# Part 6: Model Selection and Intro to ML

Chris Conlon

March 6, 2021

Applied Econometrics II

# Principal Components Regression

# Other Data Reduction Techniques

We have other data reduction techniques with a long history in Econometrics

- Principal Components
- Factor Analysis
- Partial Least Squares

# Principal Components

Suppose we have a very high dimensional $X$ where we have a high degree of correlation among the components $x_j$.

- We can start by computing the appropriate correlation matrix $C = E[\tilde{X}'\tilde{X}]$ where $\tilde{X}$ denotes we have standardized each column $x_j$ to have mean zero and variance 1.
- Diagonlize $C$ via the eigen-decomposition $V^{-1}CV = D$ where $D$ is the diagonal matrix of eigenvalues.
- Sort $D$ and the corresponding columns of $V$ in decreasing order of the eigenvalues $d_j$.
- Choose a subset of $m < K$ eigenvalues and eigenvectors and call that $V_m$ and $\lambda_m$
- Compute transformed data: $Z_m = V_m\tilde{X}$ which is of dimension $N \times m$ instead of $N \times K$.

# Principal Components

- If $m \ll K$ then we can substantially reduce the dimension of the data.
- The idea is to choose $m$ so that $(Z'Z)$ spans approximately the same space that $(X'X)$ does.
- This works because we use the principal eigenvectors (those with the largest eigenvalues).
- The first eigenvector explains most of the variation in the data, the second the most of the remaining variation, and so on.
- You may also recall that eigenvectors form an orthonormal basis, so each dimension is linearly independent of the others.
- As eigenvalues decline, it means they explain less of the variance.

# Principal Components

Output from software will inlcude

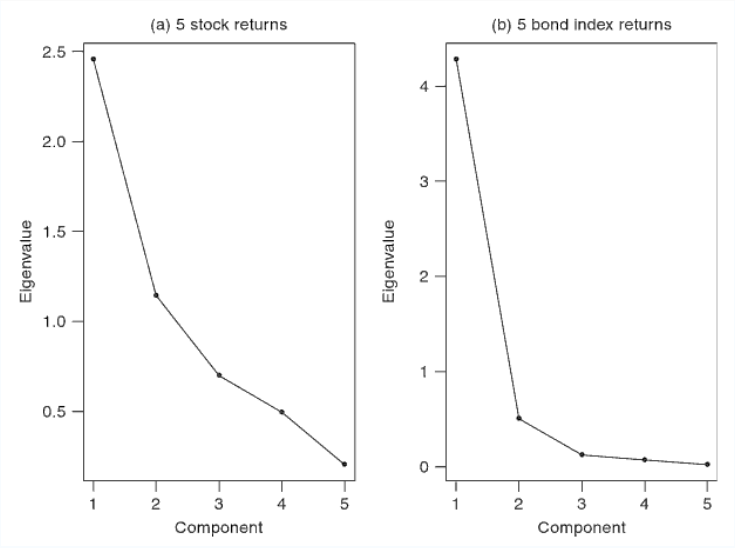
- Coefficients: these transform from $X \to Z$
- Score: these are the transformed $Z$'s
- Latent/Eigenvalue: the corresponding Eigenvalue
- Explained/Cumulative: cumulative explained variance $\sum_{j=1}^{m}(\lambda_j / \sum_k \lambda_k)$
- Stata: `pca`, Matlab: `pca`, R/stats: `princomp`.

# Principal Components

How many components?

- Choose # of components by the eigenvalues or % of variance explained
- Common cutoffs are 90-95% of variance.
- Eigenvalue based cutoff rules (only take eigenvalues $> 1$).
- Most common method is to eyeball the scree plot.

# Principal Components



# Principal Components

Table 1 Principal Components

| Principal Component | Eigenvalue | Proportion of Variance | Cumulative Variance |
|---|---|---|---|
| 1 | 75 | 26.95% | 26.95% |
| 2 | 43 | 15.45% | 42.40% |
| 3 | 30 | 10.78% | 53.18% |
| 4 | 21 | 7.55% | 60.73% |
| 5 | 19 | 6.83% | 67.55% |
| 6 | 18 | 6.47% | 74.02% |
| 7 | 17 | 6.11% | 80.13% |
| 8 | 11 | 3.95% | 84.08% |
| 9 | 10 | 3.59% | 87.67% |
| 10 | 10 | 3.41% | 91.09% |
| 11 | 9 | 3.16% | 94.25% |
| 12 | 5 | 1.80% | 96.05% |
| 13 | 4 | 1.58% | 97.63% |

# Principal Components

- We can run a regression on principal components $Z$'s and then recover the betas of the $X$'s

$$\hat{y}^{pcr}_{(M)} = \overline{y}\mathbf{1} + \sum_{m=1}^{M} \hat{\theta}_m \mathbf{z}_m$$

- Because principal components are orthogonal we can find coefficients using univariate regression $\hat{\theta}_m = \langle \mathbf{z_m}, \mathbf{y} \rangle / \langle \mathbf{z_m}, \mathbf{z_m} \rangle$.
- We can recover the $x$ coefficients because the PCA is a linear transformation:

$$\hat{\beta}^{PCR} = \sum_{m=1}^{M} \hat{\theta}_m v_m$$

# Principal Components (and Ridge)

- If $M = P$ (we use all components) then PCR = OLS.
- If $M < P$ then we discard the $p - M$ smallest eigenvalue components
- This is similar to ridge which shrinks $\beta$'s for components with small eigenvalues.
- Think about the variance matrix $X'X/n$ or $X'X = VD^2V'$.
- First component (largest eigenvalue) is $\mathbf{z_1} = \mathbf{X}v_1 = \mathbf{u_1}d_1$.
- Variance is $Var(\mathbf{z_1}) = Var(\mathbf{X}v_1) = \frac{d_1^2}{N}$ ($\mathbf{z_1}$ is first principal component of $\mathbf{X}$).

