Delft University of Technology Mekelweg 4
Faculty of Electrical Engineering, Mathematics and Computer Science 2628 CD Delft

Exam Multivariate Data Analysis (CS4070)
24 January 2023, 13:30-16:30

Probability distributions formulas:

Write $Z \sim \text{Exp}(\eta)$ if Z has the exponential distribution with parameter η. That is, its density is given by $p(z) = \eta e^{-\eta z}$ if $z \geq 0$.

Write $Z \sim N_p(\mu, \Sigma)$ if Z has the multivariate normal distribution with mean μ and covariance matrix Σ. That is, its density is given by

$$p(z) = (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu)\right).$$

Start of questions:

1. Consider the linear model

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i, \qquad 1 \leq i \leq n,$$

where $\{\varepsilon_i\} \overset{\text{iid}}{\sim} N(0, \sigma^2)$. Write $\beta = (\beta_1, \ldots, \beta_p)$ (viewed as column vector).

   (a) [1 pt]. Write down the least squares criterion for estimating β in matrix-vector notation.

   Assume for the remaining part of the question that the design matrix X has full column-rank and that the least squares estimator is given by $\hat{\beta} = (X^TX)^{-1}X^Ty$.

   (b) [1 pt]. Show that $\hat{\beta}$ is an unbiased estimator of β.

   (c) [2 pt]. Determine the variance of $\hat{\beta}^T x_{\text{new}}$ for $x_{\text{new}} \in \mathbb{R}^p$.

   (d) [1 pt]. Determine the variance of $\hat{\beta}^T x_{\text{new}} + \varepsilon_{\text{new}}$ with $\varepsilon_{\text{new}} \sim N(0, \sigma^2)$ independent of $\varepsilon_1, \ldots, \varepsilon_p$.

2. Assume

$$x_1, \ldots, x_n \mid \tau \sim N(0, \tau)$$

and $p(\tau) \propto \tau^{-A-1}e^{-B/\tau}\mathbf{1}_{[0,\infty)}(\tau)$, where $A, B > 0$ are (known) hyperparameters. We say that τ has the inverse-gamma distribution with parameters A and B.

   (a) [3 pt]. Derive the posterior distribution of τ. That is, the distribution of $\tau \mid x_1, \ldots, x_n$.
   Hint: the posterior distribution is an inverse-gamma distribution.

   (b) [2 pt]. Derive an expression for the posterior mode. If you do this by finding a stationary point, don't forget to verify that the stationary point corresponds to a maximum.

(c) [3 pt]. Determine the second derivative of the log-posterior density and the Laplace approximation of the posterior.

(d) [2 pt]. Suppose we fix $A = 1$. Explain how B can be determined by the method of empirical Bayes (also known as type II maximum likelihood). You don't need to carry out the actual computation.

(e) [3 pt]. Suppose we would use another prior on τ, say $\tau \sim \text{Exp}(\theta)$ (i.e. $p(\tau) = \theta e^{-\theta\tau}\mathbf{1}_{[0,\infty)}(\tau)$). Give the details of one step of Newton's method to numerically approximate the posterior mode.

3. Consider the model $y = X\theta + \varepsilon$ with $\varepsilon \sim N(0, I_n)$ (where I_n is the $n \times n$ identity matrix and X is an $n \times p$ matrix containing explanatory variables, considered fixed).

(a) [3 pt]. Assume $\theta \sim N(0, \gamma I_p)$ and that $\gamma > 0$ is known. Derive the posterior for θ.

(b) [3 pt]. We define a penalised estimator of θ by $\hat{\theta} = \text{argmin}_\theta \|y - X\theta\|^2 + \lambda\|\theta\|^2$ for fixed $\lambda > 0$. Derive an expression for $\hat{\theta}$. If you do this by finding a stationary point, don't forget to verify that the stationary point corresponds to a minimum.

(c) [1 pt]. Relate the posterior mean in (a) to the penalised estimator in (b).

4. Let π be a positive density on $\mathbb{R}^d$ we wish to sample from. For $\theta \in \mathbb{R}^d$ let $q(\theta, \cdot)$ be positive proposal densities on $\mathbb{R}^d$ for the Metropolis–Hastings algorithm. We accept a proposal from θ to θ° with probability

$$\alpha(\theta, \theta^\circ) = \min\left(1, \frac{\pi(\theta^\circ)}{\pi(\theta)}\frac{q(\theta^\circ, \theta)}{q(\theta, \theta^\circ)}\right).$$

Let $\bar{q}$ be the proposal adjusted by the Metropolis–Hastings acceptance rule.

