

CS761 Spring 2015 Homework 2

Assigned Mar. 13, due Mar. 27 before class

Instructions:

- Homeworks are to be done individually.
- Typeset your homework in latex using this file as template (e.g. use pdf_lat_ex). Show your derivations.
- Hand in the compiled pdf (not the latex file) online. Instructions will be provided. We do not accept hand-written homeworks.
- Homework will no longer be accepted once the lecture starts.
- Fill in your name and email below.

Name: Max Horowitz-Gelb

Email: horowitzgelb@wisc.edu

1. Let X_0, X_1, \dots, X_{M-1} denote a random sample of N -dimensional random vectors X_n , each of which has mean value m and covariance matrix R . Show that the sample mean

$$\hat{m}_t = \frac{1}{t+1} \sum_{n=0}^t X_n$$

and the sample covariance

$$S_t(\hat{m}_t) = \frac{1}{t+1} \sum_{n=0}^t (X_n - \hat{m}_t)(X_n - \hat{m}_t)^\top$$

may be written recursively as

$$\hat{m}_t = \frac{t}{t+1} \hat{m}_{t-1} + \frac{1}{t+1} X_t, \quad \hat{m}_0 = X_0,$$

and

$$S_t(\hat{m}_t) = Q_t - \hat{m}_t \hat{m}_t^\top,$$

where

$$Q_t = \frac{t}{t+1} Q_{t-1} + \frac{1}{t+1} X_t X_t^\top.$$

i.

Assume that for some $t > 0$

$$\hat{m}_{t-1} = \frac{t-1}{t} \hat{m}_{t-2} + \frac{1}{t+1} X_{t-1}$$

then

$$\begin{aligned} & \frac{t}{t+1} \hat{m}_{t-1} + \frac{1}{t+1} X_t \\ &= \frac{t}{t+1} \frac{1}{t} \sum_{n=0}^{t-1} X_n + \frac{1}{t+1} X_t \\ &= \frac{1}{t+1} \sum_{n=0}^t X_n \\ &= \hat{m}_t \end{aligned}$$

Then since clearly

$$\hat{m}_1 = \frac{1}{2} \hat{m}_0 + \frac{1}{2} X_1$$

Then for all $t > 0$

$$\hat{m}_t = \frac{t}{t+1} \hat{m}_{t-1} + \frac{1}{t+1} X_t$$

ii.

First I will show that $Q_t = \frac{1}{t+1} \sum_{n=1}^t X_n X_n^\top$

Base case:

$$Q_1 = 1/2 X_1 X_1^\top = 1/2 \sum_{n=1}^0 X_n X_n^\top + 1/2 X_1 X_1^\top$$

Induction hypothesis:

Assume $Q_{t-1} = \frac{1}{t} \sum_{n=1}^{t-1} X_n X_n^\top$

Then,

$$\begin{aligned} Q_t &= \frac{t}{t+1} * \frac{1}{t} \sum_{n=1}^{t-1} X_n X_n^\top + 1/(t+1) X_t X_t^\top \\ &= 1/(t+1) * \sum_{n=1}^t X_n X_n^\top \end{aligned}$$

Then since $Q_1 = 1/2 X_1 X_1^\top = 1/2 \sum_{n=1}^0 X_n X_n^\top + 1/2 X_1 X_1^\top$

Then by induction for all $t \geq 1$, $Q_t = 1/(t+1) * \sum_{n=1}^t X_n X_n^\top$

Then we can multiply out the dot product for $S_t(\hat{m}_t)$ and get

$$\begin{aligned} S_t(\hat{m}_t) &= \frac{1}{t+1} \sum_{n=1}^t X_n X_n^\top - 2X_n \hat{m}_t^\top + \hat{m}_t \hat{m}_t^\top \\ &= \frac{1}{t+1} \sum_{n=1}^t (X_n X_n^\top) - 2\hat{m}_t \hat{m}_t^\top + \hat{m}_t \hat{m}_t^\top \\ &= \frac{1}{t+1} \sum_{n=1}^t (X_n X_n^\top) - \hat{m}_t \hat{m}_t^\top \end{aligned}$$

Using our inductive proof from before this is equal to

$$Q_t - \hat{m}_t \hat{m}_t^\top$$

- Suppose we roll a fair 6-sided die 100 times. Let X be the sum of the outcomes. Bound $P(|X - 350| \geq 100)$ using Chebyshev and Hoeffding, respectively.

