the location on the line which is closest to the point.

Principal Component Analysis

- In this example, the first principal component can be expressed mathematically as:

$$Z_1 = 0.896(pop - \overline{pop}) + 0.544(ad - \overline{ad})$$

- Here φ_{11} = 0.896 and φ_{21} = 0.544 are the first principal component *loadings*, which define the direction referred to above.
  - $\overline{\text{pop}}$ and $\overline{\text{ad}}$ refer to the means of these features, so this expression operates on centered values of pop and ad.
- The idea is that out of every possible linear combination of pop and ad such that $\varphi_{11}^2 + \varphi_{21}^2 = 1$, this particular linear combination yields the highest variance.
  - That is, φ_{11} and φ_{21} have been chosen so that the equation above is *the* linear combination which maximizes Var(φ_{11}× (pop − $\overline{\text{pop}}$) + φ_{21} × (ad − $\overline{\text{ad}}$))
  - It is necessary to consider only linear combinations of the form $\varphi_{11}^2 + \varphi_{21}^2 = 1$, since otherwise we could increase φ_{11} and φ_{21} arbitrarily without bound to increase the variance.

Principal Component Analysis

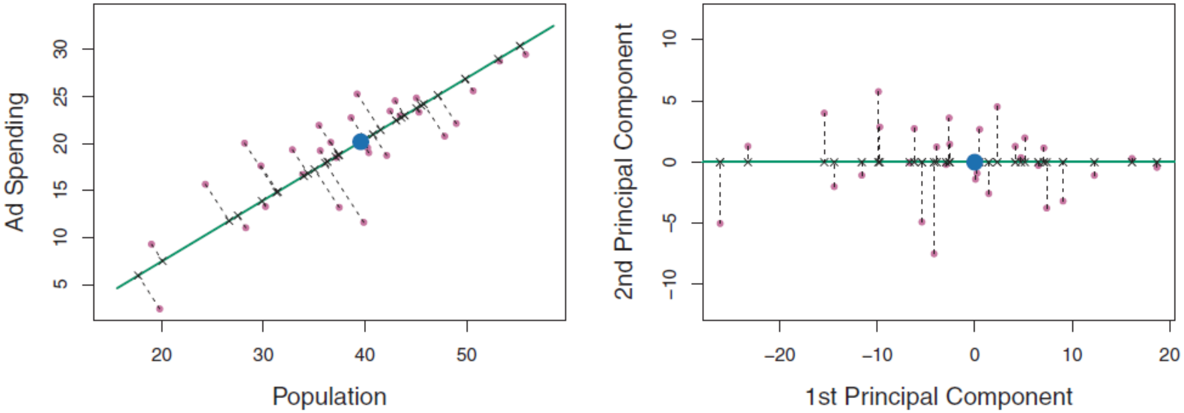
$$Z_1 = 0.896(pop - \overline{pop}) + 0.544(ad - \overline{ad})$$

- Since *n* = 100, pop and ad are vectors of length 100 in the above expression, and therefore so is Z_1. So…

$$Z_{i1} = 0.896(pop_i - \overline{pop}) + 0.544(ad_i - \overline{ad})$$

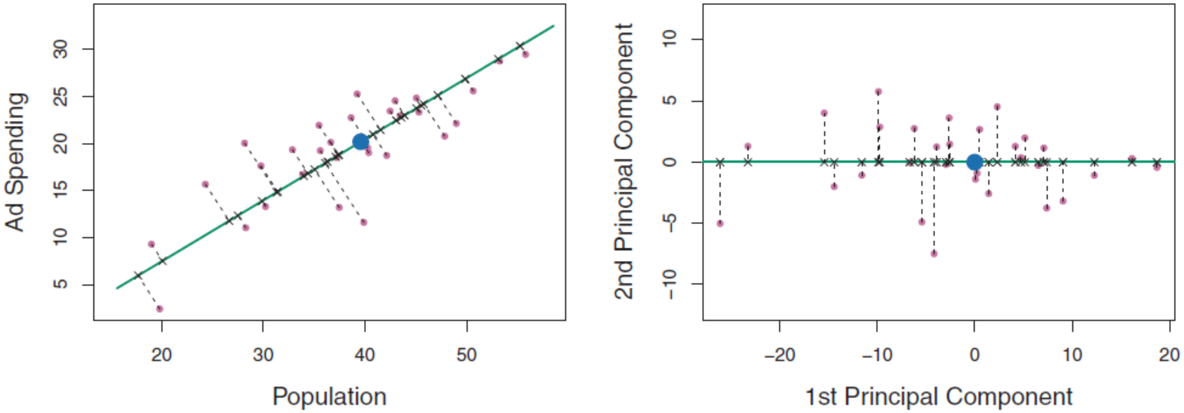
- The values of $z_{11}, \ldots, z_{n1}$ are known as the *principal component scores*.