(a) [3 pt]. Show that for all $\theta, \theta^\circ \in \mathbb{R}^d$

$$\pi(\theta)\bar{q}(\theta, \theta^\circ) = \pi(\theta^\circ)\bar{q}(\theta^\circ, \theta).$$

(b) [2 pt]. Show that when $\theta \sim \pi$ and we evolve the Metropolis–Hastings chain for one step from θ to θ° then $\theta^\circ \sim \pi$, by showing

$$\int_{\mathbb{R}^d} \pi(\theta)\bar{q}(\theta, \theta^\circ)d\theta = \pi(\theta^\circ).$$

5. [3 pt]. Let $x_1 = 1$, $x_2 = 2$ and $\sigma^2 = 5$. Suppose

$$\begin{aligned} y_i | f &\overset{\text{ind}}{\sim} N(f(x_i), \sigma^2) \\ f &\sim \text{GP}(0, K), \end{aligned}$$

so f is a Gaussian process with mean function 0 and covariance kernel K, where $K(x, \tilde{x}) = (1 + x\tilde{x})^2$, $(x, \tilde{x} \in \mathbb{R})$. Give the (joint)-distribution of

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

Answers

1. (a) Minimise $S(\beta) := \|y - X\beta\|^2$ with respect to β.

   (b) We have
   $$\mathrm{E}\,\hat{\beta} = (X^T X)^{-1} X^T \,\mathrm{E}\, y = (X^T X)^{-1} X^T X \beta = \beta.$$

   (c) We calculate
   $$\begin{aligned}\mathrm{Var}\Big(\hat{\beta}^T x_{\mathrm{new}}\Big) &= x_{\mathrm{new}}^T \mathrm{Cov}(\hat{\beta}) x_{\mathrm{new}} = x_{\mathrm{new}}^T (X^T X)^{-1} X^T \mathrm{Cov}(y) X (X^T X)^{-1} x_{\mathrm{new}} \\ &= \sigma^2 x_{\mathrm{new}}^T (X^T X)^{-1} X^T X (X^T X)^{-1} x_{\mathrm{new}} = \sigma^2 x_{\mathrm{new}}^T (X^T X)^{-1} x_{\mathrm{new}}.\end{aligned}$$

   (d) $\varepsilon_{\mathrm{new}}$ independent of $\varepsilon_1, \ldots, \varepsilon_p$ implies that $\varepsilon_{\mathrm{new}}$ is independent of $\hat{\beta}$ so we have
   $$\begin{aligned}\mathrm{Var}\Big(\hat{\beta}^T x_{\mathrm{new}} + \varepsilon_{\mathrm{new}}\Big) &= \mathrm{Var}\Big(\hat{\beta}^T x_{\mathrm{new}}\Big) + \mathrm{Var}(\varepsilon_{\mathrm{new}}) = \sigma^2 x_{\mathrm{new}}^T (X^T X)^{-1} x_{\mathrm{new}} + \sigma^2 \\ &= \sigma^2 \left(x_{\mathrm{new}}^T (X^T X)^{-1} x_{\mathrm{new}} + 1\right).\end{aligned}$$

2. (a) We have
   $$\begin{aligned}p(\tau \mid x) \propto p(x \mid \tau) p(\tau) &\propto \tau^{-n/2} \exp\left(-\frac{1}{2\tau} \sum_i x_i^2\right) \tau^{-A-1} e^{-B/\tau} \mathbf{1}_{[0,\infty)}(\tau) \\ &= \tau^{-A_p - 1} \exp\left(-B_p/\tau\right) \mathbf{1}_{[0,\infty)}(\tau),\end{aligned}$$
   where $A_p = A + n/2$ and $B_p = B + \frac{1}{2}\sum_i x_i^2$. Hence, the posterior has the InvGa distribution with parameters A_p and B_p.

   (b) There is a constant c (not depending on τ) such that
   $$\log p(\tau \mid x) = c - (A_p + 1) \log \tau - \frac{B_p}{\tau}.$$
   Taking the derivative with respect to τ and equating to zero gives
   $$\hat{\tau} = \frac{B_p}{A_p + 1}.$$
   The second derivative is negative, so this corresponds to a maximum. [Alternatively, it suffices to show that $\lim_{\tau \downarrow 0} \log p(\tau \mid x) = -\infty$ and $\lim_{\tau \to \infty} \log p(\tau \mid x) = -\infty$.]

