PCA

Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Agenda

Motivation

- One way to think about almost everything we do is as dimension reduction.
- We are trying to
  - learn from high-dimensional X some low-dimensional summaries that contain the information necessary to make good decisions.
  - model it as having been generated from a small number of components/factors.
- We are attempting to simplify X for its own sake.

Motivation

- Unsupervised learning is often much more challenging.
- The exercise tends to be more subjective, and there is no simple goal for the analysis, such as prediction of a response.
- Unsupervised learning is often performed as part of an exploratory data analysis.
- Furthermore, it can be hard to assess the results obtained from unsupervised learning methods, since there is no universally accepted mechanism for performing cross- validation or validating results on an independent data set.
- There is no way to check our work because we don’t know the true answer: the problem is unsupervised.

Agenda

Principal Component Analysis

- PCA is an unsupervised learning technique that allows to
  - reduce the dimensionality of data sets,
  - while preserving as much "variability" as possible.
- It is an unsupervised approach, it involves only a set of variables/features $X_1, X_2, \ldots, X_p$, and no associated response Y .

Principal Component Analysis

- For example:
  1. Area
  2. Rooms
  3. Bathrooms
  4. Schools
  5. Crime

Principal Component Analysis

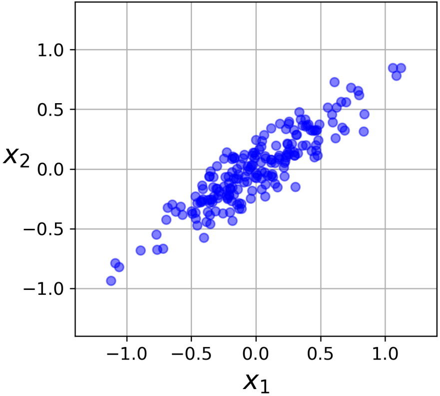
Area	Rooms	Bathrooms	Schools	Crime

PC1	PC2

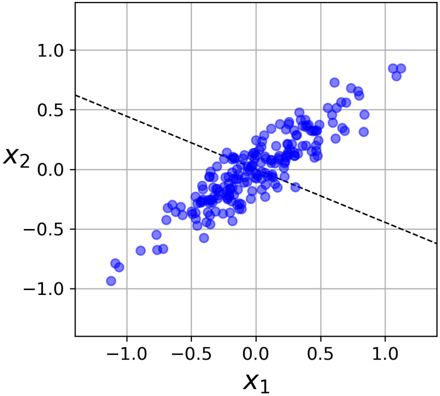
Principal Component Analysis

- PCA finds a low-dimensional representation of a data set that contains as much as possible of the variation.
- The idea is that each of the n observations lives in p-dimensional space, but not all of these dimensions are equally interesting.
- PCA seeks a small number of dimensions that are as interesting as possible, where the concept of interesting is measured by the amount that the observations vary along each dimension.

