Linear Methods for Regression: Methods using derived input directions

Statistical methods for data analysis – Machine learning

Alberto Castellini
University of Verona

Motivation

- When a **large number** of (correlated) **variables** X_j, j=1,…,p are available, they may be **linearly combined** in a **small number** of **components** (projections) Z_m, m=1,…,M, with $M \le p$.

- These **components** can be used as inputs in **regression**.

- Different **methods** are available for **constructing linear combinations** of variables
  - Principal components regression
  - Partial least squares

Motivation

- When a **large number** of (correlated) **variables** X_j, j=1,…,p are available, they may be **linearly combined** in a **small number** of **components** (projections) Z_m, m=1,…,M, with M<=p.

- These **components** can be used as inputs in **regression**.

- Different **methods** are available for **constructing linear combinations** of variables
  - Principal components regression
  - Partial least squares

Principal Component Regression (PCR)

Principal Components: SVD

Linear components Z_m are defined by **Principal Component Analysis** (PCA).

- Principal components (Karhunen-Loeve) directions of **X** are computed by **SVD** of **X** (**eigenvalue decomposition** of $\mathbf{X}^T\mathbf{X}$**,** if **X** is standardized).

- The **SVD** of the N x p matrix **X** can be written as:

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

  where:

  - U (N x p) and V (p x p) are **orthogonal** matrices
  - Columns of U span the **column space** of X
  - Columns of V span the **row space** of X
  - D is a p x p diagonal matrix with entries d1 >= d2 >= … >= dp >=0 **singular values** of X.

Principal Components: eigen decomposition

The **SVD** of the centered matrix X is another way of expressing the **principal components** of X.

In fact, the covariance matrix can be decomposed as

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

which is the **eigen decomposition** of $\mathbf{X}^T\mathbf{X}$.

- The **eigenvectors** v_j (columns of **V**) are also called **principal components** (Karhunen-Loeve) directions of **X**.

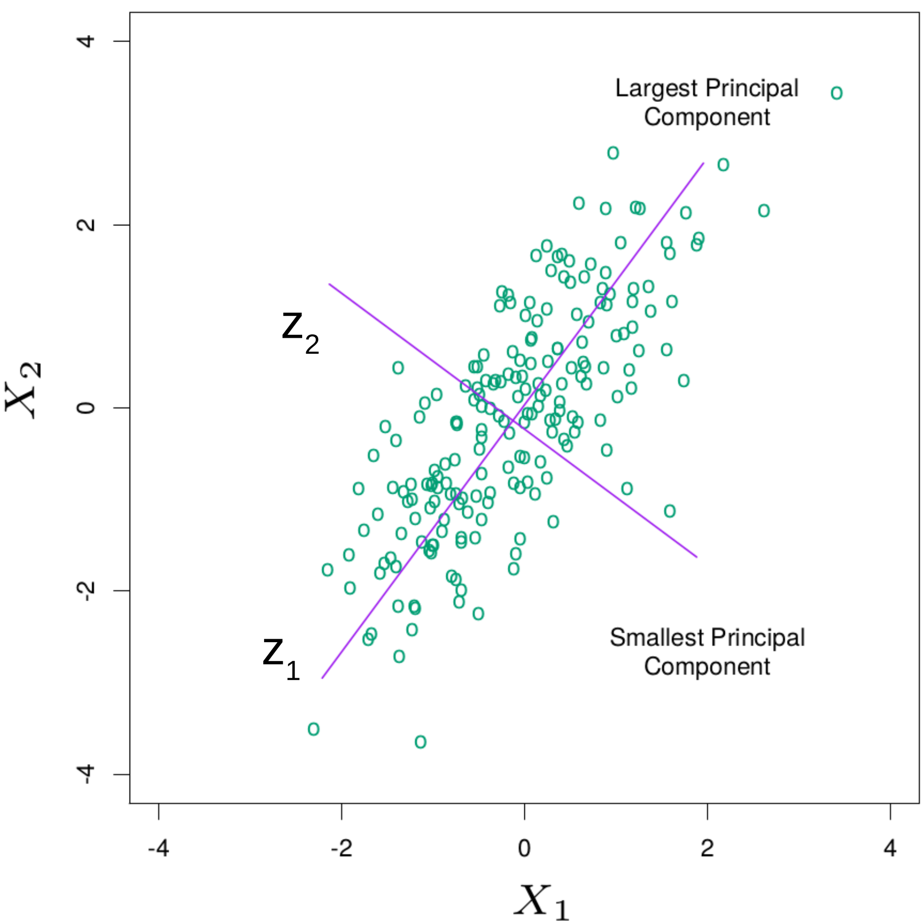
Principal Components: directions and variance

