SDS 387, Fall 2024
Homework 3

Due October 17, by midnight on Canvas.

1. The **Delta Method** is a method to derive the asymptotic distribution of a function of a random vector converging in distribution to a Gaussian. It is a consequence of the CLT. Formally, let $\mathbb{R}^d \to \mathbb{R}$ be a function continuously differentiable at a point μ on its domain and let $\{X_n\}$ be a sequence of random vectors such that

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N_d(0, \Sigma).$$

Show that

$$\sqrt{n}(f(\overline{X}_n) - f(\mu)) \xrightarrow{d} N(0, \nabla f(\mu)^\top \Sigma \nabla f(\mu)),$$

where $\nabla f(\mu))$ denotes the gradient of f evaluated at μ. This result is referred to as the delta method. *Hint: Do a first-order Taylor series expansion.*

Following the hint and performing a first-order Taylor series expansion of $f(\overline{X}_n)$ around $f(\mu)$ with the remainder in integral form, we have, using the linearity of the inner product,

$$\begin{aligned}\sqrt{n}\left(f(\overline{X}_n) - f(\mu)\right) &= \int_0^1 \langle \nabla f(\mu + u(\overline{X}_n - \mu)), \sqrt{n}(\overline{X}_n - \mu)\rangle du \\ &= \langle \int_0^1 \nabla f(\mu + u(\overline{X}_n - \mu)) du, \sqrt{n}(\overline{X}_n - \mu)\rangle.\end{aligned}$$

Then, $\int_0^1 \nabla f(\mu + u(\overline{X}_n - \mu)) du \xrightarrow{p} \nabla f(\mu)$ (it is ok to bring the limit inside the integral) and $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{p} X$, where $X \sim N_d(0, \Sigma)$ by assumption. By Slutsky's theorem, the term on the left-hand side of the above display equation converges to $\langle \nabla f(\mu), X\rangle \sim (0, \nabla f(\mu)^\top \Sigma \nabla f(\mu))$.

2. The delta method is not very useful when $\nabla f(\mu) = 0$. Here is a one-dimensional example. Suppose that $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$ and let $f(x) = x^2$. Show that $\sqrt{n}(\overline{X}_n^2 - \mu^2) \xrightarrow{d} N(0, 4\mu^2\sigma^2)$. If $\mu = 0$ the result implies that $\sqrt{n}(\overline{X}_n^2 - \mu^2) \xrightarrow{p} 0$. To obtain a limiting distribution, we need to consider a higher-order Taylor series expansion. Show that

$$n\overline{X}_n^2 \xrightarrow{d} \sigma^2 \chi_1^2$$

Hint: perform a second order Taylor series expansion and use the fact that if $X \sim N(\gamma, \sigma^2)$, then $X^2 \sim \sigma^2 \chi_1^2(\gamma^2)$.

If $g'(\mu) = 0$, we can take a second-order Taylor series expansion as follows. Suppose that $a_n(X_n - \mu) \xrightarrow{d} X$. Then for any g, continuously twice differentiable at μ, with $g''(\mu) \neq 0$,

$$g(X_n) - g(\mu) = \frac{1}{2}(X_n - \mu)^2 g''(\mu) + R(X_n - \mu),$$

where $R(x) = o(x^2)$. Multiply by a_n^2 on both sides. Then, by the continuous mapping theorem, $\frac{1}{2}a_n^2(X_n-\mu)^2 g''(\mu) \xrightarrow{d} \frac{1}{2}X^2 g''(\mu)$. As for the term $a_n^2 R(X_n-\mu)$, Lemma 2.12 in Van der Vaart's books imply that $a_n^2 R(X_n-\mu) = o_p(a_n^2(X_n-\mu)^2) = o_p(O_p(1)) = o_p(1)$ because $a_n^2(X_n-\mu)^2 = O_P(1)$ by the continuous mapping theorem. Thus, by Slutsky's theorem,

$$a_n^2(g(X_n) - g(\mu)) \xrightarrow{d} \frac{1}{2}X^2 g''(\mu).$$

When $a_n = \sqrt{n}$, $X_n = \overline{X}_n$ and $X \sim N(0, \sigma^2)$ (with $g'(\mu) = 0$)

$$\frac{g''(\mu)}{2} n(g(\overline{X}_n) - g(\mu)) \xrightarrow{d} \frac{g''(\mu)}{2}\sigma^2\chi_1(\mu^2).$$

The result follows for $g(x) = x^2$ with $\mu = 0$.