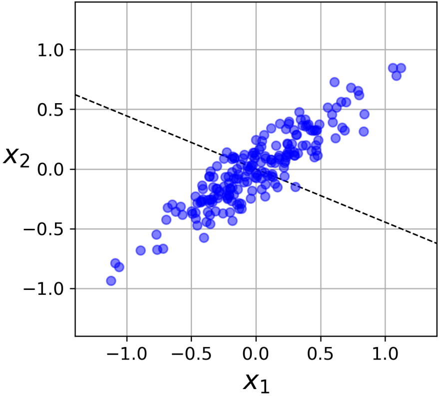
Principal Component Analysis

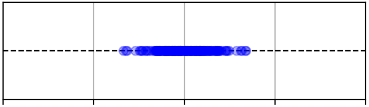


Principal Component Analysis

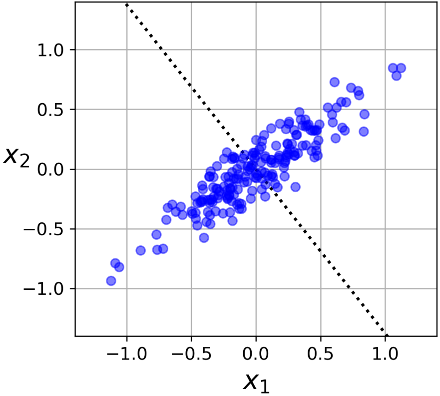


Principal Component Analysis

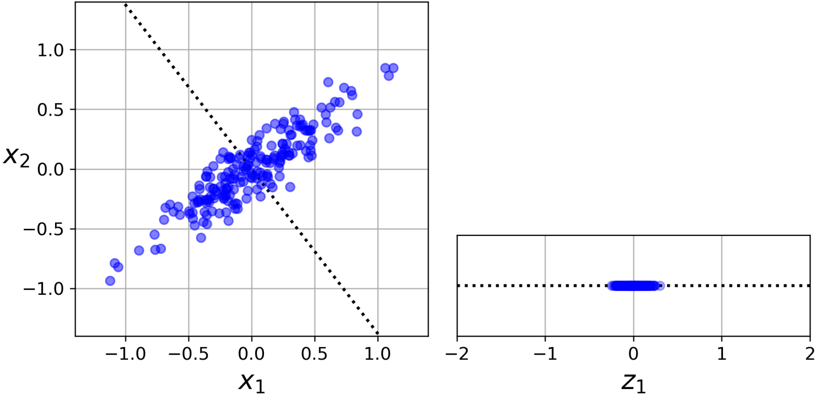




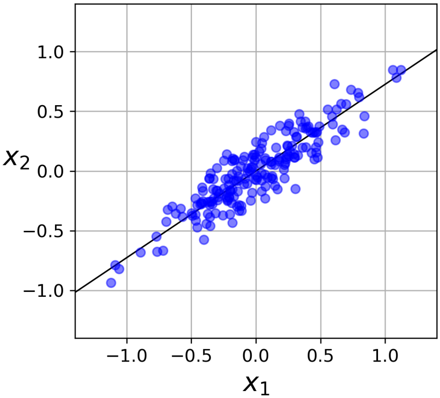
Principal Component Analysis



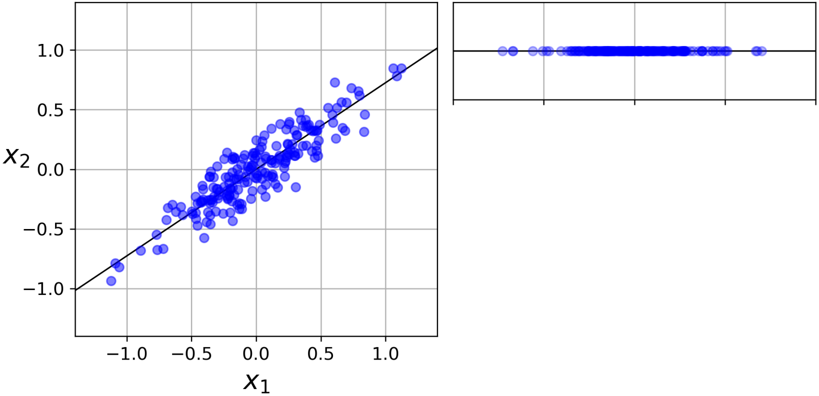
Principal Component Analysis



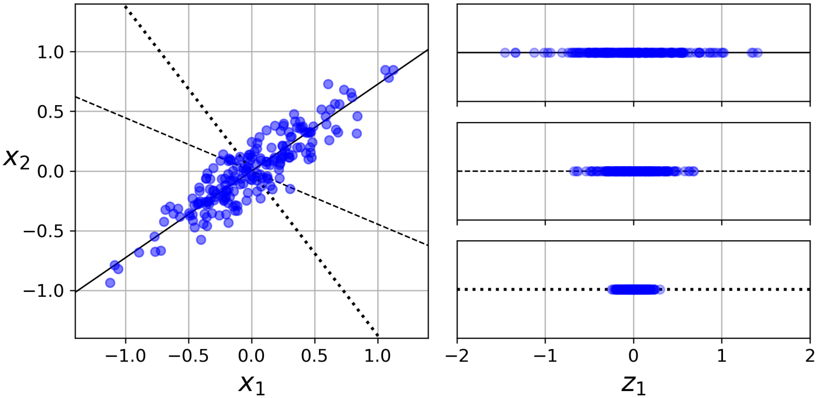
Principal Component Analysis



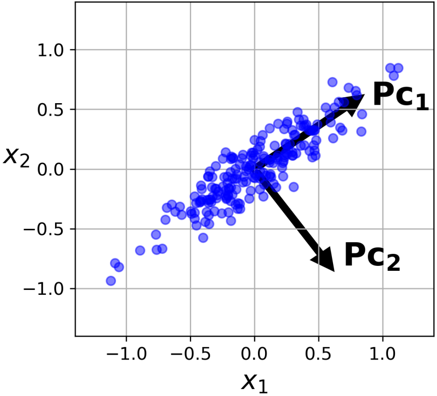
Principal Component Analysis



Principal Component Analysis



Principal Component Analysis



Principal Component Analysis

- The first principal component of a set of features $X_1, X_2, \ldots, X_p$ is the normalized linear combination of the features