# Principal Components

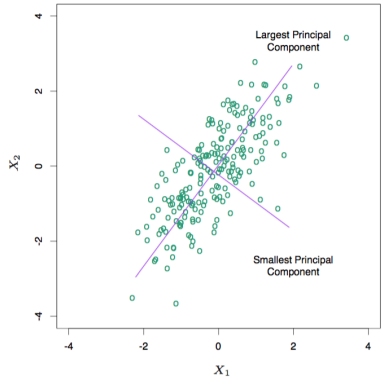



**FIGURE 3.9.** *Principal components of some input data points. The largest principal component is the direction that maximizes the variance of the projected data, and the smallest principal component minimizes that variance. Ridge regression projects* $\mathbf{y}$ *onto these components, and then shrinks the coefficients of the low–variance components more than the high-variance components.*

# Principal Components And Ridge

Consider the objective function that Ridge minimizes:

$$RSS(\lambda) = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)' + \lambda\beta'\beta$$

And the solution (which addresses multicolinearity!)

$$\hat{\beta}_{ridge} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$$

# Principal Components And Ridge

Just like we can diagonalize (some) square matrices, we can take the singular value decomposition (SVD) of any matrix $\mathbf{X}$ that is $N \times p$

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}'$$

- $\mathbf{U}, \mathbf{V}$ are $N \times p$ and $p \times p$ orthonormal matrices ($\mathbf{U}$ spans the column space, and $\mathbf{V}$ spans the row space of $X$.)
- $\mathbf{D}$ is a diagonal matrix $p \times p$ with elements corresponding to the singular values of $\mathbf{X}$.
- If $\mathbf{X}$ is a square, diagonalizable matrix the singular values are equal to the eigenvalues.

# Principal Components And Ridge

Now the least squares solution is simple

$$\mathbf{X}\hat{\beta}^{ols} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = UU'\mathbf{y}.$$

- $\mathbf{U}'\mathbf{y}$ are the values of $\mathbf{y}$ mapped into the orthonormal basis $\mathbf{U}$.
- This looks a lot like $\mathbf{QR}$ except that we have chosen a different basis.

Ridge is simple too:

$$\begin{aligned}\mathbf{X}\hat{\beta}^{ridge} = \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} &= \mathbf{U}\mathbf{D}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}'\mathbf{y} \\ &= \sum_{j=1}^{p} \mathbf{u}_j \frac{d_j^2}{d_j^2 + \lambda}\mathbf{u}_j'\mathbf{y}\end{aligned}$$

# Principal Components

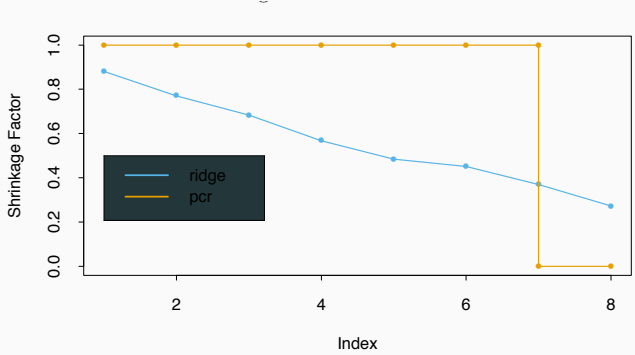



**FIGURE 3.17.** *Ridge regression shrinks the regression coefficients of the principal components, using shrinkage factors $d_j^2/(d_j^2 + \lambda)$ as in (3.47). Principal component regression truncates them. Shown are the shrinkage and truncation patterns corresponding to Figure 3.7, as a function of the principal component index.*

# Factor Models

- Related to PCA is the factor model
- These are frequently used in finance for asset pricing.

$$r_i = b_0 + b_1 f_1 + b_2 f_2 + \ldots b_p f_p + e_i$$

- Typically we choose factors so that $E[f_i] = 0$ and $E[f_i e_i] = 0$ and that $Cov(f_i, f_j) = 0$ for $i \neq j$.
- That is we choose scaled factors to form an orthogonal basis (which makes pricing assets easier).
- Instead of choosing $f$ to best explain $X'X$ we choose it to best explain $r$ by taking linear combinations of our $X$'s.
- CAPM is a single factor model (where the factor is the market return).
- Fama-French have expanded to a 5 factor model (book-to-market, market-cap, profitability, momentum)

# Factor Models: Other Examples

- *Eigenfaces* reduces your face to the first few eigenvalues – this is how face detection works!
- In psychometrics they use data from multiple tests to measure different forms of intelligence (mathematical reasoning, verbal, logical, spatial, etc.)
  - An old literature searched for general intelligence factor $g$
  - Nobody can tell what the GMAT measured!
- In marketing PCA/factor analyses are used in the construction of *perceptual maps*
- Marketing practitioners use FA/PCA more than academics these days (guess: maybe?)