MATH-412 Statistical Machine Learning

A few elements of algebra around SVD and PCA

Lecturer: Guillaume Obozinski

A key result in matrix algebra

Consider two matrices $A \in \mathbb{R}^{n \times K}$ and $B \in \mathbb{R}^{p \times K}$. Let $\boldsymbol{a}_k$ and $\boldsymbol{b}_k$ denote the kth column respectively of A and B. One simple, but key result is that we have

$$\boxed{AB^\top = \sum_{k=1}^{K} \boldsymbol{a}_k \boldsymbol{b}_k^\top} \qquad (*)$$

The simplest way to prove this result is to note that $A = \sum_{k=1}^{K} \boldsymbol{a}_k \boldsymbol{e}_k^\top$ and $B = \sum_{k=1}^{K} \boldsymbol{b}_k \boldsymbol{e}_k^\top$ where $\boldsymbol{e}_k \in \{0,1\}^K$ is the kth element of the canonical basis. We then have

$$AB^\top = \sum_{j=1}^{K} \boldsymbol{a}_j \boldsymbol{e}_j^\top \sum_{k=1}^{K} \boldsymbol{e}_k \boldsymbol{b}_k^\top = \sum_{j=1}^{K} \sum_{k=1}^{K} \boldsymbol{a}_j (\boldsymbol{e}_j^\top \boldsymbol{e}_k) \boldsymbol{b}_k^\top,$$

and hence the result since $\boldsymbol{e}_j^\top \boldsymbol{e}_k = \delta_{j,k}$.

A few applications of this:

The empirical covariance matrix

It is of the form $\quad \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top.$

So a direct application of $(*)$ shows that

- if $\bar{\mathbf{x}} = 0$, we have $\hat{\Sigma} = \frac{1}{n} X^\top X$ where $X \in \mathbb{R}^{n \times p}$ is the design matrix;
- if the data is not centered then $\widetilde{X} = (I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top)X$ is the centered data matrix and we have
$$\hat{\Sigma} = \tfrac{1}{n} \widetilde{X}^\top \widetilde{X} = \tfrac{1}{n} X^\top (I_n - \tfrac{1}{n}\mathbf{1}\mathbf{1}^\top) X,$$
given that $(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top)$ is symmetric and idempotent.

Orthogonal projectors

If $U \in \mathbb{R}^{p \times K}$ is a matrix of orthonormal columns then the projection on the subspace spanned by the $\boldsymbol{u}_i$s, i.e. the columns of U is a linear transformation whose matrix is $UU^\top$. The fact that $UU^\top = \sum_{k=1}^{K} \boldsymbol{u}_k \boldsymbol{u}_k^\top$ is true because of $(*)$ regardless of the properties of U, but here

$\boldsymbol{u}_k\boldsymbol{u}_k^\top$ is the projector on the span of $\boldsymbol{u}_k$. So, in this case, the identity has the interpretation that the sum of the projections on the $\boldsymbol{u}_k$s are equal to the projection on the subspace spanned by all of them. This is of course the main property that we seek in an orthonormal basis.

Singular value decomposition

The theory of the SVD says that any matrix $X \in \mathbb{R}^{n\times p}$ admits a decomposition, called the reduced SVD, of the form $X = USV^\top$ where $U \in \mathbb{R}^{n\times K}, V \in \mathbb{R}^{p\times K}, \quad U^\top U = V^\top V = I_K$ and S is diagonal with coefficients $s_k > 0$. It is easy to see that US is the matrix whose columns are the $s_k\boldsymbol{u}_k$ and so, using $(*)$, we have

$$USV^\top = \sum_{k=1}^{K} s_k\boldsymbol{u}_k\boldsymbol{v}_k^\top.$$

Note that the same formula can be established for the full SVD (cf the slides of the lecture on PCA).

Projection of the data on the principal directions

For the interpretations to be correct, we assume from now on that the data is centered.

With the previous notations, the *principal directions* are the $\boldsymbol{v}_k$s, and they are sorted in decreasing order of the s_ks. The simple situation is the case where $s_1 > s_2 > \ldots > s_K > 0$. Let's discuss it first.