$$PC_1 = \delta_{11} X_1 + \delta_{21} X_2 + \cdots + \delta_{p1} X_p \tag{1}$$

- The δ coefficients are called loadings or rotations—these are properties of the model and are shared across all observations.
- By normalized we mean that $\sum_{j=1}^{p} \delta_{j1}^2 = 1$

Principal Component Analysis

- The first principal component of a set of features $X_1, X_2, \ldots, X_p$ is the normalized linear combination of the features

$$PC_1 = \delta_{11} X_1 + \delta_{21} X_2 + \cdots + \delta_{p1} X_p \tag{1}$$

- The δ coefficients are called loadings or rotations—these are properties of the model and are shared across all observations.
- By normalized we mean that $\sum_{j=1}^{p} \delta_{j1}^2 = 1$
- In our example:

$$PC_1 = \delta_{11} X_1 + \delta_{21} X_2 \tag{2}$$

Principal Component Analysis

- The first principal component of a set of features $X_1, X_2, \dots, X_p$ is the normalized linear combination of the features

$$PC_1 = \delta_{11}X_1 + \delta_{21}X_2 + \cdots + \delta_{p1}X_p \tag{1}$$

- The δ coefficients are called loadings or rotations—these are properties of the model and are shared across all observations.
- By normalized we mean that $\sum_{j=1}^{p} \delta_{j1}^2 = 1$
- In our example:

$$PC_1 = \delta_{11}X_1 + \delta_{21}X_2 \tag{2}$$

- In our property values example:

$$PC_1 = \delta_{11}Area + \delta_{21}Rooms + \delta_{31}Bathrooms + \delta_{41}Schools + \delta_{51}Crime \tag{3}$$

How do we compute the first principal component

- Given a $n \times p$ data set X, how do we compute the first principal component?

Principal Component Analysis

- The problem then looks like

$$maxV(PC_1) = maxV(X\delta_1) \tag{6}$$

- where
  - $X = (x_1, \ldots, x_p)_{n \times p}$,
  - $S = V(X)$
  - δ_1 is $p \times 1$
- Let's set up the problem as

$$\underset{\delta}{max} \;\; \delta_1' V(X) \delta_1 \tag{7}$$

- What is the solution to this problem?

Principal Component Analysis

- The problem then looks like

$$maxV(PC_1) = maxV(X\delta_1) \tag{6}$$

- where
  - $X = (x_1, \ldots, x_p)_{n \times p}$,
  - $S = V(X)$
  - δ_1 is $p \times 1$
- Let's set up the problem as

$$\max_{\delta} \; \delta_1' V(X) \delta_1 \tag{7}$$

- What is the solution to this problem?
- Bring δ to infinity.

Principal Component Analysis

- Let's "fix" the problem by normalizing δ

$$\begin{aligned} \max_{\delta}\ & \delta_1' S \delta_1 \\ & \text{subject to} \\ & \delta_1 \delta_1' = 1 \end{aligned} \tag{8}$$

- Let us call the solution to this problem δ_1^*.
- $PC_1^* = X\delta_1^*$ is the 'best' linear combination of X.
- Intuition: X has P columns and $PC_1^* = X\delta_1^*$ has only one. The factor built with the first principal component is the best way to represent the P variables of X using a single single variable.

