18.650. Fundamentals of Statistics
Fall 2023. Recitation sheet 1.2

1 Multivariate random variables and limit theorems

Problem 1 Let

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}\right)$$

Compute the following quantities

1. $\mathbb{V}[X], \mathbb{V}[Y]$ **Solution:** $= 3, 2$.
2. $\mathbb{E}[(X-Y)^2]$ **Solution:** $= \mathbb{E}X^2 + \mathbb{E}Y^2 - 2\mathbb{E}XY = \mathbb{V}(X) + (\mathbb{E}X)^2 + \mathbb{V}(Y) + (\mathbb{E}Y)^2 - 2\,\mathrm{Cov}(X,Y) - 2(\mathbb{E}X)(\mathbb{E}Y) = 3 + 1 + 2 + 4 - 2 - 4 = 4$.
3. $\mathbb{V}[X+2Y]$ **Solution:** $= \mathbb{V}(X) + 4\mathbb{V}(Y) + 4\,\mathrm{Cov}(X,Y) = 3 + 8 + 4 = 15$.
4. $\mathbb{E}[X^2Y]$. (Hint: find a number a such that $Y - aX$ is uncorrelated with X) **Solution:** $\mathrm{Cov}(Y - aX, X) = \mathrm{Cov}(Y,X) - a\mathbb{V}[X] = 1 - 3a$. This is 0 when $a = 1/3$. Therefore, we can write $Y = \frac{1}{3}X + X'$ for a random variable X' which is uncorrelated with X and therefore independent of X (this is a property of Gaussians). Note that $\mathbb{E}[X'] = \mathbb{E}[Y] - \mathbb{E}[X]/3 = 2 - 1/3 = 5/3$. Now, we get

$$\begin{aligned}
\mathbb{E}[X^2Y] &= \mathbb{E}[X^2(\frac{1}{3}X + X')] = \frac{1}{3}\mathbb{E}[X^3] + \mathbb{E}[X^2]\mathbb{E}[X'] \\
&= \frac{1}{3}\mathbb{E}[(1+\sqrt{3}Z)^3] + (1^2+3)\frac{5}{3} \\
&= \frac{1}{3}\mathbb{E}[1 + 3\sqrt{3}Z + 9Z^2 + 3\sqrt{3}Z^3] + \frac{20}{3} \\
&= \frac{10}{3} + \frac{20}{3} = 10.
\end{aligned} \tag{1}$$

Problem 2 As usual, $\mathbb{E}X := (\mathbb{E}X_1, \ldots, \mathbb{E}X_d)^\top$ for a random vector $X \in \mathbb{R}^d$.

1. Let $X \in \mathbb{R}^d$ be a random vector and let $L : \mathbb{R}^d \to \mathbb{R}^k$ be a linear map. Prove that $\mathbb{E}[L(X)] = L(\mathbb{E}[X])$. **Solution:** Since L is linear it can be written as $L(x) = Ax$ for some $A \in \mathbb{R}^{k\times d}$. In particular, the $i'th$ coordinate of the vector $L(X)$ is given by

$$[L(X)]_i = \sum_{j=1}^{d} A_{ij}X_j.$$

By linearity of expectation $\mathbb{E}\big[[L(X)]_i\big] = \sum_{j=1}^{d} A_{ij}\mathbb{E}X_j = [A\mathbb{E}[X]]_i$. In other words, $\mathbb{E}[L(X)] = A\mathbb{E}[X] = L(\mathbb{E}[X])$.

2. Let Y be a random $d \times d$ matrix and A be a deterministic $d \times d$ matrix. Prove that $\mathbb{E}Tr(AY) = Tr(A\mathbb{E}[Y])$. (Recall that the trace Tr of a square matrix is the sum of its diagonal entries.) **Solution:** We apply the previous. Note that the map $y \mapsto Tr(Ay)$ is linear. Thinking of Y as a vector (by say flattening it), we can conclude by the previous part.

3. Let $X \in \mathbb{R}^d$ be a random vector such that $\mathbb{E}[X] = (0, \ldots, 0)^\top$ and $\mathbb{V}[X] = I_d$. Let A be a deterministic $d \times d$ matrix. Compute $\mathbb{E}[X^\top AX]$. **Solution:** We'll do this two ways: with and without indices. Using indices, we have

$$\mathbb{E}[X^T AX] = \mathbb{E}[\sum_{i,j} A_i jX_iX_j] = \sum_{i,j} A_{ij}\mathbb{E}[X_iX_j].$$

Now, note that since $\mathbb{E}[X]$ is the zero vector, $\mathbb{E}[X_iX_j]$ is just the covariance between X_i and X_j, which is the ijth entry of the covariance matrix $\mathbb{V}[X]$. But the covariance matrix is the identity, so $\mathbb{E}[X_iX_j] = 1$ if $i = j$ and 0 otherwise. Therefore, we get $\mathbb{E}[X^T AX] = \sum_i A_{ii} = Tr(A)$.

