SB2.1 Foundations of Statistical Inference
Problem Sheet 4

1. The risks for five decision rules $\delta_1, \dots, \delta_5$ depend on the value of a nonnegative parameter θ and are given in the table below.

	δ_1	δ_2	δ_3	δ_4	δ_5
$0 \leq \theta < 1$	10	10	7	6	8
$1 \leq \theta < 2$	8	11	8	5	10
$2 \leq \theta$	15	11	12	14	14

(a) Which decision rules strictly dominate δ_1?

(b) Which decision rules are admissible?

(c) Which decision rule is minimax?

(d) Consider a prior π on θ which is a uniform distribution on $[0, 5]$. Which rule is Bayes with respect to π and what is its Bayes risk?

Answer: (a)

$$R(\theta, \delta_1) \geq R(\theta, \delta_3) \text{ and } R(\theta, \delta_4)$$

for all $\theta > 0$ with strict inequality for some θ, so so δ_3, δ_4 strictly dominate δ_1.

(b) δ_1 and δ_5 are inadmissible and $\delta_2, \delta_3, \delta_4$ are admissible.

(c) The minimax rule chooses δ to minimize

$$\max_{\theta} R(\theta, \delta)$$

Values of the maximum risk for $\delta_1, \dots, \delta_5$ are 15,11,12,14 and 14, so δ_2 is minimax.

(d) Bayes rule minimizes

$$r(\pi, \delta) = \int_0^5 R(\theta, \delta) \cdot \frac{1}{5} d\theta.$$

Values of $5 \times$ the integral are 63,54,51,53, 60 so δ_3 is Bayes. The Bayes risk is $51/5$.

2. Let $\{X_i\}_{i=1}^n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2 I_p)$ with $\mu = (\mu_1, \dots, \mu_p)^\top$ and $X_i = (X_{i1}, \dots, X_{ip})^\top$. Consider the quadratic loss function $L(\mu, \hat{\mu}) = \|\mu - \hat{\mu}\|^2$. Assume that σ^2 is known.

(a) Find the risk of the MLE $\hat{\mu} = \bar{X}$.

(b) Find values of a for which

$$\tilde{\mu} = \left(1 - \frac{a}{\|\bar{X}\|^2}\right) \bar{X}$$

strictly dominates the maximum likelihood estimator.

(c) For the value of a for which the risk $R(\mu, \tilde{\mu})$ is minimized, show that

$$R(\mu, \tilde{\mu}) \leq \frac{p\sigma^2}{n} - \frac{(p-2)^2 \sigma^4 / n^2}{\|\mu\|^2 + p\sigma^2/n}.$$

[*Hint: Jensen's inequality* $\mathbb{E}[1/V] \geq 1/\mathbb{E}V$ *for non-negative valued random variable* V].

(d) Assume that only a small number $k \ll p$ of μ_i is non-zero and that they are bounded, i.e. $|\mu_i| \leq M$. Show that the risk $R(\mu, \tilde{\mu})$ remains bounded as $p \to \infty$ (while keeping n, k, M and σ^2 fixed).

Answer:

(a) We have

$$\mathbb{E}\|\mu - \bar{X}\|^2 = \sum_{j=1}^{p} \mathbb{E}\left[(\mu_j - \bar{X}_j)^2\right] = p\mathrm{Var}(\bar{X}_1) = \frac{p\sigma^2}{n}.$$

(b) Closely following the lectures, we have:

$$\begin{aligned}
R(\mu, \tilde{\mu}) &= \mathbb{E}_{X|\mu}\left\|\mu - \bar{X} + \frac{a}{\|\bar{X}\|^2}\bar{X}\right\|^2 \\
&= \underbrace{\mathbb{E}_{X|\mu}\left\|\mu - \bar{X}\right\|^2}_{=R(\mu,\bar{X})} + a^2\mathbb{E}_{X|\mu}\frac{1}{\|\bar{X}\|^2} + 2a\underbrace{\mathbb{E}_{X|\mu}\left(\frac{(\mu - \bar{X})^\top \bar{X}}{\|\bar{X}\|^2}\right)}_{=-\frac{(p-2)\sigma^2}{n}\mathbb{E}_{X|\mu}\frac{1}{\|\bar{X}\|^2}} \\
&= R(\mu, \bar{X}) - \left(2a\frac{(p-2)\,\sigma^2}{n} - a^2\right)\mathbb{E}_{X|\mu}\frac{1}{\|\bar{X}\|^2},
\end{aligned}$$