3. Let A be a symmetric matrix with eigendecomposition $A = U\Lambda U^\top$.

   (a) Show that, for any positive integer k

   $$A^k = U\Lambda^k U^\top$$

   and, provided that A is non-singular,

   $$A^{-k} = U\Lambda^{-k}U^\top.$$

   (If A is singular, not all hopes are lost: we would use a pseudo-inverse. But that is a topic for another homework.)
   We have

   $$\begin{aligned} A^k &= \underbrace{U\Lambda U^\top \; U\Lambda U^\top \; \ldots \; U\Lambda U^\top}_{\text{k times}} \\ &= U \underbrace{\Lambda\Lambda\ldots\Lambda}_{\text{k times}} U^\top \\ &= U\Lambda^k U^\top, \end{aligned}$$

   because $UU^\top = I$. The same argument works for the inverse A^{-1}, noting that $A^{-1} = U\Lambda^{-1}U^\top$.

   (b) The matrix exponent of a symmetric matrix A is

   $$e^A = \sum_{i=0}^{\infty} \frac{A^i}{i!}.$$

   Let $A = U\Lambda U^\top$ be the eigendecomposition of A. Show that

   $$e^A = Ue^\Lambda U^\top,$$

where e^Λ is the diagonal matrix with diagonal elements $e^{\lambda_1}, \ldots, e^{\lambda_n}$, where the λ_i'a are the eigenvalues of A.
Using the result above, we can write

$$e^A = \lim_{n\to\infty} \sum_{i=0}^{n} \frac{A^i}{i!} = \lim_{n\to\infty} U\Lambda_n U^\top = U\left(\lim_{n\to\infty} \Lambda_n\right) U^\top = Ue^\Lambda U^\top,$$

where Λ_n is the diagonal matrix whose (i,i) entry is $\sum_{i=0}^{n} \frac{\lambda^i}{i!}$.

4. Let Σ be the covariance matrix of a n-dimensional random vector X that has mean zero. If Σ has rank $r < n$, show that X takes values on a r-dimensional linear subspace (in fact, hyperplane) and finds that subspace (in fact, hyperplane).

   Because Σ is positive semidefinite, it admits the eigendecomposition $\Sigma = U_r\Lambda_r U_r^\top$, where U_r is a $n \times r$ matrix with orthonormal columns and Λ_r is a $r \times r$ diagonal matrix containing the r positive eigenvalues of Σ. Then X is supported on the r-dimensional linear subspace $\mathcal{S}$ of $\mathbb{R}^n$ spanned by the columns of U_k. If not, then there must exist a point x_0 and a $\epsilon > 0$ such that the ball $B(x_0, \epsilon) = \{x \colon \|x - x_0\| < \epsilon\}$ does not intersect $\mathcal{S}$ and $\mathbb{P}(X \in B(x_0, \epsilon)) > 0$. Then, for some unit vector v in $\mathcal{S}^\perp$, $\mathrm{Var}(v^\top X) > 0$. But this is impossible, since $\mathrm{Var}(v^\top X) = v^\top \Sigma v = 0$. Thus, X is supported on $\mathcal{S}$ (technically, this means that there exists a closed - in the subspace topology - subset of $\mathcal{S}$, say S, such that $\mathbb{P}(X \in S) = 1$).

5. Let A be a $m \times n$ matrix with SVD $U\Sigma V^\top$. Suppose we want to approximate it using a rank $r < \min\{m, n\}$ matrix. We measure the quality of the approximation by the squared Frobenius norm, i.e., we want to find a rank-r $m \times n$ matrix B such that the least squares error

   $$\|A - B\|_F^2$$

   is minimal. Find a B such that

   $$\|A - B\|_F^2 = \sum_{i>r} \sigma_i^2,$$

   where the σ_i's are the singular values of A (in decreasing order). In fact, that is the best we can do, a result known as the Eckart-Young-Mirsky theorem.

   Let $A = U\Sigma V^\top$ be the SVD of A, and set $B = U_r\Sigma_r V_r^\top$, where U_r and V_r are the submatrices of U and V containing the first r columns and Σ_r is the principal $r \times r$ submatrix of Σ. Accordingly, let U_{-r} the submatrix of U obtained by removing the first r columns of U and Σ_{-r} the diagonal matrix containing the trailing $\mathrm{rank}(A) - r$ singular values of A, in decreasing order. Then, letting $C = A - B = U_{-r}\Sigma_{-r}U_{-r}^\top$,

   $$\|A - B\|_F^2 = \mathrm{tr}(CC^\top) = \mathrm{tr}(U_{-r}\Sigma_{-r}^2 U_{-r}^\top) = \mathrm{tr}(U_{-r}^\top U_{-r}\Sigma_{-r}^2) = \mathrm{tr}(\Sigma_{-r}^2) = \sum_{i>r} \sigma^2.$$