   (c) The second derivative is
   $$\frac{d^2}{d\tau^2} \log p(\tau \mid x) = \frac{A_p + 1}{\tau^2} - 2\frac{B_p}{\tau^3}.$$
   Plugging in the posterior mode $B_p/(A_p + 1)$ gives
   $$\frac{(A_p + 1)^3}{B_p^2} - 2\frac{(A_p + 1)^3}{B_p^2} = -\frac{(A_p + 1)^3}{B_p^2}.$$

So

$$p(\tau \mid x) \approx\propto \exp\left(-\frac{1}{2}\left(\tau - \frac{B_p}{A_p+1}\right)^2 \frac{(A_p+1)^3}{B_p^2}\right).$$

The Laplace approximation is given by

$$N\left(\frac{B_p}{A_p+1}, \frac{B_p^2}{(A_p+1)^3}\right).$$

(d) Compute $p(x;B) = \int p(x \mid \tau)p(\tau;B)\mathrm{d}\,\tau$. Now maximise this expression with respect to B.

(e) The posterior mode maximises

$$\psi(\tau) = -\frac{n}{2}\log\tau - \frac{s}{2\tau} - \theta\tau.$$

Here $s = \sum_i x_i^2$.

$$\psi'(\tau) = -\frac{n}{2\tau} + \frac{s}{2\tau^2} - \theta$$

$$\psi''(\tau) = \frac{n}{2\tau^2} - \frac{s}{\tau^3}$$

Set $\tau \leftarrow \tau - \psi'(\tau)/\psi''(\tau)$.

3. (a)

$$p(\theta \mid y) \propto p(y \mid \theta)p(\theta) \propto \exp\left(-\frac{1}{2}\|y - X\theta\|^2 - \frac{1}{2\gamma}\|\theta\|^2\right).$$

This is a quadratic form in the exponent, so the posterior has a normal distribution. To find its parameters, note that

$$p(\theta \mid y) \propto \exp\left(-\frac{1}{2}\theta^T(\gamma^{-1}I + X^TX)\theta + \theta^TX^Ty\right).$$

Hence, the cov-matrix of the posterior is $\Sigma = (\gamma^{-1}I + X^TX)^{-1}$ and its mean is

$$\mu = \Sigma X^Ty = (\gamma^{-1}I + X^TX)^{-1}X^Ty = (\gamma^{-1}I + X^TX)^{-1}X^Ty.$$

(b) We want to minimise $f(\theta) := \|y - X\theta\|^2 + \lambda\|\theta\|^2$ with respect to θ. We have

$$\nabla f(\theta) = -2X^T(y - X\theta) + 2\lambda\theta.$$

Hence, equating to zero gives $X^T(y - X\theta) - \lambda\theta = X^Ty - (X^TX + \lambda I)\theta = 0$. We observe that $X^TX + \lambda I$ is positive definite and thus invertible. We derive

$$\hat{\theta} = (X^TX + \lambda I)^{-1}X^Ty.$$

Now the Hessian of f equals $X^TX + \lambda I$, which is positive definite. Hence the stationary point corresponds to a minimum.

(c) They are the same for $\lambda = 1/\gamma$.

4. (a) For $\theta = \theta^\circ$ both sides are equal. So let $\theta \neq \theta^\circ$. Then we have $\bar{q}(\theta, \theta^\circ) = q(\theta, \theta^\circ)\alpha(\theta, \theta^\circ)$. So we calculate

$$\begin{aligned}\pi(\theta)\bar{q}(\theta, \theta^\circ) &= \pi(\theta)q(\theta, \theta^\circ)\min\left(1, \frac{\pi(\theta^\circ)}{\pi(\theta)}\frac{q(\theta^\circ, \theta)}{q(\theta, \theta^\circ)}\right) \\ &= \min\left(\pi(\theta)q(\theta, \theta^\circ), \pi(\theta^\circ)q(\theta^\circ, \theta)\right) = \pi(\theta^\circ)\bar{q}(\theta^\circ, \theta).\end{aligned}$$

   (b) Using (a) we calculate

$$\int_{\mathbb{R}^d} \pi(\theta)\bar{q}(\theta, \theta^\circ)d\theta = \int_{\mathbb{R}^d} \pi(\theta^\circ)\bar{q}(\theta^\circ, \theta)d\theta = \pi(\theta^\circ)\int_{\mathbb{R}^d} \bar{q}(\theta^\circ, \theta)d\theta = \pi(\theta^\circ)$$

   since $\bar{q}(\theta^\circ, \cdot)$ is a probability density.

5. We have

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma\right),$$

   where

$$\Sigma = \begin{bmatrix} K(1,1) & K(1,2) \\ K(2,1) & K(2,2) \end{bmatrix} + \sigma^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 9 \\ 9 & 25 \end{bmatrix} + \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix} = \begin{bmatrix} 9 & 9 \\ 9 & 30 \end{bmatrix}.$$