Without indices: $\mathbb{E}[X^T AX] = \mathbb{E}[Tr(X^T AX)] = \mathbb{E}[Tr(AXX^T)] = Tr(A\mathbb{E}[XX^T]) = Tr(A)$. We used that the trace of a number is just the number itself, and that the trace is invariant under circular shifts: $Tr(ABC) = Tr(BCA)$. Finally, we applied part 2 and the fact that $\mathbb{E}[XX^T] = \mathbb{V}[X] = I_d$.

Problem 3 (AoS Exercise 5.15) Let

$$\begin{pmatrix} X_{11} \\ X_{21} \end{pmatrix}, \begin{pmatrix} X_{12} \\ X_{22} \end{pmatrix}, \ldots, \begin{pmatrix} X_{1n} \\ X_{2n} \end{pmatrix}$$

be i.i.d. random vectors with mean $\mu = (\mu_1, \mu_2) = (1, -1)$ and variance $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}$. Let

$$\bar{X}_{n,1} = \frac{1}{n}\sum_{i=1}^{n} X_{1i}, \quad \bar{X}_{n,2} = \frac{1}{n}\sum_{i=1}^{n} X_{2i}$$

and define $Y_n = \bar{X}_{n,1}/\bar{X}_{n,2}$. Find the limiting distribution of $\sqrt{n}(Y_n + 1)$. **Solution:** Let $g(x, y) = x/y$. Applying the multivariate delta method (Thm 5.15) tells us that

$$\sqrt{n}(Y_n + 1) = \sqrt{n}(g(\bar{X}_{1,n}, \bar{X}_{2,n}) - g(\mu)) \rightsquigarrow N(0, \nabla g(\mu)^\top \Sigma \nabla g(\mu)).$$

Now we compute $\nabla g(x_1, x_2) = (1/x_2, -x_1/x_2^2)$. Plugging in $(x_1, x_2) = \mu = (1, -1)$, we get $\nabla g(\mu) = (-1, -1)$. Therefore

$$\nabla g(\mu)^\top \Sigma \nabla g(\mu) = (-1\ -1) \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix} \begin{pmatrix} -1 \\ -1 \end{pmatrix} = 4.$$

So $\sqrt{n}(Y_n + 1) \rightsquigarrow \mathcal{N}(0, 4)$.

2 Parameter estimation and MSE

Problem 4 Let $X_1, \ldots, X_n$ be i.i.d. $\mathsf{Unif}(0, \theta)$ variables for some unknown $\theta > 0$.

1. Write down a valid statistical model for the resulting data. **Solution:** $\{\mathsf{Unif}(0, \theta) : \theta > 0\}$

2. Compute the cdf, pdf, expectation, and variance of $\max_i X_i$. **Solution:** For $x \in [0, \theta]$ we have $\mathsf{P}(\max_i X_i \leq x) = \mathsf{P}(X_1 \leq x)^n = (x/\theta)^n$. The pdf can be obtained by differentiating the cdf, which gives nx^{n-1}/θ^n. The expectation can be computed by integrating directly

$$\mathsf{E}[\max_i X_i] = \int_0^\theta x \times nx^{n-1}\theta^{-n} dx = n\theta^{-n} \int_0^\theta x^n dx = \frac{n\theta}{n+1}.$$

Analogously,

$$\mathsf{E}[(\max_i X_i)^2] = \int_0^\theta x^2 \times nx^{n-1}\theta^{-n} dx = n\theta^{-n} \int_0^\theta x^{n+1} dx = \frac{n\theta^2}{n+2}.$$

Therefore, the variance is

$$\mathsf{V}[\max_i X_i] = \frac{n\theta^2}{n+2} - \left(\frac{n\theta}{n+1}\right)^2 = \left(\frac{n}{n+2} - \frac{n^2}{(n+1)^2}\right)\theta^2 = \frac{n}{(n+2)(n+1)^2}\theta^2.$$