6. **PCA.** Let A be a n-dimensional positive definite matrix. For $i = 1, \ldots, n$, denote with λ_i the i-th eigenvalue, with corresponding eigenvector u_i, and, without loss of generality, assume that the eigenvalues are ordered in decreasing order. Let $U\Lambda U^\top$ be the eigendecomposition of A. The Courant-Fischer-Weyl theorem implies that the eigenvalue/eigenvector pairs can be characterized in the following way. For any $x \in \mathbb{R}^d$, let $q(x) = x^\top Ax$. Then

$$\lambda_1 = q(u_1) = \max_{\|x\|=1} q(x).$$

For $k \geq 2$, let $\mathcal{U}_k$ be the k-dimensional subspace of $\mathbb{R}^n$ spanned by the first k leading eigenvectors $u_1, \ldots, u_k$. Then

$$\lambda_k = q(u_k) = \max_{\|x\|=1, x \perp \mathcal{U}_{k-1}} q(x),$$

where $x \perp \mathcal{U}_{k-1}$ signifies that $x \in \mathcal{U}_{k-1}^\perp$.

PCA is a technique for dimensionality reduction. If X is a n-dimensional random vector with covariance matrix Σ, then the first k principal components of X are the eigenvectors $u_1, \ldots, u_k$ and their scores are the eigenvalues $\lambda_1, \ldots, \lambda_k$, respectively.

(a) Show that $\mathrm{Var}(u_k^\top X) = \lambda_k$. That is, k-th PCA indicates a direction (a one-dimensional subspace) along which to orthogonal project X, and that projection has variance λ_k. Furthermore, the first PCAs are directions of maximal variance.
This immediately follows from the fact that the eigenvectors are orthonormal:

$$u_k^\top u_i = \begin{cases} 0 & i \neq k \\ 1 & i = 1. \end{cases}$$

Thus

$$\mathrm{Var}(u_k^\top X) = u_k^\top U\Lambda U^\top u_k = \sum_{i=1}^{n} (u_k^\top u_i)^2 \lambda_i = \lambda_k.$$

(b) The *total variance* of a (possibly rank deficient) covariance matrix is the sum of its diagonal. Show that this is the same as the sum of its eigenvalue.
Easy-peasy:

$$\mathrm{tr}(\Sigma) = \mathrm{tr}(U\Lambda U^\top) = \mathrm{tr}(U^\top U\Lambda) = \mathrm{tr}(\Lambda),$$

where the second inequality is the cyclicity of the trace operator and the third uses the fact that U is orthogonal.

(c) Show that the total variance of the orthogonal projection of X onto the first k principal components is maximal, i.e. larger than the total variance of the orthogonal projection of X onto any other k-dimensional linear subspace. So, one way to think of PCA is as the best - in the sense of maximizing the total variance - linear approximation of X by an affine subspace of dimension k.
We can show this by induction. When $k = 1$ this follows from part (a), because

λ_1 is the largest eigenvalue. Indeed, we pick a direction, say v, different than u_1 then the total variance is

$$\sum_{i=1}^{n}(v^\top u_i)^2\lambda_i \leq \max_i \lambda_i = \lambda_1$$

becuase $\sum_{i=1}^{n}(v^\top u_i)^2 = 1$. By part (b), the total variance of the first $k-1$ principal components is

$$\operatorname{tr}\left(\operatorname{Var}(U_{k-1}^\top X)\right) = \sum_{i=1}^{k-1}\lambda_i.$$

In choosing the next direction, since we want to maximize the total variance, we must pick a direction orthogonal to $u_1, \ldots, u_{k-1}$. Indeed, if we pick the next direction, say v, in the linear span of $u_1, \ldots, u_{k-1}$, then the total variance will not increase, i.e. it will still be $\sum_{i=1}^{k-1}\lambda_i$. If we choose the k-th direction to be u_k, then the total variance is $\sum_{i=1}^{k}\lambda_i$. If we choose another direction, say $v \neq u_k$, then the total variance is

$$\sum_{i=1}^{k-1}\lambda_i + \sum_{i\geq k}\lambda_i(v^\top u_i)^2 \leq \sum_{i=1}^{k}\lambda_i$$

since $\sum_{i\geq k}(v^\top u_i)^2 = 1$ and $\lambda_k = \max_{i\geq k}\lambda_i$.