Principal Component Analysis



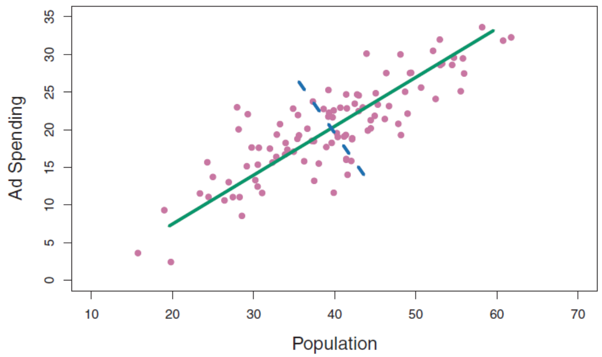
- In the right panel, the left-hand panel has been rotated so that the first principal component direction coincides with the x-axis.
- The first principal component score for the *i*$_{\text{th}}$ observation is the distance in the *x*-direction of the *i*th cross from zero.
  - So for example, the point in the bottom-left corner of the left-hand panel above has a large negative principal component score, $z_{i1} = -26.1$, while the point in the top-right corner has a large positive score, $z_{j1} = 18.7$.

Principal Component Analysis



- We can think of the values of the principal component Z_1 as single number summaries of the joint pop and ad budgets for each location.
- For example, if $z_{i1} < 0$, then this indicates a city with below-average population size and below average ad spending. A positive score suggests the opposite.
- In this case, pop and ad have approximately a linear relationship, and so we might expect that a single-number summary will work well.

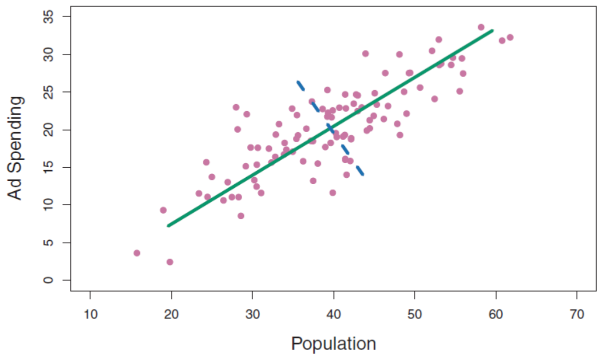
Principal Component Analysis



- So far we have concentrated on the first principal component. In general, one can construct up to *p* distinct principal components (assuming that $n \geq p$).
- The second principal component Z_2 is a linear combination of the variables that is uncorrelated with Z_1, and has largest variance subject to this constraint.
- The second principal component direction is illustrated as a dashed blue line above. It turns out that the zero correlation condition of Z_1 with Z_2 is equivalent to the condition that the direction must be *perpendicular*, or *orthogonal*, to the first principal component direction.
- The second principal component is given by the formula

$$Z_2 = 0.443(pop - \overline{pop}) - 0.897(ad - \overline{ad})$$

Principal Component Analysis



- Since the advertising data has two predictors, the first two principal components contain all of the information that is in pop and ad.
- However, by construction, the first component will contain the most information.
  - Consider, for example, the much larger variability of z_{i1} (the *x*-axis) versus z_{i2} (the *y*-axis) in the figure above.
- The fact that the second principal component scores are much closer to zero indicates that this component captures far less information, suggesting that in this case, one only needs the first principal component in order to accurately represent the pop and ad budgets.
- See PrincipalComponentAnalysisExcelExample.xlsm, sheet PCAMaxVar for the detailed calculations for this example

