Optimization-Based Data Analysis

# Recitation 3

1. True or False: Let $A \in \mathbb{R}^{m \times n}$ be a matrix of data, with each column corresponding to a datapoint. If we want to compute the principal directions is it equivalent to compute the SVD of $A$ and the eigenvalue decomposition of $AA^T$.

   *Solution.* True algebraically, false numerically. Algebraically, the eigenvectors of $AA^T$ and the left singular vectors of $A$ are the same. Computationally, it is more stable numerically to compute the SVD of $A$.

2. True or False: If you are already working with features that have been normalized to have variance 1, there is no need to whiten your data.

   *Solution.* False. The covariance matrix $\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ is standardized but not whitened with singular values 1.5 and 0.5.

3. Let $\mathbf{x}[1], \ldots, \mathbf{x}[n]$ be i.i.d. random variables taking the values $-1, 0, +1$ with probabilities 1/3 each. Let $\vec{\mathbf{x}}$ denote the random vector in $\mathbb{R}^n$ having $\mathbf{x}[i]$ as its $i$th coordinate.

   (a) Compute $E[\|\vec{\mathbf{x}}\|_2^2]$.

   (b) Compute $E[\|\vec{\mathbf{x}}\|_\infty]$.

   (c) Compute the covariance matrix of $\vec{\mathbf{x}}$.

   *Solution.*

   (a) $E[\|\vec{\mathbf{x}}\|_2^2] = \sum_{k=1}^n E[\vec{\mathbf{x}}_i^2] = 2n/3$.

   (b) $E[\|\vec{\mathbf{x}}\|_\infty] = 1 - 1/3^n$.

   (c) Let $\Sigma = \text{Cov}(\vec{\mathbf{x}})$. Then $\Sigma_{ii} = 2/3$ and $\Sigma_{ij} = 0$ for $i \neq j$ by independence.

4. If $\mathbf{x} \sim \mathcal{N}(0, 1)$ then we say that $\mathbf{x}^2 \sim \chi_1^2$ (called a chi-squared distribution with 1 degree of freedom). Give the pdf, mean, and variance of the $\chi_1^2$ distribution.

   *Solution.* Let $\mathbf{y} = \mathbf{x}^2$. To compute the pdf we use the cdf $F_{\mathbf{y}}(y)$ of $\mathbf{y}$ for $y \geq 0$:

$$
\begin{aligned}
F_{\mathbf{y}}(y) &= \mathbb{P}(\mathbf{y} \leq y) \\
&= \mathbb{P}(\mathbf{x}^2 \leq y) \\
&= \mathbb{P}(-\sqrt{y} \leq \mathbf{x} \leq \sqrt{y}) \\
&= \mathbb{P}(-\sqrt{y} < \mathbf{x} \leq \sqrt{y}) \\
&= F_{\mathbf{x}}(\sqrt{y}) - F_{\mathbf{x}}(-\sqrt{y}).
\end{aligned}
$$

The pdf of $\mathbf{y}$ is given by

$$f_{\mathbf{y}}(y) = \frac{d}{dy} F_{\mathbf{y}}(y) = \frac{f_{\mathbf{x}}(\sqrt{y})}{2\sqrt{y}} - \frac{f_{\mathbf{x}}(-\sqrt{y})}{-2\sqrt{y}} = \frac{f_{\mathbf{x}}(\sqrt{y})}{\sqrt{y}} = \frac{e^{-y/2}}{\sqrt{2\pi y}},$$

for $y > 0$ and 0 otherwise. The mean is simply the variance of a standard normal random variable, which is 1. Also note that

$$\begin{aligned}
E[\mathbf{y}^2] &= E[\mathbf{x}^4] \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^4 e^{-x^2/2}\, dx \\
&= \frac{1}{\sqrt{2\pi}} \left[ -x^3 e^{-x^2/2} \right]_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} 3x^2 e^{-x^2/2}\, dx \\
&= 3.
\end{aligned}$$

Thus $\mathrm{Var}[\mathbf{y}] = E[\mathbf{y}^2] - E[\mathbf{y}]^2 = 2$.

5. Let $A = \begin{bmatrix} 4 & -1 \\ 4 & 1 \end{bmatrix}$ and $\vec{b} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$. Suppose $\vec{\mathbf{x}} \sim \mathcal{N}(0, I)$ takes values in $\mathbb{R}^2$, and let $\vec{\mathbf{y}} = A\vec{\mathbf{x}} + \vec{b}$.

   (a) What is the distribution of $\vec{\mathbf{y}}$?

   (b) What are the marginal distributions of the components of $\vec{\mathbf{y}}$?

   (c) Are the components of $\vec{\mathbf{y}}$ independent?

   (d) What do the contour lines of the joint pdf $\vec{\mathbf{y}}$ look like?