Detour: Algebra Review

- Let $A_{m\times m}$. It exists
  - a scalar λ such that $Ap = \lambda p$ for a vector $p_{m\times 1}$,
  - if $p \neq 0$, then λ is an eigenvalue of A.
  - and p is an eigenvector of A corresponding to the eigenvalue λ.
- $A_{m\times m}$ with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_m$, then:

$$tr(A) = \sum_{i=1}^{m} \lambda_i \tag{9}$$

$$det(A) = \Pi_{i=1}^{m} \lambda_i \tag{10}$$

- If $A_{m\times m}$ has m different eigenvalues, then the associated eigenvectors are all linearly independent.
- Spectral decomposition: $A = P\Lambda P'$

Detour: Algebra Review

- Spectral decomposition:

$$A = P\Lambda P' \tag{11}$$

- where $\Lambda = diag(\lambda_1, \dots \lambda_m)$ and P is the matrix whose columns are the corresponding eigenvectors.

$$A = \begin{pmatrix} p_1 & p_2 & \dots & \dots & p_m \end{pmatrix} \begin{pmatrix} \lambda_1 & & & & 0 \\ & \lambda_2 & & & \\ & & \ddots & & \\ & & & \ddots & \\ 0 & & & & \lambda_m \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ \vdots \\ p_m \end{pmatrix} \tag{12}$$

$$A = \sum_{i=1}^{m} \lambda_i p_i p_i' \tag{13}$$

Principal Component Analysis

- Solution to the problem of the first principal component

Principal Component Analysis

- Solution to the problem of the first principal component
- Let's set the lagrangian

$$\mathcal{L} = \delta_1' S \delta_1 + \lambda_1 (1 - \delta_1' \delta_1) \tag{14}$$

- Rearranging

$$S\delta_1 = \lambda_1 \delta_1 \tag{15}$$

- At the optimum, δ is the eigenvector corresponding to the eigenvalue λ of S.
- Premultiplying by δ_1' and remembering that $\delta_1' \delta_1 = 1$:

Principal Component Analysis

$$\delta_1 S \delta_1' = \lambda_1 \tag{16}$$

- In order to maximize $\delta' S \delta$ we must choose λ equal to the maximum eigenvalue of S and δ is the corresponding eigenvector.
- The problem of finding the best linear combination that reproduces the variability of X is finding the biggest eigenvalue of S and it's corresponding eigenvector

Principal Component Analysis

- The first main component? Are there others?
- Let's consider the following problem:

$$\max_{\delta_2} \ \delta_2' S \delta_2 \tag{17}$$

$$\text{st} \tag{18}$$

$$\delta_2'\delta_2 = 1 \tag{19}$$

$$\delta_2'\delta_1 = 0 \tag{20}$$

- $PC_2^* = X\delta_2^*$ is the second principal component : the best linear combination which is orthogonal to the best initial linear combination.
- Recursively, using this logic you can form q main components.
- Note that algebraically we could construct $q = p$ factors, actually the number of PC are $min(n-1, p)$

q main components

- Let $\lambda_1, \ldots, \lambda_p$ be the eigenvalues of $S = V(X)$, ordered from highest to lowest,
- $p_1, \ldots, p_p$ the corresponding eigenvectors.
- Call P the matrix of eigenvectors.
- Then $\delta_j = p_j$, $\forall j$ ('loadings' of the principal components =ordered eigenvectors of S).

Relative importance of factors

- Now we want to know the relative importance of factors, to have a way of choosing them

- Let $PC_j = X\delta_j$, $j = 1, \ldots, K$ be the j-th principal component.

$$V(PC_j) = \delta_j' S \delta_j \tag{21}$$
$$= p_j' P' \Lambda P p_j \tag{22}$$
$$= \lambda_j \tag{23}$$

(the variance of the j-th principal component is the j-th ordered eigenvalue of S).

- We this result we can show that the total variance of X is the sum of the variances of x_j , $j = 1, ..., p$, that is $trace(S)$

Relative importance of factors

- We the above result we can show that the total variance of X is the sum of the variances of x_j , $j = 1, ..., p$, that is $trace(S)$
- Note the following:

$$trace(S) = trace(P'\Lambda P) = trace(P'P\Lambda) = \sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{p} V(PC_j) \tag{24}$$

- Then

$$\frac{\lambda_p}{\sum_{j=1}^{p} \lambda_j} \tag{25}$$

- measures the relative importance of the jth principal component.