THE PRINCIPAL COMPONENTS REGRESSION APPROACH

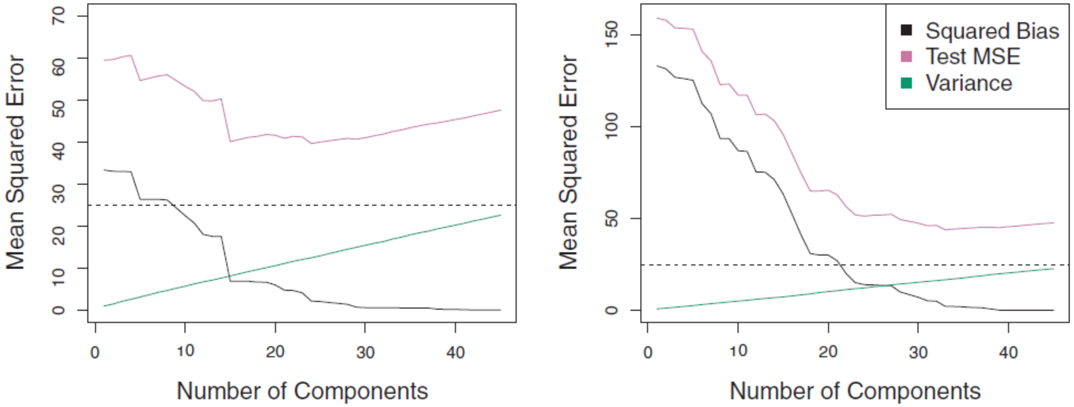
Principal Components Regression

- The *principal components regression* (PCR) approach involves constructing the first *M* principal components, $Z_1, . . ., Z_M$, and then using these components as the predictors in a linear regression model that is fit using least squares.
- The key idea is that often a small number of principal components suffice to explain most of the variability in the data, as well as the relationship with the response.

Principal Components Regression

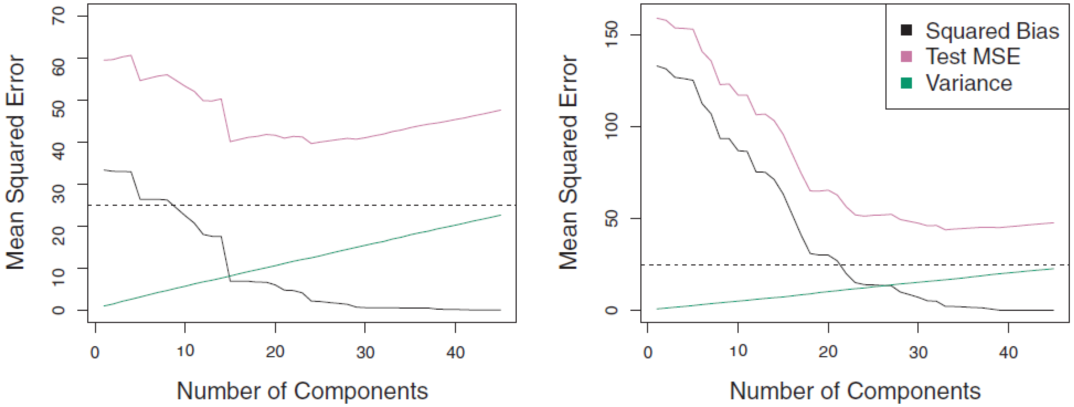
- If this assumption holds, then fitting a least squares model to $Z_1, \ldots, Z_M$ will lead to better results than fitting a least squares model to $X_1, \ldots, X_p$
  - This is so since most or all of the information in the data that relates to the response is contained in $Z_1, \ldots, Z_M$, and by estimating only $M \ll p$ coefficients we can mitigate overfitting.
- In the advertising data, the first principal component explains most of the variance in both pop and ad, so a principal component regression that uses this single variable to predict some response of interest, such as sales, will likely perform quite well.