*Solution.*

(a) $\vec{\mathbf{y}} \sim \mathcal{N}(\vec{b}, AA^T)$

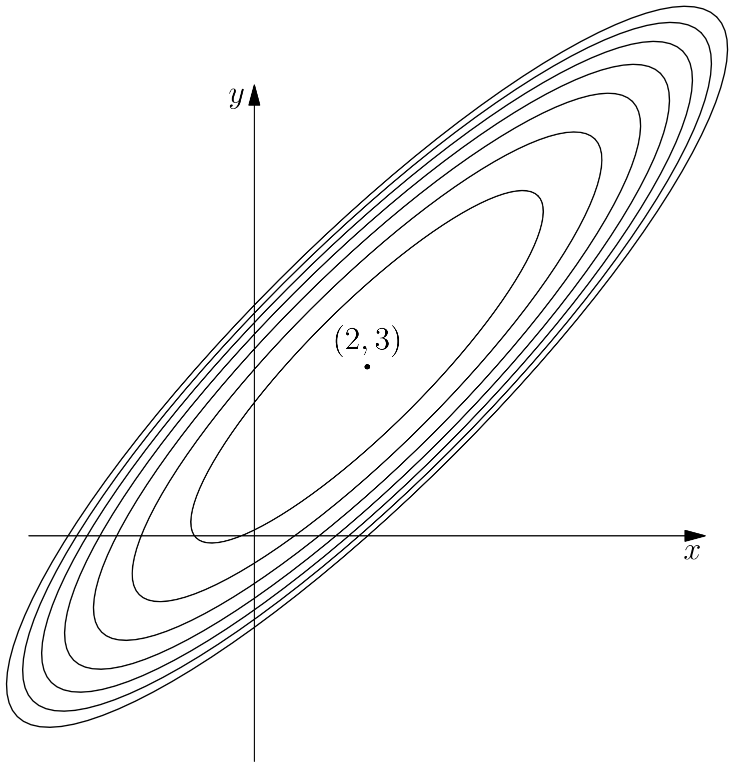
(b) Since

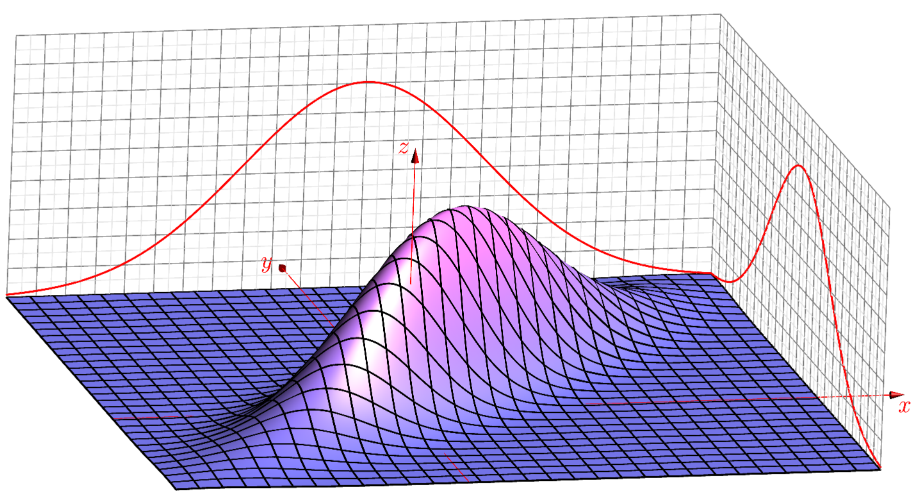
$$AA^T = \begin{bmatrix} 17 & 15 \\ 15 & 17 \end{bmatrix},$$

we have $\vec{\mathbf{y}}[1] \sim \mathcal{N}(2, 17)$ and $\vec{\mathbf{y}}[2] \sim \mathcal{N}(3, 17)$.

(c) No, as they are positively correlated.

(d) Below we give a contour plot of the joint pdf along with a 3d plot.




This can be understood by computing the SVD of $A$:

$$A = \begin{bmatrix} 4 & -1 \\ 4 & 1 \end{bmatrix} = \begin{bmatrix} 4 & -1 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} 4\sqrt{2} & 0 \\ 0 & \sqrt{2} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}^T.$$

In general, if $A = USV^T$ then $V^T$ applied to an i.i.d. Gaussian vector fixes the contours, $S$ stretches the contours, and then $U$ rotates the stretched contours. Thus, the resulting contours are always ellipsoids.

6. Let $\mathbf{x} \sim \mathcal{N}(0,1)$. Compute an upper bound on the probability that $\mathbb{P}(\mathbf{x} \geq k)$ in terms of $k > 0$. [Hint: Integration by parts.]