Selection of factors

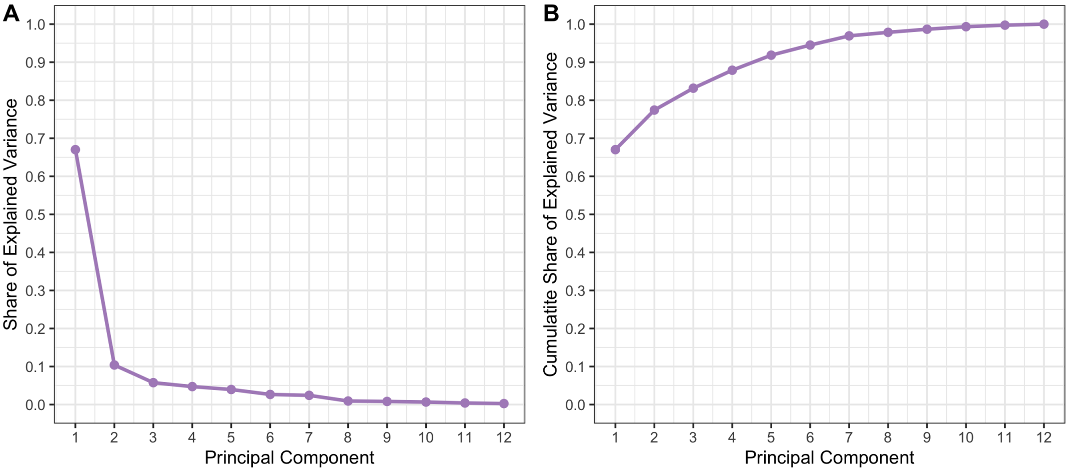
- Although a matrix X of dimension $n \times p$ generally has $min(n-1, p)$ different principal components.
- In practice, we are generally not interested in all the components, but rather stay with the first ones that allow us to visualize or interpret data.
- Indeed, we would like to keep the minimum number that allows us a good understanding of the data.
- The natural question that arises here is whether there is an established way to determine the number of principal components to use.
- Unfortunately, there is no accepted objective way in the literature to answer it.

Selection of factors

- However, there are three simple approaches that can guide you in deciding the number of relevant major components.
  - Visual examination of screeplot
  - Kaiser criterion.
  - Proportion of variance explained.

Selection of factors

Screeplot



Selection of factors

Kaiser criterion

- Let the columns of X be standardized, so that each variable has unit variance.
- In this case:

$$trace(S) = \sum_{j=1}^{p} V(PC_j) = p \tag{26}$$

- and recall $\sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{p} V(PC_j)$ then

$$\sum_{j=1}^{p} \lambda_j = p \tag{27}$$

- On average, each factor contributes one unit. When $\lambda_j > 1$, that factor it explains the total variance more than the average. $\rightarrow$ Retain the factors with $\lambda_j > 1$

Selection of factors

Proportion of variance explained

- Another approach often used in practice is to impose a threshold a priori and choose the main components based on it.
  - For example, we could define a threshold of 90%, which in the previous example plot would result in 5 main components.
  - Whereas if it were 70% we would have 2 main components.
- The threshold to be defined will depend on the application, the context, and the data set. Thresholds between 70% and 90% are typically used.

PC Computation

- Before I mentioned that data was standardized, that is, re-centered to have zero mean and scaled to have variance one.
- From a strictly mathematical point of view, there is nothing inherently wrong with making linear combinations of variables with different units of measurement.
- However, when we use PCA we seek to maximize variance and the variance is affected by the units of measurement.
- This implies that the principal components based on the covariance matrix S will change if the units of measure of one or more variables change.

PC Computation

- To prevent this from happening, it is common practice to standardize the variables. That is, each X value is re-centered and divided by the standard deviation:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{28}$$

- where $\bar{x}_j$ is the mean and s_j is the standard deviation of column j.
- Then the initial data matrix X is replaced by the standardized data matrix Z.
- Note also that when standardizing the data matrix, the covariance matrix S is simply the original data correlation matrix. This is sometimes referred to in the literature as the PCA correlation matrix.

PC Computation

Uniqueness of the main components

- It is necessary to warn that the ”loadings” of the main components δ are unique except for a sign change.
- This implies that depending on the implementation we can obtain different results from two libraries.
- The ”loadings” will be the same but the signs may differ.
- The signs may differ because each weight specifies a direction in k-dimensional space and the change of sign has no effect on the direction.

PC Computation

- As a practical aside, note that for really big sparse X, R will run out of memory.
- A big data strategy for PCA is to first calculate the covariance matrix for X and then obtain PC rotations as the eigenvalues of this covariance matrix.
  - The first step can be done using sparse matrix algebra.
  - The rotations are then available as

```
## eigen( xvar, symmetric = TRUE)$vec.
```

- There are also approximate PCA algorithms available for fast factorization on big data. See, for example, the `irlba` package for R.