X is the sum of n iid outcomes, $Y_1 \dots Y_n$.

$$E[Y_i] = 7/2$$

$$\text{Var}[Y_i] = 35/12$$

Therefore

$$E[X] = \sum_{i=1}^{100} E[Y_i] = 350$$

$$\text{Var}[X] = \sum_{i=1}^{100} \text{Var}[Y_i] = 291 + 2/3$$

Therefore by the Chebyshev inequality

$$\Pr(|X - 350| \geq 100) = \Pr(|X - 350| \geq \frac{100}{\sqrt{291 + 2/3}} \sqrt{291 + 2/3}) \leq \frac{291 + 2/3}{10000}$$

And by the Hoeffding inequality

$$\Pr(|X - 350| \geq 100) \leq 2 \exp\left(-\frac{20000}{\sum_{n=1}^{100} (6-1)^2}\right) = 2 \exp(-8)$$

3. Let \mathcal{X} be the vector space of *finitely* nonzero sequences $X = (x_1, x_2, \dots, x_n, 0, 0, \dots)$. Define the norm on \mathcal{X} as $\|X\| = \max |x_i|$. Let X_n be a point in \mathcal{X} (a sequence) defined by

$$X_n = \left(1, \frac{1}{2}, \frac{1}{3}, \dots, \frac{1}{n}, 0, 0, \dots\right).$$

- Show that the sequence X_n is a Cauchy sequence.

The sequence X_n is a Cauchy sequence. Let $\epsilon > 0$ be given. Then let $N = \lceil 1/\epsilon \rceil$. Then for any $l, s > N$ such that $l \leq s$

$$\|x_s - x_l\| = \|(0, 0, \dots, 1/(l+1), \dots, 1/s, 0, \dots, 0)\| = 1/(l+1) < 1/N \leq \epsilon$$

- Show that \mathcal{X} is not complete.

\mathcal{X} is not complete since $\|X_n - X_{n-1}\|$ converges to 0 which implies that as $n \rightarrow \infty$, $X_{n-1} \rightarrow X_n$ and this would imply that the number of non-zero elements of X_{n-1} would have to go to ∞ , which is not finite, and so X_{n-1} is not in \mathcal{X} .

4. Determine the range and nullspace of the following linear operators (matrices):

$$A = \begin{bmatrix} 1 & 0 \\ 5 & 4 \\ 2 & 4 \end{bmatrix} \quad B = \begin{bmatrix} 1 & 0 & 1 \\ 5 & 4 & 9 \\ 2 & 4 & 6 \end{bmatrix}$$

A has a null space of $\vec{0}$ and range equal to $\text{span}([1, 5, 2]^\top, [0, 4, 4]^\top)$

B has a null space equal to $\text{span}([-1, -1, 1]^\top)$ and a range equal to $\text{span}([1, 5, 2]^\top, [0, 4, 4]^\top)$.

5. Let

$$A = \begin{bmatrix} 1 & 4 & 5 & 6 \\ 6 & 7 & 2 & 1 \end{bmatrix} \quad b = \begin{bmatrix} 48 \\ 30 \end{bmatrix}.$$

One solution to $Ax = b$ is $x = [1, 2, 3, 4]^\top$. Compute the least-squares solution using the SVD (explain how), and compare. Why was the solution chosen?

Using SVD A is equivalent to

$$U\Sigma V$$

where U and V are orthogonal matrices and Σ is a diagonal matrix with diagonal values σ_1, σ_2 . Then solving $Ax = b$ is equivalent to minimizing $\|Ax - b\|^2$ which is equivalent to,

$$\|U\Sigma Vx - b\|^2$$

Since U^\top is orthogonal it does not change the norm so the above is equivalent to,

$$\|U^\top U\Sigma Vx - U^\top b\|^2 = \|\Sigma Vx - U^\top b\|^2$$

Then let $z = Vx$ and the above is equivalent to

$$(\sigma_1 z_1 - U_1^\top b)^2 + (\sigma_2 z_2 - U_2^\top b)^2 + (U_3^\top b)^2 + (U_4^\top b)^2$$

Since this is a minimization we can remove the summands that don't involve z and get

$$(\sigma_1 z_1 - U_1^\top b)^2 + (\sigma_2 z_2 - U_2^\top b)^2$$

We can then clearly minimize this by setting the free variables $z_3 = z_4 = 0$ and $z_1 = \frac{U_1^\top b}{\sigma_1}$, $z_2 = \frac{U_2^\top b}{\sigma_2}$. Solving for z we get

$$z = [-4.68423189, -2.75801453, 0, 0]$$

and

$$x = V^\top z = [0.54424779, 2.40265487, 3.09292035, 3.7300885]$$

Since $\|z\| = \|Vx\| = \|x\|$, Then since z_1, z_2 are constrained and z_3, z_4 are free, then by setting $z_3 = z_4 = 0$ we minimize $\|z\|$ and thereby minimize $\|x\|$. Therefore SVD gives the minimum norm solution to $Ax = b$.