7. **Distance between equidimensional linear subspaces.** Let $\mathcal{F}$ and $\mathcal{E}$ be two r-dimensional subspaces of $\mathbb{R}^d$ with orthogonal projection matrices $\mathrm{P}_\mathcal{F}$ and $\mathrm{P}_\mathcal{E}$, respectively. To measure the distance between them, a very commonly used metric is the sin-θ distance:

$$\frac{1}{\sqrt{2}}\|\mathrm{P}_\mathcal{F} - \mathrm{P}_\mathcal{E}\|_F.$$

(The fact that this is a distance is immediate and follow from the fact that the Frobenius norm is a norm. The division by $\sqrt{2}$ is made out of convenience and is immaterial. To learn more about this topic, see Chapter 5 of the book "Matrix Perturbation Theory" by Stewart and Sun). Show that the squared sin-θ distance is equal to

$$\|\mathrm{P}_\mathcal{F}(I_d - \mathrm{P}_\mathcal{E})\|_F^2 = \|\mathrm{P}_\mathcal{E}(I_d - \mathrm{P}_\mathcal{F})\|_F^2.$$

When $r = 1$ show that the above expression reduces to

$$1 - (e^\top f)^2,$$

where e and f are unit vectors spanning $\mathcal{E}$ and $\mathcal{F}$ respectively. It is, of course, not a coincidence that in this case the squared sin-θ distance is 1 minus the squared cosine of the angle between the vectors f and e.

By linearity of trace, and using repeatedly the facts that orthogonal projection matrices are idempotent and symmetric,

$$\begin{aligned}\|\mathrm{P}_{\mathcal{F}} - \mathrm{P}_{\mathcal{E}}\|_F^2 &= \mathrm{tr}\left((\mathrm{P}_{\mathcal{F}} - \mathrm{P}_{\mathcal{E}})(\mathrm{P}_{\mathcal{F}} - \mathrm{P}_{\mathcal{E}})\right) \\ &= \mathrm{tr}(\mathrm{P}_{\mathcal{F}} + \mathrm{P}_{\mathcal{E}} - \mathrm{P}_{\mathcal{F}}\mathrm{P}_{\mathcal{E}} - \mathrm{P}_{\mathcal{E}}\mathrm{P}_{\mathcal{F}}) \\ &= \mathrm{tr}\left(\mathrm{P}_{\mathcal{F}}(I - \mathrm{P}_{\mathcal{E}}) + \mathrm{P}_{\mathcal{E}}(I - \mathrm{P}_{\mathcal{F}})\right) \\ &= \|\mathrm{P}_{\mathcal{F}}(I - \mathrm{P}_{\mathcal{E}})\|_F^2 + \|\mathrm{P}_{\mathcal{E}}(I - \mathrm{P}_{\mathcal{F}}\|_F^2.\end{aligned}$$

Next,

$$\begin{aligned}\|\mathrm{P}_{\mathcal{F}}(I - \mathrm{P}_{\mathcal{E}})\|_F^2 &= \mathrm{tr}\left(\mathrm{P}_{\mathcal{F}}(I - \mathrm{P}_{\mathcal{E}})\right) \\ &= r + \mathrm{tr}\left(\mathrm{P}_{\mathcal{F}}\mathrm{P}_{\mathcal{E}}\right) \\ &= r + \mathrm{tr}\left(\mathrm{P}_{\mathcal{E}}\mathrm{P}_{\mathcal{F}}\right) \\ &= \mathrm{tr}\left(\mathrm{P}_{\mathcal{E}}(I - \mathrm{P}_{\mathcal{F}})\right) \\ &= \|\mathrm{P}_{\mathcal{E}}(I - \mathrm{P}_{\mathcal{F}}\|_F^2,\end{aligned}$$

where in the second identity, we have used the fact that the trace of an orthogonal projection matrix onto a subspace of dimension r is r, and in the third identity we have used cyclicity of the trace. Thus,

$$\frac{1}{2}\|\mathrm{P}_{\mathcal{F}} - \mathrm{P}_{\mathcal{E}}\|_F^2 = \|\mathrm{P}_{\mathcal{F}}(I_d - \mathrm{P}_{\mathcal{E}})\|_F^2 = \|\mathrm{P}_{\mathcal{E}}(I_d - \mathrm{P}_{\mathcal{F}})\|_F^2.$$

When $r = 1$ and e and f are unit vectors spanning $\mathcal{E}$ and $\mathcal{F}$, respectively, this quantity reduces to

$$1 + \mathrm{tr}(ee^\top ff^\top) = 1 + (e^\top f)\mathrm{tr}(ef^\top) = 1 - (e^\top f)^2.$$