Agenda

PC Interpretation

Caveat

- Component interpretation is hard because PCA focuses on variance, not meaning
- The technique optimally compresses information, but translating this compression into human-understandable concepts requires domain knowledge, simplifying assumptions, and statistical insight.
- As a result, many practitioners focus on explaining variance or ranking the importance of variables instead of finding exact meanings for each component.

Factor Model Interpretation

- Suppose we have p regressors and K=1

$$x_i = hf_i \tag{29}$$

- h is $p \times 1$
- f_i 1×1 and is the factor
- h are the factor loadings
- In this model, the factor f_i affects all regressors x_{ji}
- But the magnitude is specific to the regressor and captured by h

Factor Model Interpretation

Test Scores

$$x_i = hf_i \tag{30}$$

- x_i is a set of test scores for an individual student
- f_i is the student's latent ability
- h is how ability affects the different test scores
  - Some tests may be highly related to ability
  - Some tests may be less related
  - Some may be unrelated (random?)

Factor Model Interpretation

Test Scores

$$x_i = \sum_{m=1}^{k} h_m f_{mi} \tag{31}$$

- There are more than one form of ability
- i.e. literary and mathematical
- In labor economics, there has been hypothesized a distinction between cognitive and non-cognitive ability which has been very useful in explaining wage patterns (some jobs require one or the other, and some both (e.g. surgeon)

Factor Interpretation: Examples

Agenda

1. What are PCAs?
2. PC Interpretation
3. Principal Component Regression (PCR)

Principal Component Regression (PCR)

- Now that you've learned how to fit PCA models, what are they good for?
- In some settings, as in the previous political science example, the factors themselves have clear meaning and can be useful in their own right for understanding complex systems.
- More commonly, unfortunately, the factors are of dubious origin or interpretation.
- However, they can still be useful as inputs to a regression system.
- Indeed, this is the primary practical function for PCA, as the first stage of principal components regression (PCR).

Principal Component Regression (PCR)

- The concept of PCR is simple:
  - Instead of doing $y \to X$,
  - Use a lower-dimension set of principal components as covariates.
- This is a fruitful strategy for a few reasons:
  - PCA reduces dimension, which is usually good.
  - The PCs are independent, so you have no multicollinearity and the final regression is easy to fit.

Principal Component Regression (PCR)

- The disadvantage of PCR is that PCA will be driven by the dominant sources of variation in X.
- If the response is connected to these dominant sources of variation, PCR works well.
- If it is more of a “needle in the haystack response,” driven by a small number of inputs, then PCR will not work well.
- In practice, you do not know what scenario you are in

Principal Component Regression (PCR)

- How many PC do we use?
  - When PCA was used as a dimensionality reduction tool *per se* we had some guidelines...
- Should we do the same here?

Principal Component Regression (PCR)

- How many PC do we use?
  - When PCA was used as a dimensionality reduction tool *per se* we had some guidelines...
- Should we do the same here?
- In PCR the approach is slightly different
  - Construct $min(n-1, p)$ components
  - Use K fold crossvalidation adding 1 PC at a time
  - Choose the model with the lowest out of sample MSE
- Because the PCs are ordered (by their variance) and independent, this works better than subset selection on the raw dimensions of X_i.

Principal Component Regression (PCR)

- An alternative mechanism is run a lasso on the full set of PCs (works best in practice).
- This procedure makes it easy to incorporate other information in addition to the PCs.
- For example, one tactic that works well in practice is to put both PC and Xs into the lasso model matrix.
  - This then allows the regression to make use of the underlying factor structure in X and still pick up individual X_j signals that are related to y.
  - This hybrid strategy is a solution to the disadvantage of PCR mentioned earlier—that it will only pick up dominant sources of variation in X.

Principal Component Regression (PCR)

Summary of the steps

- Given a sample of regression input observations x_i, accompanied by output labels y_i for some subset of these observations:

  1. Fit PCA on the full set of X inputs to obtain PC of length $min(n-1, p)$.
  2. For the labeled subset, run a lasso regression for y on f (PC).
     - Alternatively, regress y on f and Xs to allow simultaneous selection between PCs and raw inputs.
  3. To predict for a new X_{new}, use the rotations from step 1 to get $f = \delta X_{new}$ and then feed these scores into the regression fit from step 2.

PCR Example