3. Suppose we use $\hat{\theta} = \max_i X_i$ as an estimator for θ. Compute the MSE of this estimator. **Solution:**

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{bias}(\hat{\theta})^2 + \mathsf{V}[\hat{\theta}] = \left(\frac{n\theta}{n+1} - \theta\right)^2 + \frac{n}{(n+2)(n+1)^2}\theta^2 = \frac{2}{(n+2)(n+1)}\theta^2.$$

4. Find $a > 0$ such that $a \max_i X_i$ is an unbiased estimator of θ and compute its MSE. How does it compare to the MSE of $\hat{\theta} = \max_i X_i$? **Solution:** We debias $\hat{\theta}$ by choosing $a = (n+1)/n$. Since the bias is zero, the MSE is just the variance:

$$\mathrm{MSE}(\frac{n+1}{n}\hat{\theta}) = \mathsf{V}[\frac{n+1}{n}\hat{\theta}] = \frac{(n+1)^2}{n^2}\frac{n}{(n+2)(n+1)^2}\theta^2 = \frac{1}{n(n+2)}\theta^2.$$

This MSE is smaller. To see this note that $1/(n(n+2)) < 2/((n+1)(n+2))$ for $n \geq 2$.

Problem 5 Let $X_1, \ldots, X_n \sim N(\theta, 1)$. Define

$$Y_i = \begin{cases} 1 & \text{if } X_i > 0, \\ 0 & \text{if } X_i \leq 0. \end{cases}$$

We're interested in estimating $\psi = \mathsf{P}(Y_1 = 1)$.

1. Find f such that $\psi = f(\theta)$. **Solution:** Write $X_1 = \theta + Z$ for a standard normal Z. Then $\mathsf{P}(Y_1 = 1) = \mathsf{P}(X_1 > 0) = \mathsf{P}(Z > -\theta) = 1 - \Phi(-\theta) = \Phi(\theta)$.

2. This motivates using the "plug-in" estimator $\hat{\psi} := f(\bar{X}_n)$ of ψ. Find $\mathrm{se}(\hat{\psi})$. **Solution:** By the delta method, $\Phi(\bar{X}_n) \approx \mathcal{N}(\Phi(\theta), \frac{\Phi'(\theta)^2}{n})$. I.e. the variance is approximately $\Phi'(\theta)^2/n$, so the standard error is $\Phi'(\theta)/\sqrt{n} = \phi(\theta)/\sqrt{n}$, where ϕ is the standard normal pdf.

3. Construct an approximate 95% confidence interval for ψ. **Solution:** We use $\phi(\theta)/\sqrt{n} \approx \phi(\bar{X}_n)/\sqrt{n}$. So

$$\sqrt{n}\frac{\Phi(\bar{X}_n) - \psi}{\phi(\bar{X}_n)}$$

is approximately standard normal. Since $\mathsf{P}(Z \in [-2, 2]) \approx 0.95$, we get the confidence interval

$$\left(\hat{\psi} - 2\frac{\phi(\bar{X}_n)}{\sqrt{n}}, \hat{\psi} + 2\frac{\phi(\bar{X}_n)}{\sqrt{n}}\right).$$

4. Let $\bar{\psi} = \bar{Y}_n$ be another estimator of ψ. Find $\mathrm{se}(\bar{\psi})$ **Solution:** Note that $Y_i \sim \mathrm{Ber}(\Phi(\theta))$. Therefore $\mathsf{V}[\bar{Y}_n] = \Phi(\theta)(1 - \Phi(\theta))/n$, so $\mathrm{se}(\bar{Y}_n) = \sqrt{\Phi(\theta)(1 - \Phi(\theta))/n}$.

5. Suppose $\theta = 0$. Which is smaller: $\mathrm{se}(\bar{\psi})$ or $\mathrm{se}(\hat{\psi})$? **Solution:** We have $\mathrm{se}(\bar{\psi}) = \sqrt{\frac{1}{2} \times \frac{1}{2}/n} = 1/(2\sqrt{n})$, and $\mathrm{se}(\hat{\psi}) \approx \phi(0)/\sqrt{n} = 1/\sqrt{2\pi n}$. Since $1/\sqrt{2\pi} \leq 1/2$, the se of $\hat{\psi}$ is smaller than that of $\bar{\psi}$ (this holds for large n).