- The **first principal components** direction v_1 (**eigenvector** of $\mathbf{X}^T\mathbf{X}$) has the property that $\mathbf{z}_1 = \mathbf{X}^* v_1$ has the **largest sample variance** amongst all normalized linear combinations of columns of **X**

$$\mathrm{Var}(\mathbf{z}_1) = \mathrm{Var}(\mathbf{X}^* v_1) = d_1^2/N,$$

where d_1 is the eigenvalue of $\mathbf{X}^T\mathbf{X}$ with maximum absolute value and N is the total number of observations.

- **Subsequent principal components** $\mathbf{z}_j$ have **maximum variance** and are **orthogonal** to the earlier ones.



Principal Component Regression: parameter learning

Principal Component Regression forms the **derived input columns**

$$\mathbf{z}_m = \mathbf{X} * v_m$$

and then regresses **y** on $\mathbf{z}_1$, $\mathbf{z}_2$,…,$\mathbf{z}_M$, for some **M<=p**

- Since the z_m are **orthogonal**, this regression is a sum of univariate regressions:

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

Parameter on the m-th principal component

Inner product

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

- Since the z_m are linear combinations of the original x_j, the coefficients of variables $\mathbf{x}_j$ can be written as

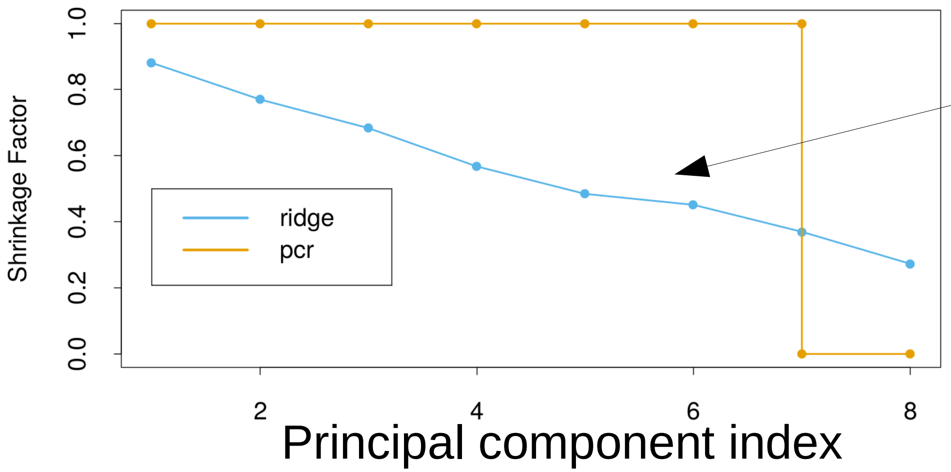
$$\hat{\beta}^{\text{pcr}}(M) = \sum_{m=1}^{M} \hat{\theta}_m v_m.$$

Observations

- Data **standardization** is needed (as in ridge regression) since principal components depend on variable scale.
- If **M=p** then PCR corresponds to OLS since the columns of **Z**=**UD** span the column space of **X**.

Similarities between ridge regression and PCR:

- **Both** operate on **principal components** of **X**
- **Ridge shrinks more** the components with **small eigenvalues** (directions with **smaller variance**)
- **PCR discards** the p-M **smallest eigenvalue** components



$$\frac{d_j^2}{d_j^2 + \lambda}$$

PCR on the prostate cancer dataset

Regression Coefficients

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

Exercise: Prediction on the prostate cancer dataset

See text of Exercise 5

References

[Hastie 2009] Trevor Hastie, Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (second edition). Springer. 2009.