Principal Components Regression



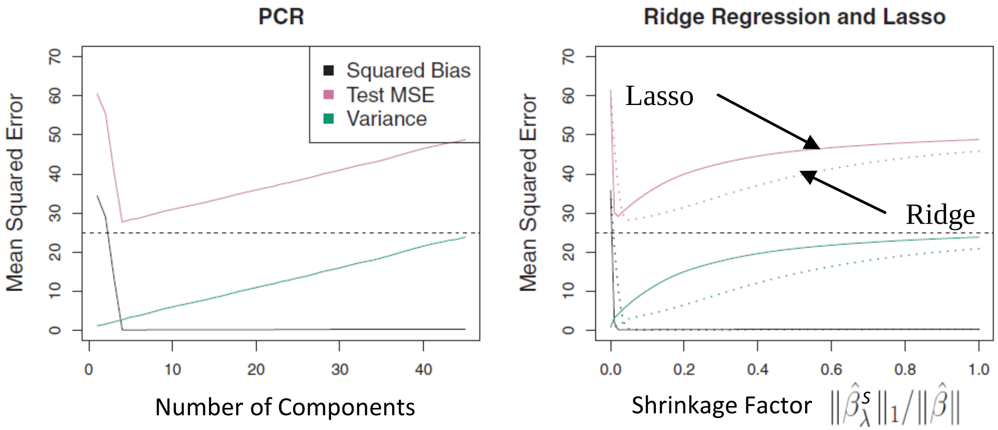
- PCR was applied to a simulated data set with $n = 50$ and $p = 45$.
- In the left-hand panel, the response is truly related to all 45 predictors, while in the right-hand panel, the response is truly related to only 2 of the predictors.
- When $M = p = 45$, then PCR amounts simply to a least squares fit using all of the original predictors.

Principal Components Regression



- Observe that in both cases, as more principal components are used in the regression model, the bias decreases, but the variance increases. This results in a typical U-shape for the mean squared error.
- Recalling our analysis of these datasets using Ridge regression and the Lasso, both of these approaches out-performed PCR.
- But this data was generated in such a way that many principal components are required in order to adequately model the response.
- In contrast, PCR will tend to do well in cases when the first few principal components are sufficient to capture most of the variation in the predictors as well as the relationship with the response.

Principal Components Regression

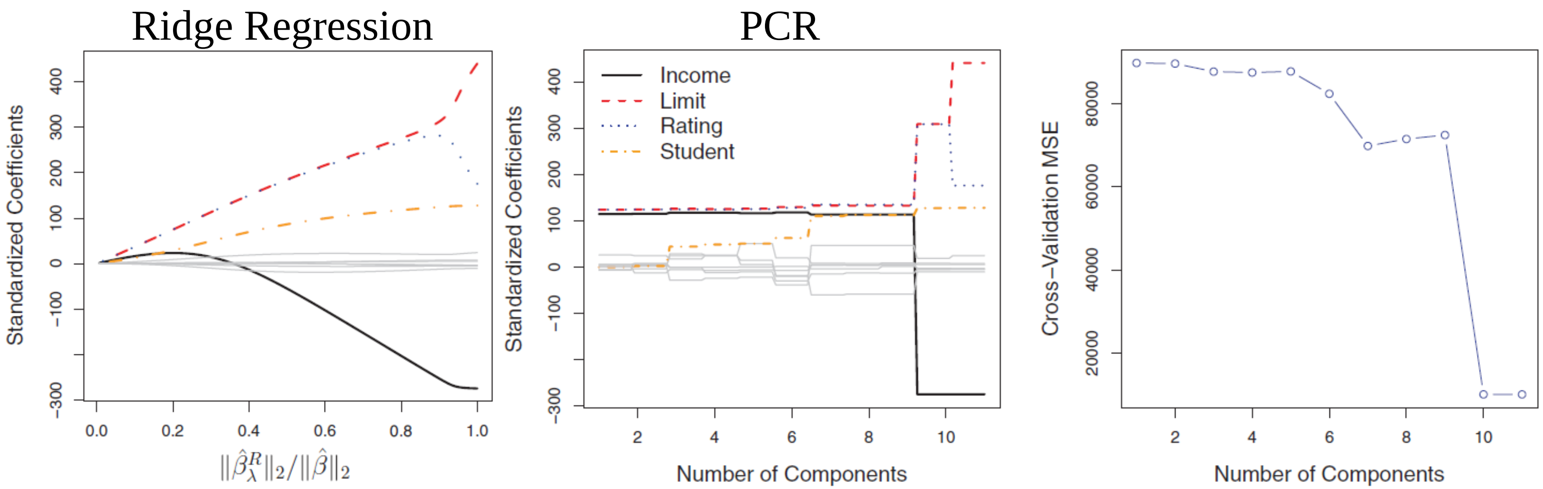


- The left-hand panel above illustrates the results from another simulated data set designed to be more favorable to PCR.
- Here the response depends exclusively on the first five principal components.
- Now the bias drops to zero rapidly as *M*, the number of principal components used in PCR, increases.
- The mean squared error displays a clear minimum at $M = 5$.
- The right-hand panel displays the results on these data using Ridge regression and the Lasso.
- All three methods offer a significant improvement over least squares. However, PCR and Ridge regression slightly outperform the Lasso.