   *Solution.* Note that

$$\begin{aligned}
\mathbb{P}(\mathbf{x} \geq k) &= \frac{1}{\sqrt{2\pi}} \int_k^\infty e^{-x^2/2}\,dx \\
&= \frac{1}{\sqrt{2\pi}} \int_k^\infty \frac{x}{x} e^{-x^2/2}\,dx \\
&= -\frac{1}{\sqrt{2\pi}} \left[\frac{e^{-x^2/2}}{x}\right]_k^\infty - \frac{1}{\sqrt{2\pi}} \int_k^\infty \frac{e^{-x^2/2}}{x^2}\,dx \\
&\leq \frac{e^{-k^2/2}}{k\sqrt{2\pi}}.
\end{aligned}$$

   This bound is good if $k$ isn't close to zero.

7. Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from a random sample from a Bernoulli($p$) distribution. How large must $n$ be to guarantee that

$$\mathbb{P}(|\overline{\mathbf{x}}_n - p| < 0.01) \geq 0.98?$$

   Here $\overline{\mathbf{x}}_n$ is defined to be the sample mean:

$$\overline{\mathbf{x}}_n = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i.$$

   *Solution.* Note that $\mathbb{P}(|\overline{\mathbf{x}}_n - p| < 0.01) \geq 0.98$ if and only if $\mathbb{P}(|\overline{\mathbf{x}}_n - p| \geq 0.01) \leq 0.02$. If we apply Chebyshev's inequality, we obtain

$$\begin{aligned}
\mathbb{P}(|\overline{\mathbf{x}}_n - p| \geq 0.01) &\leq \frac{\mathrm{Var}(\overline{\mathbf{x}}_n)}{0.01^2} \\
&= \frac{10000(p(1-p))}{n}.
\end{aligned}$$

   We can guarantee $\frac{10000(p(1-p))}{n} \leq 0.02$ if

$$n \geq 500000p(1-p) \geq 125000,$$

   as $p(1-p)$ is maximized at $p = 1/2$. If instead we approximate

$$\sqrt{\frac{n}{p(1-p)}}(\overline{\mathbf{x}}_n - p) \approx \mathcal{N}(0,1),$$

then

$$\mathbb{P}(|\overline{\mathbf{x}}_n - p| < 0.01) = \mathbb{P}\left(-0.01\sqrt{\frac{n}{p(1-p)}} < \sqrt{\frac{n}{p(1-p)}}(\overline{\mathbf{x}}_n - p) < 0.01\sqrt{\frac{n}{p(1-p)}}\right)$$
$$\approx 1 - 2\Phi\left(-0.01\sqrt{\frac{n}{p(1-p)}}\right),$$

where $\Phi$ is the cdf of the standard normal distribution. This is larger than 0.98 when

$$\begin{aligned}
0.02 \geq 2\Phi\left(-0.01\sqrt{\frac{n}{p(1-p)}}\right) &\iff \Phi^{-1}(0.01) \geq -0.01\sqrt{\frac{n}{p(1-p)}} \\
&\iff -100\Phi^{-1}(0.01) \leq \sqrt{\frac{n}{p(1-p)}} \\
&\iff 232.6348 \leq \sqrt{\frac{n}{p(1-p)}} \\
&\iff 54119p(1-p) \leq n,
\end{aligned}$$

giving a bound of $n \geq 13530$. While this is a much better bound, it only holds approximately. To get a precise bound we can appeal to a stronger version of the CLT like the Berry-Esseen theorem. As an alternative, we can also extend our Chebyshev proof using something called Chernoff bounds:

$$\mathbb{P}(\mathbf{y} \geq a) = \mathbb{P}(e^{t\mathbf{y}} \geq e^{ta}) \leq e^{-ta}E[e^{t\mathbf{y}}],$$

for all $t > 0$. Thus we can bound the tail probability of a random variable $\mathbf{y}$ by bounding its moment generating function $\varphi(t) = E[e^{t\mathbf{y}}]$. If you follow through with this technique on our Bernoulli samples we obtain Hoeffding's inequality:

$$\mathbb{P}(|\overline{\mathbf{x}}_n - p| \geq 0.01) \leq 2e^{-2n(0.01)^2}.$$

Solving we see

$$2e^{-2n(0.01)^2} \leq 0.02 \iff -2n(0.01)^2 \leq \log(0.01) \iff n \geq -5000\log(0.01) = 23025.85.$$

While this is worse than our CLT bound, it holds for all $n$.

8. In this question we prove a version of Hoeffding's inequality. It can easily be extended to general independent random variables taking values in a bounded interval $[a, b]$ by an affine transformation.

   Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be i.i.d. random variables taking the values $-1, +1$ with probabilities $1/2$ each. Let $\mathbf{s}_n = \mathbf{x}_1 + \cdots + \mathbf{x}_n$.