6. Consider the following process. A probability vector $p = (p_1, \dots, p_d)$ is drawn from a Dirichlet distribution with parameter vector α . Then, a vector of category counts $x = (x_1, \dots, x_d)$ is drawn from a multinomial distribution with probability vector p and number of trials N . Give an analytic form of $P(x | \alpha)$.

Since our p vector is random then

$$\begin{aligned} P(x|\alpha) &= \int_p \text{multinomial}(x|p) * \text{dirichlet}(p|\alpha) dp \\ &= \frac{\Gamma(\sum_i \alpha_i) d!}{\prod_i \Gamma(\alpha_i) * \prod_i x_i!} \int_p \prod_i p_i^{\alpha_i-1} \prod_i p_i^{x_i} \\ &= \frac{\Gamma(\sum_i \alpha_i) d!}{\prod_i \Gamma(\alpha_i) * \prod_i x_i!} \int_p \prod_i p_i^{\alpha_i+x_i-1} \end{aligned}$$

We then noticed that the integral is that of an unnormalized dirichlet pmf with $\alpha'_i = \alpha_i + x_i$ Therefore the above becomes,

$$\frac{\prod_i (\Gamma(\alpha_i + x_i) \Gamma(\sum_i \alpha_i) d!)}{\Gamma(\sum_i \alpha_i + x_i) \prod_i \Gamma(\alpha_i) * \prod_i x_i!}$$

7. Let X_1, X_2, \dots, X_m be a random sample, where $X_i \sim U(0, \theta)$ the uniform distribution.

- Show that $\hat{\theta}_{ML} = \max X_i$.
- Show that the density of $\hat{\theta}_{ML}$ is $f_{\theta}(x) = \frac{m}{\theta^m} x^{m-1}$.
- Find the expected value of $\hat{\theta}_{ML}$.
- Find the variance of $\hat{\theta}_{ML}$.

Under the uniform distribution the likelihood for $\theta > 0$ is equal to

$$L(\theta|x_1, \dots, x_m) = \prod_{i=1}^m \mathbb{1}(x_i \leq \theta) * 1/\theta$$

Therefore clearly since θ is positive the maximum likelihood estimator is $\max X_i$

Since $\hat{\theta}_{ML}$ is the max order statistic then it has density,

$$f_{\theta}(x) = m * F(x)^{m-1} * f(x) = m * (x/\theta)^{m-1} * 1/\theta = \frac{m}{\theta^m} x^{m-1}$$

$$E[\hat{\theta}_{ML}] = \int_0^{\theta} x * \frac{m}{\theta^m} x^{m-1} = \frac{m}{m+1} \theta$$

$$E[\hat{\theta}_{ML}^2] = \int_0^\theta = x^2 * \frac{m}{\theta^m} x^{m-1} = \frac{m}{m+2} \theta^2$$

$$Var[\hat{\theta}_{ML}] = E[\hat{\theta}_{ML}^2] - E[\hat{\theta}_{ML}]^2 = (\frac{m}{m+2} - \frac{m^2}{(m+1)^2}) \theta^2$$

8. Let X_1, \dots, X_n be a sample from $N(\mu, \sigma^2)$.

- Show that the MLE of σ^2 is

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Since the maximum likelihood of σ^2 is dependent on μ we must consider a likelihood function that is a function of $\hat{\sigma}^2$ and $\hat{\mu}$. So for $\hat{\mu}$ we will use the MLE of μ , which is \bar{X} . Then the likelihood function is

$$L(X|\sigma^2 = \hat{\sigma}^2, \mu = \hat{\mu}) = (2\pi\hat{\sigma}^2)^{-n/2} \exp(-1/2\hat{\sigma}^2 * \sum_i (X_i - \hat{\mu})^2)$$

Minimizing this is equivalent to minimizing the log,

$$-n/2 * \log(2\pi) - n/2 * \log(\hat{\sigma}^2) - 1/2\hat{\sigma}^2 * \sum_i (X_i - \hat{\mu})^2$$

The derivative of this with respect to $\hat{\sigma}^2$ is ,

$$\frac{-n}{2\hat{\sigma}^2} + \frac{\sum_i (X_i - \hat{\mu})^2}{2\hat{\sigma}^4}$$

which when $\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 = n^{-1} \sum_i (X_i - \bar{X})^2$ is equal to

$$\frac{-n^2}{2 \sum_i (X_i - \bar{X})^2} + \frac{n^2 * \sum_i (X_i - \bar{X})^2}{2 (\sum_i (X_i - \bar{X})^2)^2} = 0$$