Principal Components Regression

- Note that PCR provides a simple way to perform regression using $M < p$ predictors, it is *not* a feature selection method.
- This is because each of the M principal components used in the regression is a linear combination of all p of the *original* features.
- In this sense, PCR is more closely related to Ridge regression than to the Lasso.
- Although it is beyond our scope, one can show that Ridge regression is a continuous version of PCR!

Principal Components Regression



Ridge Regression

PCR

- In PCR, the number of principal components, *M*, is typically chosen by cross-validation.
- The results of applying Ridge regression and PCR to the Credit data set are shown in above in the two left-hand panels.
- The right-hand panel displays the cross-validation errors obtained, as a function of *M*.
- On these data, the lowest cross validation error occurs when there are M = 10 components; this corresponds to almost no dimension reduction at all, since PCR with M = 11 is equivalent to simply performing least squares.

Principal Components Regression

- When performing PCR, we generally *standardize* each predictor prior to generating the principal components.
- In the absence of standardization, the high-variance variables will tend to play a larger role in the principal components obtained, and the scale on which the variables are measured will ultimately have an effect on the final PCR model.
- However, if the variables are all measured in the same units (say, kilograms, or inches), then one might choose not to standardize them.

THE PARTIAL LEAST SQUARES REGRESSION APPROACH

Partial Least Squares

- The PCR approach involves identifying linear combinations, or *directions*, that best represent the predictors $X_1, . . .,X_p$.
- These directions are identified in an *unsupervised* way, since the response *Y* is not used to help determine the principal component directions.
    - That is, the response does not *supervise* the identification of the principal components.
- Consequently, PCR suffers from a drawback: there is no guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response.

Partial Least Squares

- Partial least squares (PLS) is a *supervised* alternative to PCR.
- Like PCR, PLS is a dimension reduction method, which first identifies a new set of features $Z_1, \ldots,$ Z_M that are linear combinations of the original features, and then fits a linear model via least squares using these M new features.
- But unlike PCR, PLS identifies these new features in a supervised way—that is, it makes use of the response Y in order to identify new features that not only approximate the old features well, but also that are related to the response.
  - Roughly speaking, the PLS approach attempts to find directions that help explain both the response and the predictors.

