GI07/COMPM012: Mathematical Programming and Research Methods (Part 2)

2. Least Squares and Principal Components Analysis

Massimiliano Pontil

Today's plan

- SVD and principal component analysis (PCA)
- Connection between PCA and linear regression
- Low rank matrix approximation
- Application of SVD to least squares and ridge regression
- Generalized solution and pseudoinverse
- Role of the regularization parameter

Principal component analysis (PCA)

We are given data points $x_1, \ldots, x_n \in \mathbb{R}^d$ (training data)

Dimension reduction: we wish to find a lower dimensional representation of the data, ie. for visualization purposes, for cluster analysis, or as a preprocessing step in supervised learning

PCA is an instance of dimension reduction, which finds a k-dimensional subspace S of the "ambient" space $\mathbb{R}^d$, such that the projection on S retains most of the variance in the data

In PCA, the lower dimensional representation is a *linear* function of the input data

PCA optimization problem

For simplicity, we assume that the data points have zero mean: $\sum_{i=1}^{n} x_i = 0$ (otherwise subtract the mean)

Our goal is to maximize the variance of the projected data,

$$\text{var}(P) = \frac{1}{n} \sum_{i=1}^{n} \|Px_i\|^2$$

over the set of k-dimensional orthogonal projections P:

$$\max \left\{ \text{var}(P) : P \in \mathbb{R}^{d \times d},\ P^2 = P,\ P^\top = P,\ \text{rank}(P) = k \right\}$$

PCA optimization problem (cont.)

We write $P = QQ^\top$ where $Q = [q_1, \ldots, q_k]$ and the vectors $q_1, \ldots, q_k$ are o.n. (they form a basis for the subspace S we wish to project to)

We reformulate the above problem as an optimization problem in Q:

$$\max\left\{\frac{1}{n}\sum_{i=1}^{n}\|QQ^\top x_i\|^2 : Q \in \mathbb{R}^{d\times k}, Q^\top Q = I_{k\times k}\right\} \tag{1}$$

1-dimensional projection

When $k = 1$, $Q = q$ (a d-dimensional column vector). We have

$$\frac{1}{n}\sum_{i=1}^{n} \|qq^\top x_i\|^2 = \frac{1}{n}\sum_{i=1}^{n} (q^\top x_i)^2 = q^\top \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right) q = q^\top C q$$

where C is the data covariance: $C = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top$

We see that problem (1) is the same as maximizing the Rayleigh quotient

$$\frac{q^\top C q}{q^\top q}$$

whose solution is the leading eigenvector of the data covariance

Case $k > 1$

In the general case, similarly to the case $k = 1$, we derive that

$$\frac{1}{n}\sum_{i=1}^{n} \|Q^\top x_i\|^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}(q_j^\top x_i)(x_i^\top q_j) = \sum_{j=1}^{k} q_j^\top C q_j$$

The optimization problem (1) is now more difficult to analyze

We show that $q_1, \ldots, q_k$ are the k leading eigenvectors of C. They are also called the principal components of the data

Diagonal covariance

Suppose $C = \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_d)$ with $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$

$$\sum_{j=1}^{k} q_j^\top C q_j = \sum_{j=1}^{k} \sum_{\ell=1}^{d} q_{j\ell}^2 \lambda_\ell = \sum_{\ell=1}^{d} \lambda_\ell \sum_{j=1}^{k} q_{j\ell}^2$$

We will show that the maximum is attained at $q_1 = e_1, \ldots, q_k = e_k$

Proof: We use the fact that $\sum_{j=1}^{k} q_{j\ell}^2 \leq 1$ and $\sum_{\ell=1}^{d} \sum_{j=1}^{k} q_{j\ell}^2 \leq k$ (can you argue these inequalities are true?). These and $k \leq d$ give the upper bound

$$\sum_{\ell=1}^{d} \lambda_\ell \sum_{j=1}^{k} q_{j\ell}^2 \leq \sum_{\ell=1}^{k} \lambda_\ell$$

which is attained for $q_1 = e_1, \ldots, q_k = e_k$

General case

Let $C = V\Lambda V^\top$ and note that

$$\sum_{j=1}^{k} q_j^\top C q_j = \sum_{j=1}^{k} q_j^\top V\Lambda V^\top q_j$$

We can reduce this to the diagonal case by letting $\tilde{q}_j = V^\top q_j$

This transformation does not change the problem because V is orthogonal

We know that the solution is: $\tilde{q}_1 = e_1, \ldots, \tilde{q}_k = e_k$

We conclude that $q_j = Ve_j = v_j,\ j = 1, \ldots, k$

Connection to linear regression

We proceed to show that the principal components provide a sequence of best linear approximations to the data

Since

$$\|x\|^2 = \|(I - QQ^\top)x\|^2 + \|QQ^\top x\|^2$$

we see that maximizing the variance of the projected data is equivalent to minimizing

$$\sum_{i=1}^{n} \|(I - QQ^\top)x_i\|^2 \qquad (2)$$

ie. the variance associated with the complementary projection

Connection to linear regression (cont.)