   (a) Give the upper bound for $\mathbb{P}(|\mathbf{s}_n| \geq a\sqrt{n})$ given by Chebyshev's inequality.
   (b) Use the central limit theorem to approximate $\mathbb{P}(|\mathbf{s}_n| \geq a\sqrt{n})$. This is valid for large $n$.

(c) Let $\varphi_{\mathbf{y}}(t) = E[e^{t\mathbf{y}}]$ denote the moment generating function for a random variable $\mathbf{y}$ (where the expectation is finite). Prove that

$$\mathbb{P}(\mathbf{s}_n \geq a\sqrt{n}) \leq e^{-ta\sqrt{n}}\varphi_{\mathbf{s}_n}(t),$$

for all $t > 0$ using Markov's inequality. [This is called a Chernoff bound.]

(d) Show that $\varphi_{\mathbf{s}_n}(t) = \varphi_{\mathbf{x}_1}(t)^n$.

The remaining parts are more advanced.

(e) Prove that $\varphi_{\mathbf{x}_1}(t) \leq \cosh(t)$ by using the fact that $f(x) = e^{tx}$ is convex and $x \in [-1, 1]$ giving

$$e^{tx} \leq \frac{1-x}{2}e^{-t} + \frac{1+x}{2}e^{t}.$$

(f) Prove that $\cosh(t) \leq e^{t^2/2}$ by comparing Taylor series.

(g) Combining earlier results, show that

$$\mathbb{P}(\mathbf{s}_n \geq a\sqrt{n}) \leq e^{-ta\sqrt{n}}e^{nt^2/2},$$

for all $t > 0$.

(h) Optimizing over $t$ in the previous part, conclude Hoeffding's lemma:

$$\mathbb{P}(\mathbf{s}_n \geq a\sqrt{n}) \leq e^{-a^2/2},$$

and

$$\mathbb{P}(|\mathbf{s}_n| \geq a\sqrt{n}) \leq 2e^{-a^2/2}.$$

*Solution.*

(a) $\mathbb{P}(|\mathbf{s}_n| \geq a\sqrt{n}) \leq \frac{\mathrm{Var}(\mathbf{s}_n)}{a^2 n} = \frac{1}{a^2}$

(b) We approximate $|\mathbf{s}_n|/\sqrt{n}$ by $\mathcal{N}(0,1)$ to get

$$\mathbb{P}(|\mathbf{s}_n| \geq a\sqrt{n}) = 2\mathbb{P}(\mathcal{N}(0,1) \geq a) = 2\int_a^\infty e^{-x^2/2}\,dx \leq 2\int_a^\infty \frac{x}{a}e^{-x^2/2}\,dx = \frac{2e^{-a^2/2}}{a}.$$

(c) Note that

$$\mathbb{P}(\mathbf{s}_n \geq a\sqrt{n}) = \mathbb{P}(e^{t\mathbf{s}_n} \geq e^{ta\sqrt{n}}) \leq e^{-ta\sqrt{n}}E[e^{t\mathbf{s}_n}],$$

by Markov's inequality. Note that $E[e^{t\mathbf{s}_n}]$ exists for all $t > 0$ since $S_n$ is bounded.

(d) Note that

$$E[e^{t\mathbf{s}_n}] = E\left[\prod_{k=1}^n e^{t\mathbf{x}_k}\right] = \prod_{k=1}^n E\left[e^{t\mathbf{x}_k}\right] = E[e^{t\mathbf{x}_1}]^n,$$

by independence.

(e) To see that $e^{tx}$ is convex, note it has a positive second derivative everywhere. The hinted inequality comes directly from the definition of convexity since

$$\frac{1-x}{2}+\frac{1+x}{2}=1 \quad \text{and} \quad -(1-x)/2+(1+x)/2=x.$$

Taking expectations on both sides we have

$$E[e^{t\mathbf{x}_1}] \leq E\left[\frac{1-\mathbf{x}_1}{2}e^{-t}+\frac{1+\mathbf{x}_1}{2}e^{t}\right]=\frac{e^{-t}+e^{t}}{2}=\cosh(t),$$

since $E[\mathbf{x}_1]=0$.

(f) Note that

$$\cosh(t)=\sum_{k=0}^{\infty}\frac{t^{2k}}{(2k)!} \leq \sum_{k=0}^{\infty}\frac{t^{2k}}{k!2^k}=e^{t^2/2},$$

since $k!2^k$ is the product of the even numbers up to $2k$ and $(2k)!$ is the product of all of them.

(g) Plugging in we have

$$\mathbb{P}(\mathbf{s}_n \geq a\sqrt{n}) \leq e^{-ta\sqrt{n}}\varphi_{\mathbf{x}_1}(t)^n \leq e^{-ta\sqrt{n}}e^{nt^2/2}.$$

(h) Optimizing the quadratic in $t$ in the exponent we obtain $t=a/\sqrt{n}$. Plugging in gives the result. For the absolute value, we apply symmetry.