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

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

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.*

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

$$n(\overline{X}_n^2 - \mu^2) \xrightarrow{d} \sigma^2(\chi_1^2(\gamma_n^2) - \gamma_n^2),$$

where $\chi_1^2(\gamma_n^2)$ denotes a chi-squared distribution with one degree of freedom and non-centrality parameter γ_n^2 and

$$\gamma_n = \sqrt{n}\mu/(2\sigma).$$

What is interesting about this problem is that $\sqrt{n}(\overline{X}_n^2 - \mu^2)$ has a non-trivial limiting distribution for all $\mu \neq 0$ but not when $\mu = 0$, an odd discontinuity of sort. On the other hand $n(\overline{X}_n^2 - \mu^2)$ does not suffer from this issue.
Hint: perform a second order Taylor series expansion and use the fact that if $X \sim N(\gamma, \sigma^2)$, then $X^2 \sim \sigma^2\chi_1^2(\gamma^2)$.

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.)

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

$$e^A = \sum_{i=1}^{\infty} \frac{A^i}{i!}.$$

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.

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 $n - r$ dimensional linear subspace and finds that subspace.

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.

6. **PCA.** Let A be a n-dimensional positive definite matrix. For $i = 1, \ldots, n$, denote with λ_i be 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 $\text{Var}(u_k^\top X) = \lambda_k$. That is, k-th PCA indicates a direction (a one-dimensional subspace) along which to project X, and that projection has variance λ_k. Furthermore, the first PCAs are directions of maximal variance.

(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.

(c) Show that the total variance of the projection of X onto the first k principal components is maximal, i.e. larger than the total variance of the 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.

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.