where we put $g_j(x) = \frac{x_j}{\|x\|^2}$ and apply Stein's lemma noting $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}I_p)$ as follows:

$$\begin{aligned}
\mathbb{E}\left(\frac{(\mu - \bar{X})^\top \bar{X}}{\|\bar{X}\|^2}\right) &= \sum_{j=1}^{p}\mathbb{E}\left[(\mu_j - \bar{X}_j)\frac{\bar{X}_j}{\|\bar{X}\|^2}\right] = \sum_{j=1}^{p}\mathbb{E}\left[(\mu_j - \bar{X}_j)\,g_j(\bar{X})\right] \\
&= -\frac{\sigma^2}{n}\sum_{j=1}^{p}\mathbb{E}\left[\frac{\partial g_j}{\partial x_j}(\bar{X})\right] = -\frac{\sigma^2}{n}\sum_{j=1}^{p}\mathbb{E}\left[\frac{\|\bar{X}\|^2 - 2\bar{X}_j^2}{\|\bar{X}\|^4}\right] \\
&= -\frac{(p-2)\sigma^2}{n}\mathbb{E}\frac{1}{\|\bar{X}\|^2}.
\end{aligned}$$

We see that the values $a \in \left(0, \frac{2(p-2)\sigma^2}{n}\right)$ lead to a strictly lower risk for all μ and that the risk is minimized for $a = \frac{(p-2)\sigma^2}{n}$ and equal to $\frac{p\sigma^2}{n} - \frac{(p-2)^2\sigma^4}{n^2}\mathbb{E}\frac{1}{\|\bar{X}\|^2}$.

(c) From Jensen's inequality,

$$\mathbb{E}\frac{1}{\|\bar{X}\|^2} \geq \frac{1}{\mathbb{E}\,\|\bar{X}\|^2} = \frac{1}{\sum_{j=1}^{p}\mathbb{E}\bar{X}_j^2} = \frac{1}{\sum_{j=1}^{n}\left(\mu_j^2 + \frac{\sigma^2}{n}\right)} = \frac{1}{\|\mu\|^2 + \frac{p\sigma^2}{n}},$$

and inserting back gives the desired bound.

(d) We have $0 \leq \|\mu\|^2 \leq kM^2$. Now

$$\begin{aligned}
R(\mu, \tilde{\mu}) &\leq \frac{(p\sigma^2/n)\|\mu\|^2 + (\sigma^4/n^2)(p^2 - (p-2)^2)}{\|\mu\|^2 + p\sigma^2/n} \\
&\leq \|\mu\|^2 + \frac{\sigma^2}{n}\frac{4p-4}{p} \\
&\leq kM^2 + 4\frac{\sigma^2}{n}.
\end{aligned}$$

This is in contrast to MLE whose risk diverges regardless of the form of μ.

3. Consider the linear regression model

$$Y_i = \beta^\top x_i + \epsilon_i, \quad \beta \in \mathbb{R}^p, \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0,1)$$

with $x_i = [x_{i,1}, \cdots, x_{i,p}]^\top$, $i = 1, \ldots, n$.

(a) Write the model as $Y = X\beta + \epsilon$ where $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$ and $\epsilon \in \mathbb{R}^n$. Find the maximum likelihood estimator $\hat{\beta}$ of β and compute its bias and variance-covariance matrix.

(b) Assume that $X^\top X = nI_p$ where I_p is the p-dimensional identity matrix. Show that $\hat{\beta}$ is not admissible as soon as $p > 2$ and find an estimator which strictly dominates $\hat{\beta}$.

[*Hint: Re-express likelihood in terms of* $Z = X^\top Y$ *and use Question* 2].

Answer:

(a) We have the likelihood given by

$$L(\beta) \propto e^{-\|Y-X\beta\|^2/2},$$

so that MLE is

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y,$$

with

$$\mathbb{E}[\hat{\beta}] = \beta, \mathrm{Var}(\hat{\beta}) = (X^\top X)^{-1}.$$

i.e. it is unbiased.