3 Maximum Likelihood Estimators

Problem 6 In each of the following cases write the likelihood function and compute the maximum likelihood estimator of the parameter based on the X_i:

1. $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$. **Solution:** The pdf for the normal is

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

so the log likelihood is

$$\ell_n(\mu, \sigma^2) = \sum_{i=1}^n \log f_{\mu,\sigma^2}(X_i) = \text{const.} - \frac{n}{2}\log(\sigma^2) - \sum_{i=1}^n \frac{(X_i - \mu)^2}{2\sigma^2}$$

We compute the derivative of ℓ_n with respect to μ and set it to zero:

$$0 = \partial_\mu \ell_n(\mu, \sigma^2) = -\frac{1}{2\sigma^2}\partial_\mu \sum_{i=1}^n (X_i - \mu)^2 = -\frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \mu)$$

We cancel the σ^2 and are left with

$$0 = \sum_{i=1}^n (X_i - \mu) = \left(\sum_{i=1}^n X_i\right) - n\mu,$$

so $\mu = \bar{X}_n$. Next, we compute the partial derivative of ℓ_n with respect to $\tau = \sigma^2$. For μ we use $\mu = \bar{X}_n$, since we already know this is the optimizer.

$$\partial_\tau \ell_n(\bar{X}_n, \tau) = -\frac{n}{2\tau} + \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{2\tau^2} = 0.$$

Solving gives

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

2. $X_1, \ldots, X_n$ are i.i.d. Pois(λ) random variables. **Solution:** The pdf for the Poisson distribution is

$$f_\lambda(k) = e^{-\lambda}\frac{\lambda^k}{k!},$$

so the log likelihood is

$$\begin{aligned} \ell_n(\lambda) &= \sum_{i=1}^n \log f_\lambda(X_i) = \sum_{i=1}^n [-\lambda + X_i \log\lambda - \log(X_i!)] \\ &= -n\lambda + \sum_i X_i \log\lambda + \text{const.} \end{aligned} \tag{2}$$

We take the derivative w.r.t. λ and set it to zero:

$$\ell_n'(\lambda) = -n + \frac{\sum_i X_i}{\lambda} = 0,$$

which gives $\lambda = \bar{X}_n$.

3. $X_1, \ldots, X_n$ are i.i.d. $\mathsf{Exp}(\lambda)$ random variables. **Solution:** The pdf for the exponential distribution is

$$f_\lambda(x) = \frac{1}{\lambda} e^{-x/\lambda}$$

so the log likelihood is

$$\begin{aligned} \ell_n(\lambda) &= \sum_{i=1}^n \log f_\lambda(X_i) = \sum_{i=1}^n [-\log \lambda - X_i/\lambda] \\ &= -n \log \lambda - (\sum_{i=1}^n X_i)/\lambda \end{aligned} \tag{3}$$

We take the derivative w.r.t. λ and set it to zero:

$$\ell_n'(\lambda) = -\frac{n}{\lambda} + \frac{\sum_{i=1}^n X_i}{\lambda^2} = 0,$$

which gives $\lambda = \bar{X}_n$.

4. $X_1, \ldots, X_n$ are i.i.d. $\mathrm{Unif}([0, \theta])$ random variables. **Solution:** The pdf for the uniform distribution is

$$f_\theta(x) = \begin{cases} \frac{1}{\theta} & x \in [0, \theta] \\ 0 & \text{otherwise} \end{cases}$$

Therefore the *likelihood* function is:

$$L_n(\theta) = \prod_{i=1}^n f_\theta(X_i) = \begin{cases} \frac{1}{\theta^n} & X_i \in [0, \theta] \text{ for all } i = 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

In other words,

$$L_n(\theta) = \begin{cases} \frac{1}{\theta^n} & \theta \geq \max_i X_i \\ 0 & \text{otherwise} \end{cases}$$

This function is maximized at $\theta = \max_i X_i$, so this is the MLE.

5. $X_1, \ldots, X_n$ are i.i.d. Bernoulli(p) **Solution:** The pdf for the Bernoulli distribution is

$$f_p(x) = p^x (1-p)^{1-x}$$

so the log likelihood is

$$\begin{aligned}
\ell_n(\lambda) &= \sum_{i=1}^{n} \log f_p(X_i) = \sum_{i=1}^{n} [X_i \log p + (1 - X_i) \log(1-p)] \\
&= \sum_i X_i \log p + (n - \sum_i X_i) \log(1-p),
\end{aligned} \tag{4}$$

We take the derivative w.r.t. p and set it to zero:

$$0 = \frac{\sum_i X_i}{p} - \frac{n - \sum_i X_i}{1-p}$$

, so

$$0 = (1-p) \sum_i X_i - (n - \sum_i X_i)p = \sum_i X_i - np,$$

so $p = \bar{X}_n$ is the MLE.

