Statistical Genetics, Spring 2024

Class notes 6: PCA

Principal component analysis — PCA and its use for stratification correction

General PCA idea: Given a data matrix $X_{n\times p}$, we can consider it as n points in $\mathbb{R}^p$. Assume the columns are centered: $1_n^T X = 0_p^T$, and consider a unit vector $v \in \mathbb{R}^p,\ \|v\|_2 = 1$. Then

$$\frac{1}{n}\|Xv\|_2^2 = \hat{Var}(Xv),$$

is the "spread" of the data in the direction v.

PCA seeks to find directions of maximum spread, the first PC direction is defined as:

$$\begin{aligned} v_1 &= \arg\max_v \|Xv\| \\ &\quad \text{s.t. } \|v\| = 1, \end{aligned}$$

the direction of maximum spread. The first PC projection is Xv_1.

The second PC continues this thinking:

$$\begin{aligned} v_2 &= \arg\max_v \|Xv\| \\ &\quad \text{s.t. } \|v\| = 1\ ,\ \ v \perp v_1, \end{aligned}$$

the next direction of maximum spread, orthogonal to v_1 — we can think that we eliminated v_1 from the linear space and now look for the best direction, and so only for $v_3, v_4, \ldots$.

Finding PCs is based on the SVD of $X = UDV^T$, where:

- $U_{n\times r}$ has orthonormal columns, with $r = \min(n, p)$
- $D = diag(d_1, ..., d_r)$ has non-negative elements, and we assume WLOG $d_1 \geq d_2 \geq \ldots \geq d_r$, by permuting columns of U, V
- $V_{p\times r}$ also has orthonormal columns

Every matrix X can be written in this way. We may also derive from this the eigendecompositions of $X^T X$ or XX^T :

$$X^T X_{p\times p} = VD^2V^T\ ,\ \ (XX^T)^{n\times n} = UD^2U^T.$$

The well known result, which is also easy to prove is that with this SVD representation: $v_1 = V_{\downarrow 1}$, the first column of V, $v_2 = V_{\downarrow 2}$, etc.

Proof for v_1 : Assume for simplicity of notation that $p < n$, so $r = p$. Then $V_{\downarrow 1}, \ldots, V_{\downarrow p}$ is a basis of $\mathbb{R}^p$. For a vector $v \in \mathbb{R}^p$, $\|v\|_2 = 1$, write it in this basis:

$$v = \sum_{j=1}^{p} \alpha_j V_{\downarrow j}, \ \|\alpha\|_2 = 1.$$

Then we can write:

$$Xv = UDV^T \left(\sum_{j=1}^{p} \alpha_j V_{\downarrow j} \right) = \sum_{j=1}^{p} \alpha_j UD \left(V^T V_{\downarrow j} \right) = \sum_{j=1}^{p} \alpha_j U_{\downarrow j} d_j.$$

Now the norm gives us:

$$\|Xv\|_2^2 = \| \sum_{j=1}^{p} \alpha_j U_{\downarrow j} d_j \|_2^2 = \sum_{j} (\alpha_j^2 d_j^2) \leq d_1$$

but we know that $v = V_{\downarrow 1}$ attains equality with $\alpha_1 = 0, \alpha_2 = \ldots = \alpha_p = 0$. ■

Some notes:

1. Note that because of orthogonality we have $X = \sum_{j=1}^{r} d_j U_{\downarrow j} V_{\downarrow j}^T$. This means we can use the PCs to get the **best low-rank approximation** of X, since for $q < r$, $Y = \sum_{j=1}^{q} d_j U_{\downarrow j} V_{\downarrow j}^T$ is immediately seen to be the solution of:

$$Y = \arg\min \|X - Y\|_{Fro}, \quad s.t.\ rank(Y) \leq q.$$

2. Simple linear algebra gives us that the total "variance" of the data is the trace of D^2

$$\sum_{ij} X_{ij}^2 = tr(XX^T) = tr(X^T X) = tr(VD^2 V^T) = tr(D^2) = d_1^2 + \ldots + d_r^2.$$

This means that d_j^2 **can be interpreted as the % of variance explained by the** j**th PC**.

Using PCA to model stratification in genetics

We expect (and indeed see) that the ethnic structure and stratification aspects will come out as top PCs in our data. This can be used in several ways:

1. The first or several first PC's can be used as covariates in the regression, like the output of the EM

2. Remove the top PCs from the matrix X before doing the testing:

$$\tilde{X} = X - d_1 U_{\downarrow 1} V_{\downarrow 1}^T - \ldots - d_k U_{\downarrow k} V_{\downarrow k}^T,$$

for some properly chosen k. This is equivalent to regressing the PCs out of the X's

3. It is common to also regress the PCs out of the response Y, which is basically equivalent to adding the PCs as covariates to the regression, in terms of both the estimates it gives for non-PC variables, specifically SNPs, and the t-based inference it gives in linear regression (why?)