Partial Least Squares

Recall that in simple regression, $\hat{\beta}_1 = \frac{Cov(x,y)}{Var(x)}$

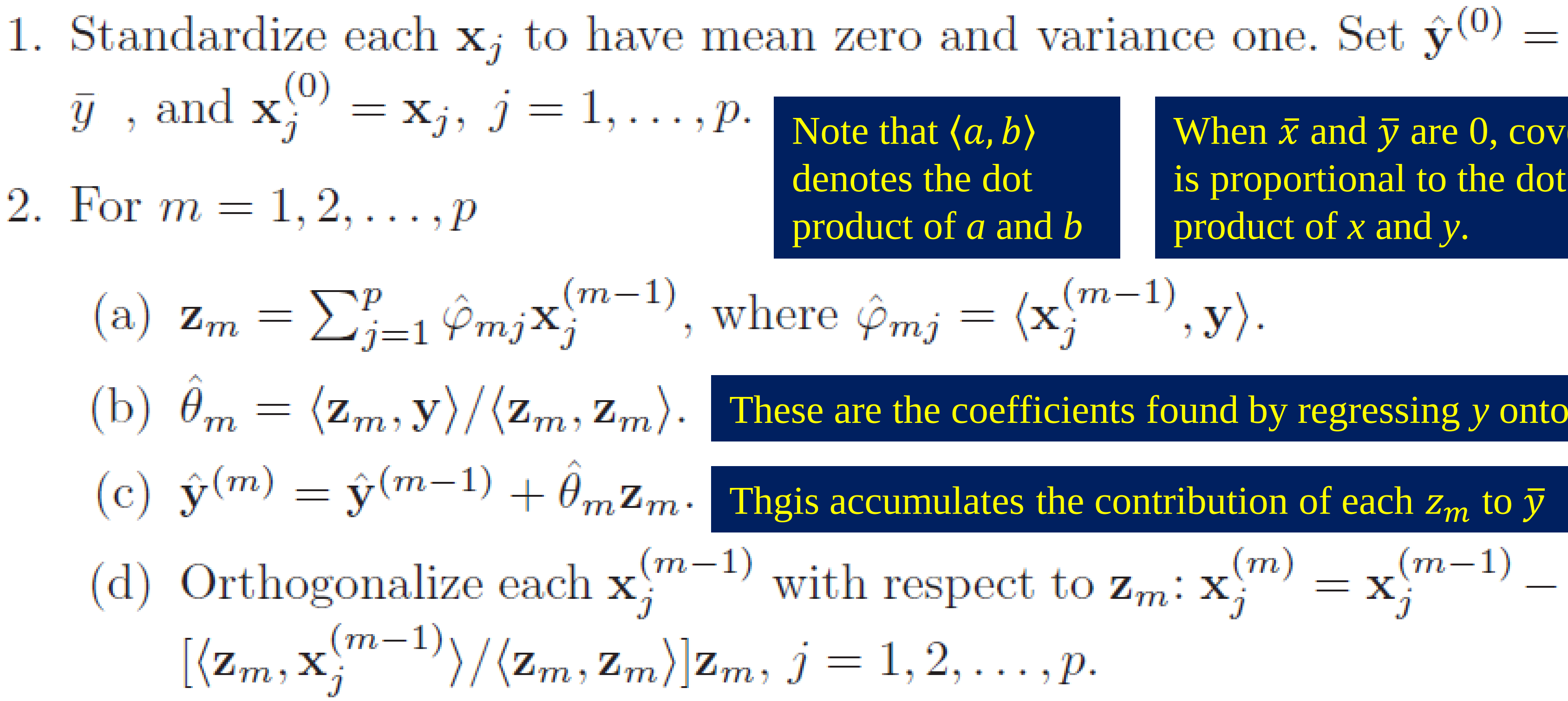
The steps in PLS regression are as follows:

- PLS is not scale invariant, so we assume that each x_j is standardized to have mean 0 and variance 1.
- PLS begins by computing $\varphi_{1j} = cov(x_j, y)$ for each j
  - These are just the coefficients of the j simple regressions of the y onto the individual x's.
- From this we construct the derived input $z_1 = \sum_{j-1}^{p} \varphi_{1j} x_j$ (just as we did with PCA, where there the φ_{1j} was the first principal component vector). This is called the first partial least squares direction.
  - Note that in the construction of each z_m, the inputs are weighted by the strength of their univariate effect on y.
- To identify the second PLS direction we first *adjust* each of the variables for z_1 by regressing each variable on Z_1 and taking *residuals*.
  - These residuals can be interpreted as the remaining information that has not been explained by the first PLS direction.
- We then compute z_2 using this *orthogonalized* data in exactly the same fashion as z_1 was computed based on the original data.
- We continue this process, until $M \leq p$ directions have been obtained.

Partial Least Squares

- In this manner, partial least squares produces a sequence of derived, orthogonal inputs or directions z_1, z_2,…, z_M.
- As with principal-component regression, if we were to construct all $M = p$ directions, we would get back a solution equivalent to the usual least squares estimates; using M < p directions produces a reduced regression.
- The procedure is described more fully in the next slide.

Partial Least Squares



1. Standardize each $\mathbf{x}_j$ to have mean zero and variance one. Set $\hat{\mathbf{y}}^{(0)} = \bar{y}$, and $\mathbf{x}_j^{(0)} = \mathbf{x}_j$, $j = 1, \ldots, p$.

2. For $m = 1, 2, \ldots, p$

Note that $\langle a, b \rangle$ denotes the dot product of a and b

When $\bar{x}$ and $\bar{y}$ are 0, cov(x, y) is proportional to the dot product of x and y.

   (a) $\mathbf{z}_m = \sum_{j=1}^{p} \hat{\varphi}_{mj} \mathbf{x}_j^{(m-1)}$, where $\hat{\varphi}_{mj} = \langle \mathbf{x}_j^{(m-1)}, \mathbf{y} \rangle$.