Definition 1 *(Fisher Information for Random Variables) Let X be a random variable with pdf $f_\theta(x)$ where θ is some parameter. The Fisher Information $I(\theta)$ of θ is defined by:*

$$I(\theta) = V_\theta \left(\frac{\partial \log f_\theta(x)}{\partial \theta} \right)$$

Under certain regularity conditions, one can show that this is equivalent to:

$$I(\theta) = -\mathsf{E}_\theta \left[\frac{\partial^2 \log f_\theta(x)}{\partial \theta^2} \right] = -\int \frac{\partial^2 \log f_\theta(x)}{\partial \theta^2} f_\theta(x) dx$$

Theorem 1 *(Asymptotic Normality of MLE) Let $\hat{\theta}_n$ be the maximum likelihood estimator of θ. Then under appropriate regularity conditions:*

$$\sqrt{n}(\hat{\theta}_n - \theta) \to^d N(0, I(\theta)^{-1})$$

Amongst these regularity conditions is the condition that the support of $f_\theta(x)$ does not depend on θ.

Problem 7 Consider the maximum likelihood estimators evaluated in parts 1-4 of problem 6.

1. For each part determine whether or not the corresponding Fisher information is well-defined. **Solution:** The Fisher information is not well-defined for the uniform distribution as the support changes with θ.

2. For each part where the Fisher information is well-defined determine it and the asymptotic variance of the estimator. **Solution:**
Normal distribution. For this part, it will be more convenient to think of the normal distribution as parameterized by μ, σ rather than by μ, σ^2. Therefore, we will compute
$$I(\theta) = -\mathsf{E}_{\mu,\sigma}\left[\nabla^2_{\mu,\sigma} \log f_{\mu,\sigma}(X)\right].$$
Now, the pdf for the normal is
$$f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$
Taking the log, we get
$$\log f_{\mu,\sigma}(x) = \text{const} - \log(\sigma) - (x-\mu)^2/(2\sigma^2).$$
We now take the partial derivatives. We get
$$\begin{aligned}\partial^2_\mu \log f_{\mu,\sigma} &= -1/\sigma^2,\\ \partial_\mu\partial_\sigma \log f_{\mu,\sigma} &= 2(\mu-x)/\sigma^3,\\ \partial^2_\sigma \log f_{\mu,\sigma} &= 1/(\sigma^2) - 3(x-\mu)^2/\sigma^4\end{aligned} \tag{5}$$
Putting the second partial derivatives together in a matrix and taking the expectation, we get
$$\begin{aligned}I(\mu,\sigma) &= -\mathsf{E}_{\mu,\sigma}\left[\nabla^2_{\mu,\sigma} \log f_{\mu,\sigma}(X)\right]\\ &= -\mathsf{E}_{\mu,\sigma}\left[\begin{pmatrix} -\frac{1}{\sigma^2} & \frac{2(\mu-X)}{\sigma^3} \\ \frac{2(\mu-X)}{\sigma^3} & \frac{1}{\sigma^2} - 3\frac{(X-\mu)^2}{\sigma^4}\end{pmatrix}\right] = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2}\end{pmatrix}.\end{aligned} \tag{6}$$
Here, $\mathsf{E}_{\mu,\sigma}[f(X)]$ denotes the expectation of $f(X)$, where $X \sim \mathcal{N}(\mu, \sigma^2)$. The theorem therefore tells us that asymptotically, the covariance matrix of $(\hat{\mu}, \hat{\sigma})$ is
$$\mathsf{V}[(\hat{\mu}, \hat{\sigma})] \approx \frac{1}{n} I(\mu,\sigma)^{-1} = \frac{1}{n}\begin{pmatrix} \sigma^2 & 0 \\ 0 & \frac{\sigma^2}{2}\end{pmatrix},$$
$\hat{\mu}, \hat{\sigma}$ are the MLEs for μ and σ.
Poisson distribution. The pmf of the Poisson distribution is
$$f_\lambda(x) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$$
Taking the log, we get
$$\log f_\lambda(x) = -\lambda + x \log \lambda - \log(x!).$$

Taking the second derivative with respect to λ, we get

$$\frac{\partial^2}{\partial\lambda^2}\log f_\lambda(x) = \frac{\partial^2}{\partial\lambda^2}(x\log\lambda - \lambda) = -\frac{x}{\lambda^2}$$

Thus the Fisher Information is:

$$I(\lambda) = \mathsf{E}_\lambda[X/\lambda^2] = \lambda/\lambda^2 = \frac{1}{\lambda}$$

By the preceding theorem, the asymptotic variance is then λ.