(b) If $X^\top X = nI_p$, then $\hat{\beta} = n^{-1}X^\top Y = n^{-1}Z$, denoting $Z = X^\top Y$.

By expanding $\|Y - X\beta\|^2$ and keeping terms that depend on β, we have

$$\begin{aligned} \|Y - X\beta\|^2 &= \text{const} + \beta^\top X^\top X\beta - 2Y^\top X\beta \\ &= \text{const} + n\beta^\top\beta - 2Z^\top\beta \\ &= \text{const} + n\|\beta - n^{-1}Z\| \end{aligned}$$

so that the likelihood can be written as

$$L(\beta) \propto e^{-\frac{n\|n^{-1}Z-\beta\|^2}{2}}$$

so it is a Gaussian likelihood for a single observation $n^{-1}Z \sim \mathcal{N}(\beta, I_p/n)$. Using the previous question, we obtain that the MLE $\hat{\beta}$ is not admissible and that the estimator

$$\tilde{\beta} = \left(1 - \frac{(p-2)}{n\|n^{-1}Z\|^2}\right) n^{-1}Z = \left(1 - \frac{p-2}{n\|\hat{\beta}\|^2}\right)\hat{\beta}$$

strictly dominates $\hat{\beta}$.

4. Consider a hierarchical model

$$\{X_{ij}\}_{j=1}^{n_i} | \theta_i \sim \text{Bern}(\theta_i), \ \{\theta_i\}_{i=1}^k | \alpha, \beta \sim \text{Beta}(\alpha, \beta).$$

Write $\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$ and denote the observed sample mean and sample variance of $\{\bar{X}_i\}_{i=1}^k$ by m and v, respectively.

(a) Find $\hat{\alpha}$ and $\hat{\beta}$ using empirical Bayes, i.e. by matching the mean and variance of Beta (α, β) to m and v.

(b) Show that the posterior mean $\mathbb{E}\left[\theta_i | \{X_{ij}\}, \hat{\alpha}, \hat{\beta}\right]$ is a weighted average of m and $\bar{X}_i$.

Answer: The mean and variance of Beta (α, β) are given by $\frac{\alpha}{\alpha+\beta}$ and $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. Setting to m and v and solving for α and β gives

$$\hat{\alpha} = \frac{m\left(m - m^2 - v\right)}{v}, \quad \hat{\beta} = \frac{(1-m)\left(m - m^2 - v\right)}{v}.$$

The posterior $p\left(\theta_i | \{X_{ij}\}, \hat{\alpha}, \hat{\beta}\right)$ of θ_i is Beta $\left(\hat{\alpha} + n_i\bar{X}_i, \hat{\beta} + n_i - n_i\bar{X}_i\right)$ and the posterior mean is

$$\begin{aligned}
\mathbb{E}\left[\theta_i | \{X_{ij}\}, \hat{\alpha}, \hat{\beta}\right] &= \frac{\hat{\alpha} + n_i\bar{X}_i}{\hat{\alpha} + \hat{\beta} + n_i} \\
&= \frac{\frac{(m-m^2-v)}{v} m + n_i\bar{X}_i}{\frac{(m-m^2-v)}{v} + n_i},
\end{aligned}$$

which takes form of the weighted average.

5. Let θ be a real-valued parameter and $f(x \mid \theta)$ be the probability density function of an observation x, given θ. Let H_0 be the hypothesis that $\theta = \theta_0$ and H_1 be the hypothesis that $\theta \neq \theta_0$ and consider a prior on H_1 given by $\theta \sim g(\theta)$. The prior probability for H_0 is β.

   (a) Write down expressions for the (joint) distribution $P(H, \theta)$ (observe that H can take only two values), the marginal distribution $P(x)$ and $P(H_0, \theta|x)$ and $P(H_1, \theta|x)$.

   (b) Derive an expression for $\pi(H_1 \mid x)$, the posterior probability of H_1.

   Now, suppose that $x_1, \ldots, x_n$ is a sample from a normal distribution with mean θ and variance v. Let $\beta = 1/2$ and let

   $$g(\theta) = (2\pi w^2)^{-1/2} \exp\left\{-\theta^2/(2w^2)\right\}$$

   for $-\infty < \theta < \infty$.