The term under the summation in (2) can be interpreted as a linear regression:

$$\|(I - QQ^\top)x_i\|^2 = \min_{w_i} \|x_i - Qw_i\|^2$$

where the minimizing $w_i = Q^\top x_i$

Thus minimizing (2) is the same as minimizing

$$\sum_{i=1}^{n} \min_{w_i} \|x_i - Qw_i\|^2 \tag{3}$$

We conclude that PCA provides a sequence of best (over Q) linear approximations to the data

Summary

The k principal components represent a generic data point $x \in \mathbb{R}^d$ by the lower dimensional feature vector

$$w = V_k^\top x$$

where $V_k = [v_1, \ldots, v_k]$ is the matrix formed by the k leading eigenvectors of the training data covariance

The matrix V_k minimizes the reconstruction error (3) over all $k \times d$ orthogonal matrices

PCA as best low rank approximation

Denote by X the $n \times d$ matrix whose rows are the points $x_1^\top, \ldots, x_n^\top$

Recall the singular value decomposition (SVD) of X,

$$X = U \Sigma V^\top$$

where U and V are $n \times n$ and $d \times d$ orthogonal matrices, respectively, and Σ is the $n \times d$ diagonal matrix with diagonal entries $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d \geq 0$

At last, we show that PCA provides the best low rank matrix approximation of the data matrix X

PCA and best low rank approximation (cont.)

Recall the definition of the Frobenius norm and note that

$$\sum_{i=1}^{n} \|x_i - Qw_i\|^2 = \|X^\top - QW\|_F^2$$

where $W = [w_1, \ldots, w_n]$. The matrix QW has rank at most k

Hence the PCA problem is equivalent to

$$\min\{\|X - Z\| : \text{rank}(Z) \leq k\}$$

From the above discussion we conclude that the best rank k matrix approximation is: $Z = XV_kV_k^\top = U\Sigma V^\top V_kV_k^\top = U\Sigma V_k^\top$

Least squares

Problem: We wish to find a function $f(x) = w^\top x$ which best fits a data set $S = \{(x_1, y_1), \ldots, (x_n, y_n)\} \subseteq \mathbb{R}^d \times \mathbb{R}$

Assume that there exists some $w_* \in \mathbb{R}^d$ such that $y_i \approx w_*^\top x_i$

We find w by minimizing the residual sum of squares (RSS) on the data

$$R(w) = \sum_{i=1}^{n} \left(y_i - w^\top x_i\right)^2$$

To compute the minimum we need to solve the equations

$$\nabla R(w) = 0, \qquad \text{where } \nabla = \left(\frac{\partial}{\partial w_j}\right)_{j=1}^{d}$$

Normal equations

A direct computation gives the **linear system** of equations

$$\sum_{i=1}^{n} x_i x_i^\top w = \sum_{i=1}^{n} x_i y_i$$

or, in matrix notation

$$X^\top X w = X^\top y \tag{4}$$

where

$$X^\top = \begin{bmatrix} x_{11} & \cdots & x_{n1} \\ \vdots & \ddots & \vdots \\ x_{1d} & \cdots & x_{nd} \end{bmatrix} \equiv \left[x_1, \cdots, x_n \right], \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

(**Note:** we may also write $R(w) = \|y - Xw\|^2$ and differentiating over w to directly obtain equation (4))

Existence of solution

There always exists a solution to the normal equations (4)

To see this, we uniquely decompose y as $y = \bar{y} + y_\perp$ where $\bar{y} \in \text{range}(X)$, $y_\perp \in \text{range}(X)^\perp$ so that

$$R(w) = \|y_\perp\|^2 + \|\bar{y} - Xw\|^2$$

It follows that

$$\min R(w) = \|y_\perp\|^2$$

and the set of solutions is formed by the vectors w which interpolate $\bar{y}$:

$$Xw = \bar{y}$$

Overdetermined case ($n \geq d$)

If $n \geq d$ and X is full rank (ie. span$\{x_1, \ldots, x_n\} = \mathbb{R}^d$) then matrix $X^\top X$ is invertible and equation (4) has a **unique** solution:

$$w = (X^\top X)^{-1} X^\top y$$

In particular if $n = d$ then $w = X^{-1} y$
(Note: if $x_1, \ldots, x_n$ are in "generic positions" then rank$(X) = d$)