- Show that $\hat{\sigma}^2$ has a smaller mean squared error than

$$(n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

We know that the above is the sample variance, which we will call $\hat{\sigma}_{sv}^2$, has the properties,

$$E[\hat{\sigma}_{sv}^2] = \sigma^2$$

$$Var(\hat{\sigma}_{sv}^2) = 2\sigma^4/(n-1)$$

Since

$$\hat{\sigma}^2 = (n-1)/n * \hat{\sigma}_{sv}^2$$

then,

$$E[\hat{\sigma}^2] = (n-1)/n * \sigma^2$$

$$Var(\hat{\sigma}^2) = (n-1)^2/n^2 * 2\sigma^4/(n-1) = \frac{2(n-1)\sigma^4}{n^2}$$

The MSE of an estimator is its variance plus bias squared so,

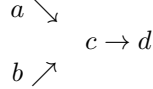
$$MSE(\hat{\sigma}_{sv}^2) = 2\sigma^4/(n-1)$$

and,

$$MSE(\hat{\sigma}^2) = \frac{2(n-1)\sigma^4}{n^2} + (\sigma^2/n)^2 = \frac{\sigma^4(2n-1)}{n^2} < 2\sigma^4/n$$

for positive n , $2\sigma^4/n < 2\sigma^4/(n-1)$ and then $MSE(\hat{\sigma}^2) < MSE(\hat{\sigma}_{sv}^2)$.

9. Consider the directed graphical model in which none of the variables is observed.



Show that $a \perp b | \emptyset$ by using a probability argument. Suppose we now observe the variable d . Show that in general $a \not\perp b | d$ (you can use a counterexample).

For any directed graphical model, the local markhov assumptions state that for all X , $(X \perp Y | Parents(X))$ for all Y that are nondescendants of X . Therefore since a and b parents are equal to \emptyset and are not descendants of each other, then $a \perp b | \emptyset$

- ii. For example assume a, b, c, d are boolean random variables. Then we will define probabilities as such.

$$p(a = T) = 0.5$$

$$p(b = T) = 0.5$$

$$p(c = T | a, b) = \begin{cases} 1 & a = T, b = T \\ 0 & \text{else} \end{cases}$$

10. Consider two discrete random variables $x, y \in \{A, B, C\}$. Construct a joint distribution $p(x, y)$ with the following properties:

- \hat{x} is the maximizer of the marginal $p(x)$
- \hat{y} is the maximizer of the marginal $p(y)$
- $p(\hat{x}, \hat{y}) = 0$.

Let our joint probability be

$$P(x, y) = \begin{cases} 1/3 & x = A, y \neq B \\ 1/6 & y = C, x \neq A \\ 0 & \text{else} \end{cases}$$

For this distribution A is a maximizer for x with $p(x = A) = 2/3$ and B is a maximizer of y with $p(y = B) = 1/3$. But $p(X = A, Y = B) = 0$

11. Logistic regression for $y \in \{-1, 1\}$ is defined by

$$p(y \mid x; w, b) = \frac{1}{1 + e^{-y(x^\top w + b)}}.$$

Show that logistic regression is in the exponential family, that is, the probability distribution can be written in the form

$$p(y \mid x; \tilde{w}) = \frac{1}{Z(x, \tilde{w})} e^{\phi(y, x)^\top \tilde{w}}.$$

Note the mapping ϕ depends only on y, x , but not on w or b .

Note that the above probability is equal to

$$\begin{aligned} & \frac{e^{1/2 y (x^\top w + b)}}{e^{1/2 y (x^\top w + b)}} * \frac{1}{1 + e^{-y(x^\top w + b)}} \\ &= \frac{e^{1/2 y (x^\top w + b)}}{e^{1/2 y (x^\top w + b)} + e^{-1/2 y (x^\top w + b)}} \end{aligned}$$

Since $y \in \{1, -1\}$ this is equivalent to

$$\frac{e^{1/2 y (x^\top w + b)}}{e^{1/2 (x^\top w + b)} + e^{-1/2 (x^\top w + b)}}$$

Then let $\tilde{w} = [w, b]$, $Z(x, \tilde{w}) = e^{1/2 (x^\top w + b)} + e^{-1/2 (x^\top w + b)}$ and $\phi(y, x) = [yx/2, y/2]$,

Then the above is equivalent to

$$\frac{1}{Z(x, \tilde{w})} e^{\phi(y, x)^\top \tilde{w}}$$

and therefore the probability belongs to the exponential family.