Exponential distribution. The pdf of the exponential distribution is

$$f_\lambda(x) = \frac{1}{\lambda}e^{-x/\lambda}.$$

Taking the log, we get

$$\log f_\lambda(x) = -\log\lambda - x/\lambda.$$

Taking the second derivative with respect to λ, we get

$$\frac{\partial^2}{\partial\lambda^2}\log f_\lambda(x) = \frac{\partial^2}{\partial\lambda^2}(-\log\lambda - \frac{x}{\lambda}) = \frac{1}{\lambda^2} - \frac{2x}{\lambda^3}$$

Thus the Fisher Information is:

$$I(\lambda) = -\mathsf{E}_\lambda\left[\frac{1}{\lambda^2} - \frac{2X}{\lambda^3}\right] = -\frac{1}{\lambda^2} + 2\frac{\lambda}{\lambda^3} = \frac{1}{\lambda^2}.$$

By the preceding theorem, the asymptotic variance is then λ^2.

Problem 8 Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$. Let τ be such that $\mathsf{P}(X < \tau) = 0.95$.

1. Write τ in terms of μ and σ. **Solution:** Let $Z \sim N(0,1)$, so $(X-\mu)/\sigma \sim Z$. We have:

$$\mathsf{P}(X < \tau) = 0.95 \tag{7}$$

$$\mathsf{P}\left(\frac{X-\mu}{\sigma} < \frac{\tau-\mu}{\sigma}\right) = 0.95 \tag{8}$$

$$\mathsf{P}\left(Z < \frac{\tau-\mu}{\sigma}\right) = 0.95, \tag{9}$$

$$\frac{\tau-\mu}{\sigma} = z_{5\%} \tag{10}$$

$$\tau = \mu + z_{5\%}\sigma. \tag{11}$$

2. Write down an estimator $\hat{\tau}$ of τ in terms of the MLEs $\hat{\mu}, \hat{\sigma}$ for μ and σ, respectively. **Solution:** We plug in $\hat{\mu}$ for μ and $\hat{\sigma}$ for σ in the formula for τ from the previous part: $\hat{\tau} = \hat{\mu} + z_{5\%}\hat{\sigma}$, where $\hat{\mu}, \hat{\sigma}$ are the MLEs for the normal distribution parameters:

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

3. Find an expression for an approximate $1-\alpha$ confidence interval for τ. Hint: use the asymptotic variance of $\hat{\mu}, \hat{\sigma}$ computed in part 2 of problem 7. Then apply the delta method to get the asymptotic variance of τ. **Solution:** Let's use the multivariate delta method. We have $\hat{\tau} = g(\hat{\mu}, \hat{\sigma}) = \hat{\mu} + z_{5\%}\hat{\sigma}$, so

$$\nabla g = \begin{pmatrix} \partial g/\partial \hat{\mu} \\ \partial g/\partial \hat{\sigma}, \end{pmatrix} = \begin{pmatrix} 1 \\ z_{5\%} \end{pmatrix}$$

Now, using the Fisher information matrix computed in part 2 of problem 7, we have that the asymptotic covariance matrix of $(\hat{\mu}, \hat{\sigma})$ is

$$J_n = \frac{1}{n}\begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}^{-1} = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{\sigma^2}{2n}. \end{pmatrix}.$$

The standard error estimate for our new parameter $\hat{\tau}$ is therefore

$$\hat{\mathrm{se}}(\hat{\tau}) = \sqrt{\nabla g^T \hat{J}_n \nabla g} = \sqrt{\frac{\hat{\sigma}^2}{n} + \frac{\hat{\sigma}^2}{2n}z_{5\%}^2} = \hat{\sigma}\sqrt{n^{-1}(1 + z_{5\%}^2/2)}.$$

A $1-\alpha$ confidence interval for $\hat{\tau}$, then, is

$$C_n = \left([\hat{\mu} + \hat{\sigma}z_{5\%}] - z_{\alpha/2}\hat{\sigma}\sqrt{n^{-1}(1 + z_{5\%}^2/2)},\quad [\hat{\mu} + \hat{\sigma}z_{5\%}] + z_{\alpha/2}\hat{\sigma}\sqrt{n^{-1}(1 + z_{5\%}^2/2)}\right)$$