If $V_{[k]} \in \mathbb{R}^{p\times k}$ is the matrix formed by the k first columns of V, by definition its columns form an orthonormal basis of the *principal subspace of dimension* k. The projector on that subspace is $V_{[k]}V_{[k]}^\top$. The projection of $\mathbf{x}_i$ is therefore $V_{[k]}V_{[k]}^\top\mathbf{x}_i = \sum_{j=1}^k \boldsymbol{v}_k\boldsymbol{v}_k^\top\mathbf{x}_i \in \mathbb{R}^p$. Note that the projection is still a vector in the same space as $\mathbf{x}_i$. To compute the projection of all datapoints at once we can do the same calculation on the design matrix

$$XV_{[k]}V_{[k]}^\top = USV^\top V_{[k]}V_{[k]}^\top = \sum_{j=1}^{K} s_j\boldsymbol{u}_j\boldsymbol{v}_j^\top \sum_{\ell=1}^{k} \boldsymbol{v}_\ell\boldsymbol{v}_\ell^\top = \sum_{j=1}^{k} s_j\boldsymbol{u}_j\boldsymbol{v}_j^\top = U_{[k]}S_{[k]}V_{[k]}^\top,$$

because again $\boldsymbol{v}_j^\top\boldsymbol{v}_\ell = \delta_{j,\ell}$.

The situation is more complicated if $s_i = s_{i+1}$ because in that case the subspace associated with s_i is of dimension $d_i > 1$, and the *principal directions* are not unique because any orthonormal basis of that subspace is acceptable. If the subspace associated with a singular value is of dimension greater than 1 it makes sense to consider this subspace as a whole, when projecting the data, i.e. we would prefer to only consider the projections with $V_{[k]}V_{[k]}^\top$ for k such that $s_k > s_{k+1}$. That way the projections computed depend only on the subspaces and not on the choice of any particular basis. Apart from that, all the algebra above stays the same.

Principal components

As before, we assume that the data is centered.

Imagine that the data lives in $\mathbb{R}^3$ and that we project the data on a generic plane going through the origin. If we think of the plane as a sheet of paper, and if we want to have the best view of the projection on the sheet of paper, it makes sense to put the sheet of paper flat on the table and to look at it from top. This "rotation" or *change of basis* in space corresponds to keeping the coordinates of the projection of a datapoint $\mathbf{x}_i$ in the basis $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$ but to use these k coordinate in the canonical basis. The j th coordinates is $\boldsymbol{v}_j^\top \mathbf{x}_i$ which is the component of the projection of $\mathbf{x}_i$ on $\boldsymbol{v}_j$. The set of coordinates $(\boldsymbol{v}_j^\top \mathbf{x}_i)_{j=1}^p$ are called the *principal components.*

For a datapoint $\mathbf{x}_i$ the vector formed by its k first principal components is

$$V_{[k]}^\top \mathbf{x}_i \in \mathbb{R}^k.$$

Note that this is now a vector of dimension k. If the goal is to visualise the data we typically consider $k = 2$ and plot $\boldsymbol{v}_1^\top \mathbf{x}_i$ vs $\boldsymbol{v}_2^\top \mathbf{x}_i$ for all i as a scatter plot.

Now, exactly as before we can do this for the whole matrix at a time and we obtain

$$XV_{[k]} = U_{[k]} S_{[k]},$$

with the same calculations as for the projections.

Remarks:

- This shows that the principal components of $\mathbf{x}_i$ can be read on the ith row of $U_{[k]} S_{[k]}$.
- This also shows that the principal components can be computed directly from $\hat{\Sigma}$ whose eigenvalue decomposition is $\hat{\Sigma} = \frac{1}{n} U_{[k]} S_{[k]}^2 U_{[k]}$, although the principal components are inaccessible from $\hat{\Sigma}$.

Principal variables

As before, we assume that the data is centered.

If we look at the design matrix X from the point of view of the variables and not of the datapoints, then each variable corresponds to a column of X which lives in $\mathbb{R}^n$. With this perspective, $X\boldsymbol{v}_k \in \mathbb{R}^n$, which is the vector of values of the kth principal component for all the datapoints, defines a new variable from all the others. But $X\boldsymbol{v}_k = s_k \boldsymbol{u}_k$, so again the SVD provides a direct representation of this new variable.

Furthermore, as we often like to have variables that are normalized, we can divide the column by its standard deviation, which turns out to be $s_k/\sqrt{n}$, and since $\boldsymbol{u}_k$ is a unit vector $\sqrt{n}\boldsymbol{u}_k$ is of variance 1. We therefore see that, up of a constant factor $\sqrt{n}$, the $\boldsymbol{u}_k$s can be interpreted as a collection of new variables that are all perfectly decorrelated from each other empirically (since they are orthogonal) and such that $\boldsymbol{u}_1$ captures the largest amount of information about all the variables, then $\boldsymbol{u}_2$, etc.

In particular, U can be viewed as an interesting orthogonal basis for the original set of variables, on which we can project the variables. The *principal variables* are to the variables (colums of X) what the principal directions are to the datapoints (rows of X) ! In particular if we would like to study the correlation structure between the variables we can project all the columms of X on $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$. Correlated pairs of variables tend to be closeby in this representation. This representation is called the *circle of correlations* because if all variables are normalized appropriately their all fall in 2D in a circle of radius 1 (after dividing by $\sqrt{n}$).