Two sub-cases:

- If $y \in$ range(X) then $y_\perp = 0$ and $\min R(w) = 0$ (perfect fit)

- If $y \notin$ range(X) then $y_\perp \neq 0$ and $\min R(w) > 0$

Underdetermined case ($n < d$)

If $n < d$ (or just rank$(X) < d$) then the solution is **not unique**. Again, we have two sub-cases:

- If $y \in$ range(X) we can interpolate the data: $\min R(w) = 0$ and any interpolant is a solution
- If $y \notin$ range(X) (e.g. this could be the case if $x_1 = x_2$ but $y_1 \neq y_2$) we cannot interpolate the data. As we saw above any vector w which interpolates $\bar{y}$ is a solution

Strict convexity of RSS

Another perspective: the function R is a convex quadratic function. To see this note that the Hessian of R at any vector w is the positive definite matrix $X^\top X$. Since R is lower bounded and grows at infinity, there is a minimum

- If $\text{rank}(X) = d$ then the $X^\top X$ is strictly positive definite. In this case the error function R is strictly convex, so the minimum is unique.

- If $\text{rank}(X) < d$ then R is not strictly convex and the minimum is not unique

These observations can be extended to a generic error function of the type $R(w) = \sum_{i=1}^n L(y_i, w^\top x_i)$, where L is a loss function

Statistical perspective

Fitting the data with a linear function makes especially sense if we know that the data has been generated by a linear function

$$y = Xw_* + \epsilon$$

where ϵ is some small noise error

We obtain that

$$\min R(w) = \epsilon^\top (I - P)\epsilon > 0$$

where $P = X(X^\top X)^{-1}X^\top$ is the orthogonal projection on range(X)

Ridge regression

In general, the problem of finding (learning) w from the data is *ill-posed*, i.e. at least one of the following conditions (which define a well-posed problem) is violated: (1) a solution exists; (2) the solution is unique; (3) the solution depends continuously on the data

We minimize the regularized error function

$$R_\lambda(w) := \sum_{i=1}^{n}(y_i - w^\top x_i)^2 + \lambda \sum_{\ell=1}^{d} w_\ell^2 \equiv (y - Xw)^\top(y - Xw) + \lambda w^\top w$$

The positive parameter λ defines a trade-off between the error on the data and the norm of the vector w (degree of regularization)

The objective function is now strictly convex. There is a unique minimum, which depends continuously on the data X and y

Ridge regression (cont.)

Setting $\nabla R_\lambda(w) = 0$, we obtain the modified normal equations

$$X^\top (Xw - y) + \lambda w = 0$$

whose solution (called *regularized solution*) is

$$w = (X^\top X + \lambda I)^{-1} X^\top y$$

It is interesting to analyze how this solution depends on λ and study how to choose this parameter in practice (we come back to this point later)

Singular value decomposition

We use the singular value decomposition (SVD) of X,

$$X = U\Sigma V^\top$$

where recall U and V are $n \times n$ and $d \times d$ orthogonal matrices, respectively, and Σ is the $n \times d$ matrix with leading diagonal entries $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d \geq 0$

We have

$$(X^\top X + \lambda I)w = X^\top y \iff V(\Sigma^2 + \lambda I)V^\top w = V\Sigma U^\top y$$

from which we obtain the solution

$$w_\lambda = V(\Sigma^2 + \lambda I)^{-2}\Sigma U^\top y \quad (5)$$

Generalized solution

In vector notations (5) becomes

$$w_\lambda = \sum_{i=1}^{r} \frac{\sigma_i}{\sigma_i^2 + \lambda} v_i u_i^\top y$$

where $r = \text{rank}(X)$, $U = [u_1, \ldots, u_n]$, $V = [v_1, \ldots, v_d]$

When λ goes to zero w tends to the **generalized solution**

$$w^{(0)} := \sum_{i=1}^{r} \sigma_i^{-1} v_i u_i^\top y = X^+ y \qquad (6)$$

Matrix $X^+ = \sum_{i=1}^{r} \sigma_i^{-1} u_i v_i^\top$ is called the pseudoinverse of X

If $n = d$ and X is full rank then $X^+ = X^{-1}$

Interpretation of $w^{(0)}$

We saw before than if rank$(X) < d$ then the RSS has not a unique minimum

The solution set is given by $\{w : X^\top X w = X^\top y\}$

The generalized solution $w^{(0)}$ is the vector which, among those which minimize $R(w)$ has the smallest norm