4 Method of Moments and EM

Problem 9 The gamma distribution is given by

$$f_{k,\theta}(x) = \frac{1}{\Gamma(k)\theta^k}x^{k-1}e^{-x/\theta},$$

where $\Gamma(k)$ is the gamma function. The parameter k is known as a shape parameter, and θ is called the scale parameter. The first and second moments are

$$\begin{aligned} \alpha_1 &= \mathsf{E}_{k,\theta}[X] = k\theta, \\ \alpha_2 &= \mathsf{E}_{k,\theta}[X^2] = (k+k^2)\theta^2 \end{aligned} \tag{12}$$

Given $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f_{k,\theta}$ find the method of moments estimators $\hat{k}, \hat{\theta}$

Solution: We solve the system of equations

$$\begin{cases} \alpha_1 = k\theta \\ \alpha_2 = (k+k^2)\theta^2 \end{cases}$$

for k and θ. Note that $\alpha_2 - \alpha_1^2 = k\theta^2$, and $\alpha_1 = k\theta$. Therefore,

$$\theta = \frac{k\theta^2}{k\theta} = \frac{\alpha_2 - \alpha_1^2}{\alpha_1}.$$

Now that we have θ, we can solve for k:

$$k = \alpha_1/\theta = \frac{\alpha_1^2}{\alpha_2 - \alpha_1^2}.$$

We now substitute

$$\hat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j, \quad j = 1, 2$$

for α_1, α_2 to conclude.

Problem 10 Let $X_1, \ldots, X_n$ be i.i.d. with pdf

$$f_\mu(x) = \frac{1}{2}\left[\frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}}\right] + \frac{1}{2}\left[\frac{1}{\sqrt{2\pi}} e^{-\frac{(x+\mu)^2}{2}}\right],$$

where $\mu \geq 0$. This is an equally-weighted mixture of the two Gaussian distributions $\mathcal{N}(\mu, 1)$ and $\mathcal{N}(-\mu, 1)$.

1. What is the log likelihood $\ell_n(\mu)$? **Solution:**

$$\ell_n(\mu) = \sum_{i=1}^{n} \log f_\mu(X_i) = \sum_{i=1}^{n} \log\left(\frac{1}{2}\left[\frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i-\mu)^2}{2}}\right] + \frac{1}{2}\left[\frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i+\mu)^2}{2}}\right]\right)$$

2. Let $Y_i \in \{-1, 1\}$ be the hidden variable telling us the membership of X_i to either $\mathcal{N}(-\mu, 1)$ or $\mathcal{N}(\mu, 1)$. What is the log likelihood $\ell_n(\mu)$ in the case that we observe both the X_i's and the Y_i's? What is the MLE for μ in this case? **Solution:** First note that

$$f_\mu(X_i \mid Y_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i - Y_i\mu)^2}{2}}.$$

Then

$$f_\mu(X_i, Y_i) = \mathsf{P}(Y_i) f_\mu(X_i \mid Y_i) = \frac{1}{2}\frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i - Y_i\mu)^2}{2}}.$$

From here, we get

$$\begin{aligned} \ell_n(\mu) &= \sum_{i=1}^n \log f_\mu(X_i, Y_i) \\ &= \text{const} - \sum_{i=1}^n \frac{(X_i - Y_i\mu)^2}{2} \\ &= \text{const} + \left[\sum_{i=1}^n Y_i X_i\right] \mu - \frac{n}{2}\mu^2, \end{aligned} \tag{13}$$

using that $Y_i^2 = 1$. Maximizing, we get

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i Y_i = \frac{1}{n}\left(\sum_{i:Y_i=1} X_i + \sum_{i:Y_i=-1} (-X_i)\right).$$