   Show that, if $\theta_0 = 0$ and the sample mean is observed to be $10(v/n)^{1/2}$ then

   (c) The likelihood ratio frequentist test of size $\alpha = 0.05$ will reject H_0 for any value of n;

   (d) The posterior probability of H_0 converges to 1, as $n \to \infty$.

   (e) Comment on the apparent contradiction between (c) and (d).

Answer: The answer touches on some theory of distributions from outside the course. Let the sample space for (H, θ) be $\Psi = (H_0, \theta_0) \cup \{(H_1, \theta), \theta \in \Theta\}$. On this space

$$\begin{aligned}
P(H, \theta) &= P(\theta|H)P(H) \\
&= \begin{cases} \beta & \text{if } (H, \theta) = (H_0, \theta_0), \text{ and} \\ g(\theta)(1-\beta) & \text{if } (H, \theta) \in \{(H_1, \theta), \theta \in \Theta\}. \end{cases}
\end{aligned}$$

where the meaning of the above expression is that the distribution of (H, θ) is the sum of a point mass at (H_0, θ_0) and a continuous density on $\{(H_1, \theta), \theta \in \Theta\}$. One way to think about such distributions is to say that this is the distribution of the variable X obtained in the following way:

with probability β we have $X = (H_0, \theta_0)$ and with probability $(1-\beta)$ we take $X = (H_1, Y)$ where Y is drawn accordingt to g. Thus, we have that

$$\begin{aligned} P(x) &= P(x|H_0)P(H_0) + P(x|H_1)P(H_1) \\ &= L(x,\theta_0)\beta + (1-\beta)\int_\Theta L(x,\theta)g(\theta)d\theta. \end{aligned}$$

and obviously

$$\begin{aligned} P(H,\theta|x) &= \frac{P(x|\theta,H)P(H,\theta)}{P(x)} \\ &\propto L(\theta;x)P(\theta|H)P(H) \\ &= \begin{cases} L(\theta_0;x)\beta & \text{if } (H,\theta) = (H_0,\theta_0), \text{ and} \\ L(\theta;x)g(\theta)(1-\beta) & (H,\theta) \in \{(H_1,\theta), \theta \in \Theta\}. \end{cases} \end{aligned}$$

(b)

$$P(H_1 \mid \mathbf{x}) = P(H_1)\frac{\int_\Theta L(\theta;x)g(\theta)d\theta}{P(x)}$$

$$P(H_1 \mid \mathbf{x}) = \frac{(1-\beta)\int_\Theta L(\theta;x)g(\theta)d\theta.}{L(\theta_0;x)\beta + (1-\beta)\int_\Theta L(\theta;x)g(\theta)d\theta.}$$

(c) Assuming $\theta_0 = 0$ in the classical test we reject H_0 if

$$|\bar{x}/\sqrt{v/n}| > 1.96$$

If $\bar{x} = 10(v/n)^{1/2}$ then $\bar{x}/\sqrt{v/n} > 10$, so H_0 is rejected for any value of n.

(d)

$$\begin{aligned} \pi(H_0 \mid \mathbf{x}) &= \frac{e^{-\frac{1}{2}\sum x_i^2/v}}{e^{-\frac{1}{2}\sum x_i^2/v} + \int_{-\infty}^{\infty} e^{-\frac{1}{2}\sum_1^n (x_i-\theta)^2/v}(2\pi w^2)^{-1/2}e^{-\theta^2/2w^2}d\theta} \\ &= \frac{1}{1 + \int_{-\infty}^{\infty} e^{-\frac{1}{2v}(n\theta^2 - 2n\bar{x}\theta)}(2\pi w^2)^{-1/2}e^{-\theta^2/2w^2}d\theta} \end{aligned}$$

The integral in the denominator converges to zero, because of the Lebesgue dominated convergence theorem and

$$e^{-\frac{1}{2v}(n\theta^2 - 20\sqrt{nv}\theta)} \to 0, \quad e^{-\frac{1}{2v}(n\theta^2 - 20\sqrt{nv}\theta)} = e^{-n(\theta - 10\sqrt{v}/\sqrt{n})^2/2v}e^{50} \leq e^{50}$$

(e) The classical test only considers the tails of the distribution under H_0, whereas the Bayesian approach considers the prior under H_1 and does not have a fixed type I error proabability.