   (b) $\hat{\theta}_m = \langle \mathbf{z}_m, \mathbf{y} \rangle / \langle \mathbf{z}_m, \mathbf{z}_m \rangle$.

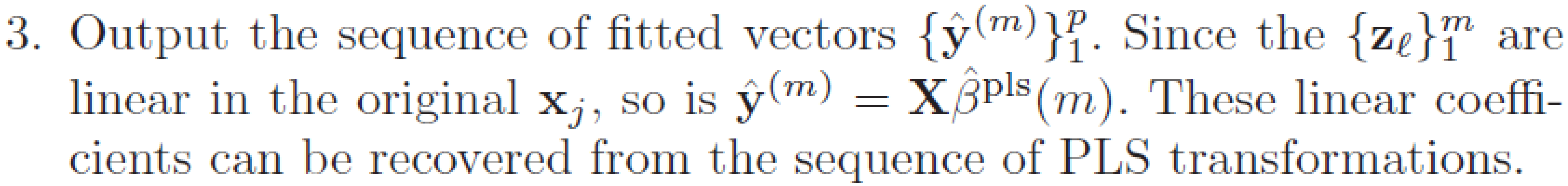
These are the coefficients found by regressing y onto z_m

   (c) $\hat{\mathbf{y}}^{(m)} = \hat{\mathbf{y}}^{(m-1)} + \hat{\theta}_m \mathbf{z}_m$.

Thgis accumulates the contribution of each z_m to $\bar{y}$

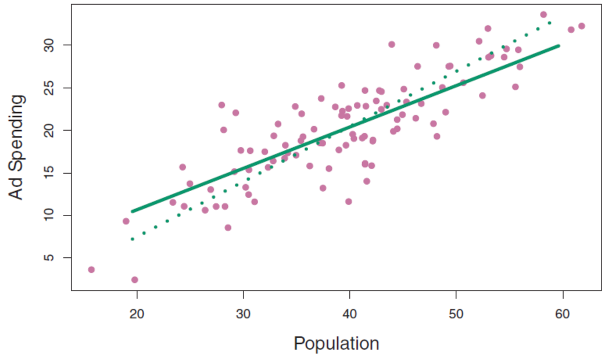
   (d) Orthogonalize each $\mathbf{x}_j^{(m-1)}$ with respect to $\mathbf{z}_m$: $\mathbf{x}_j^{(m)} = \mathbf{x}_j^{(m-1)} - [\langle \mathbf{z}_m, \mathbf{x}_j^{(m-1)} \rangle / \langle \mathbf{z}_m, \mathbf{z}_m \rangle] \mathbf{z}_m$, $j = 1, 2, \ldots, p$.

$x_j^{(m)}$ is the vector of residuals in the regression of the previous set of residuals $x_j^{(m-1)}$ onto z_m - this amounts to the portion of y that remains unexplained so far. The residual vector is always orthogonal to the prediction vector.



3. Output the sequence of fitted vectors $\{\hat{\mathbf{y}}^{(m)}\}_1^p$. Since the $\{\mathbf{z}_\ell\}_1^m$ are linear in the original $\mathbf{x}_j$, so is $\hat{\mathbf{y}}^{(m)} = \mathbf{X}\hat{\beta}^{\text{pls}}(m)$. These linear coefficients can be recovered from the sequence of PLS transformations.

Partial Least Squares



- The chart above displays an example of PLS on the advertising data.
  - The solid green line indicates the first PLS direction, while the dotted line shows the first principal component direction.
  - PLS has chosen a direction that has less change in the ad dimension per unit change in the pop dimension, relative to PCA.
- Again pop is identified as being more highly correlated with the response than is ad.
- Note that the PLS direction does not fit the predictors as closely as does PCA, but it does a better job explaining the response.

Partial Least Squares

- As with PCR, the number *M* of partial least squares directions used in PLS is a tuning parameter that is typically chosen by cross-validation.
- In practice it often performs no better than ridge regression or PCR.
- While the supervised dimension reduction of PLS can reduce bias, it also has the potential to increase variance, since it is more closely dependent on the training y’s, so that the benefit of supervision in PLS of often cancelled out relative to PCR.