3. Let μ_k be our current guess for μ. Compute $\hat{Y}_i = \mathsf{E}_{\mu_k}[Y_i \mid X_i]$. This is the E-step of the EM algorithm.) Use this to construct an approximation $\hat{\ell}_n(\mu)$ to the log likelihood in the observed-label case. **Solution:** Using Bayes Rule,

$$\begin{aligned} \mathsf{P}_\mu(Y_i = 1 \mid X_i) &= \frac{\mathsf{P}_\mu(Y_i = 1, X_i)}{\mathsf{P}_\mu(X_i)} \\ &= \frac{f_\mu(X_i \mid Y_i = 1)\mathsf{P}_\mu(Y_i = 1)}{f_\mu(X_i \mid Y_i = 1)\mathsf{P}_\mu(Y_i = 1) + f_\mu(X_i \mid Y_i = -1)\mathsf{P}_\mu(Y_i = -1)} \\ &= \frac{f_\mu(X_i \mid Y_i = 1)}{f_\mu(X_i \mid Y_i = 1) + f_\mu(X_i \mid Y_i = -1)} \\ &= \frac{e^{-\frac{1}{2}(X_i - \mu)^2}}{e^{-\frac{1}{2}(X_i - \mu)^2} + e^{-\frac{1}{2}(X_i + \mu)^2}} = \frac{e^{X_i\mu}}{e^{X_i\mu} + e^{-X_i\mu}}. \end{aligned} \tag{14}$$

Analogously,

$$\mathsf{P}(Y_i = -1 \mid X_i) = \frac{e^{-X_i\mu}}{e^{X_i\mu} + e^{-X_i\mu}}.$$

Therefore,

$$\hat{Y}_i = \mathsf{E}[Y_i \mid X_i] = \frac{e^{X_i\mu} - e^{-X_i\mu}}{e^{X_i\mu} + e^{-X_i\mu}} = \tanh(X_i\mu).$$

Substituting $\hat{Y}_i$ for Y_i in (13), we get

$$\hat{\ell}_n(\mu) = \text{const} + \left[\sum_{i=1}^{n} \tanh(X_i\mu_k)X_i\right]\mu - \frac{n}{2}\mu^2$$

4. Find $\mu_{k+1} = \operatorname{argmax}_\mu \hat{\ell}_n(\mu)$ to get an update rule $\mu_k \mapsto \mu_{k+1}$. (This is the M step of the EM algorithm.) Check that this rule makes sense intuitively, and compare it to the estimator $\hat{\mu}$ from part 1. **Solution:** Solving

$$0 = \hat{\ell}_n'(\mu) = \sum_{i=1}^{n} \tanh(X_i\mu_k)X_i - n\mu,$$

we get

$$\mu_{k+1} = \operatorname*{argmax}_\mu \hat{\ell}_n(\mu) = \frac{1}{n}\sum_{i=1}^{n} \tanh(X_i\mu_k)X_i.$$

This update rule makes sense intuitively. Comparing to the ideal estimator from part 1, we see that the weight $\tanh(X_i\mu_k)$ is a "soft" version of the "hard" weight $Y_i = \pm 1$, which we don't have access to. (Note that tanh increases monotonically from -1 at $-\infty$ to +1 at $+\infty$.)

5. What happens if you initialize the algorithm at $\mu_0 = 0$? **Solution:** Since $\tanh(0) = 0$, we get $\mu_1 = 0$, so we're stuck at zero.

6. Compute $\mathsf{E}[X_1^2]$ in terms of μ. Based on this, construct the method of moments estimator $\hat{\mu}$ of μ. **Solution:** $\mathsf{E}[X_1^2] = \mu^2 + 1$. Therefore, we replace $\mathsf{E}[X_1^2]$ by the plug-in estimator and solve the resulting equation for μ:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \mu^2 + 1 \implies \hat{\mu} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - 1}.$$

7. Suppose $\mu = 1$. What is the asymptotic variance of $\hat{\mu}$? **Solution:** Let $Y_i = X_i^2 - 1$, so that $\hat{\mu} = \sqrt{\bar{Y}_n}$. Note $\mathsf{E}[Y_i] = \mu^2 = 1$ and $\mathsf{V}[Y_i] = \mathsf{E}[(X_i^2-1)^2] - \mu^4 = \mathsf{E}[((Z+1)^2-1)^2] - 1 = \mathsf{E}[Z^4 + 4Z^3 + 4Z^2] - 1 = 6$.
Therefore by the CLT, $\bar{Y}_n \approx \mathcal{N}(1, 6/n)$. Now let $g(y) = \sqrt{y}$, which has derivative $g'(y) = 1/2\sqrt{y}$, so $g'(1) = 1/2$. Therefore by the delta method, $\hat{\mu} = \sqrt{\bar{Y}_n} \approx \mathcal{N}(1, \left(\frac{1}{2}\right)^2 \frac{6}{n})$, so we get an asymptotic variance of $3/2n$.

	Second decimal place of Z									
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

The table lists $P(Z \leq z)$ where $Z \sim N(0,1)$ for positive values of z.