Technische Universiteit Delft
Faculteit Elektrotechniek, Wiskunde en Informatica
Van Mourik Broekmanweg 6
2628 XE Delft

Exam Multivariate Data Analysis (CS4070)
January 2022

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. Assume $Y_1, \ldots, Y_n$ are independent and $Y_i \sim \text{Exp}(\psi(\theta^T x_i))$. Here $x_1, \ldots, x_n$ are vectors in $\mathbb{R}^p$ of predictor variables and $\theta \in \mathbb{R}^p$ is an unknown parameter vector. Furthermore, $\psi\colon \mathbb{R} \to (0, \infty)$ is fixed and specified.

   (a) [2 pt]. Give expressions for the likelihood $L(\theta)$ and loglikelihood $\ell(\theta)$.

   (b) [3 pt]. Derive an expression for $\frac{\partial \ell(\theta)}{\partial \theta_j}$. Show that if $\psi(x) = e^x$ this expression simplifies to $\sum_{i=1}^n x_{ij}(1-\psi_i y_i)$, where $\psi_i \equiv \psi(x_i^T\theta)$ and x_{ij} denotes the j-th element in the vector x_i. In the remainder we'll assume $\psi(x) = e^x$.

   (c) [2 pt]. Derive an expression for the Hessian matrix $H(\theta)$ of the mapping $\theta \mapsto \ell(\theta)$.

   (d) [1 pt]. Explain one step of Newton's algorithm for optimising the (log)likelihood.

   (e) [2 pt]. Suppose we would take the Bayesian point of view and provide a prior for $\theta \sim \text{N}_p(0, \alpha I_{p\times p})$. How would you have to adjust the answer to the previous question to numerically approximate the posterior mode?

2. Assume $X_1, \ldots, X_p$ are independent conditional on $\Theta_1, \ldots, \Theta_p$. Suppose $X_i \mid \Theta_i = \theta_i \sim Unif(0, \theta_i)$ (the uniform distribution on $(0, \theta_i)$). Consider estimation of the parameters $\Theta_1, \ldots, \Theta_p$ based on data $X_1, \ldots, X_p$.

   (a) [3 pt]. Model $\Theta_1, \ldots, \Theta_p$ as independent with common density

   $$f_\Theta(\theta) = \theta\lambda^2 e^{-\lambda\theta}\mathbf{1}_{[0,\infty)}(\theta).$$

   This implies $\mathbb{E}\Theta_i = 2/\lambda$. Show that $\Theta_1, \ldots, \Theta_p$ are aposteriori independent. Find their posterior distribution and verify that

   $$\mathbb{E}[\Theta_i \mid X_i] = X_i + 1/\lambda.$$

   (b) [2 pt]. Verify that $\mathbb{E}[X_i] = 1/\lambda$.
   Hint; use $\mathbb{E}[X_i] = \mathbb{E}[\mathbb{E}[X_i \mid \Theta_i]]$.

(c) [2 pt]. If we use the posterior mean as estimator, then the performance of the estimator is highly depend on the choice of the hyperparameter λ. The method of empirical Bayes consists of plugging in an estimator for λ, based on

$$f_{X_1,\ldots,X_p}(x_1,\ldots,x_p) = \int f_{X_1,\ldots,X_n|\Theta_1,\ldots,\Theta_p}(x_p,\ldots,x_p \mid \theta_1,\ldots,\theta_p) \\ f_{\Theta_1,\ldots,\Theta_p}(\theta_1,\ldots,\theta_p)\mathrm{d}\,\theta_1,\ldots\mathrm{d}\,\theta_p.$$

Derive an expression for $\lambda \mapsto f_{X_1,\ldots,X_p}(x_1,\ldots,x_p)$. Note that the dependence on λ is suppressed from the notation, but enters via the prior distribution.

(d) [2 pt]. Determine an estimator for λ as the (a) maximiser of

$$\lambda \mapsto f_{X_1,\ldots,X_p}(x_1,\ldots,x_p).$$

(e) [1 pt]. Combine parts (a) and (d) to find empirical Bayes estimators for $\Theta_1,\ldots,\Theta_p$.

3. Consider the model $y \mid z \sim N(Xz, \lambda I_n)$ (where I_n is the $n \times n$ identity matrix and X is an $n \times p$ matrix containing explanatory variables, considered fixed). Assume $z \sim N(0, \alpha I_p)$ and that λ and α are known.

(a) [3 pt]. Derive the posterior for z.

(b) [2 pt]. Derive an expression for the covariance matrix of y.

4. [3 pt]. Suppose $i \in \{1,\ldots,n\}$ and

$$\begin{aligned} y_i \mid \theta_i &\overset{\text{ind}}{\sim} Ber\left(u(\theta_i)\right) \\ \theta_1,\ldots,\theta_n \mid \eta &\overset{\text{iid}}{\sim} N(0,\eta) \\ \eta &\sim Exp(1), \end{aligned}$$

where u is a function that maps $\mathbb{R}$ to $(0,1)$ and Ber denotes the Bernoulli distribution. Suppose we want to use the Metropolis-Hastings algorithm to draw from the posterior of $(\theta_1,\ldots,\theta_n,\eta)$. Suppose we use the Gibbs sampler and iteratively update $(\theta_1,\ldots,\theta_n)$ and η. Give the details necessary for implementing the update step for η. This includes specifying a proposal distribution for the Metropolis-Hastings algorithm.

5. [2 pt]. Suppose $f \sim \mathrm{GP}(0,K)$, so f is a Gaussian process with mean function 0 and covariance kernel K, where $K(x,y) = e^{-|x-y|}$, $(x,y \in \mathbb{R})$. Give the (joint)-distribution of

$$\begin{bmatrix} f(1) \\ f(3) \end{bmatrix}$$

6. [2 pt]. Why is Gaussian process regression computationally expensive for large datasets? Explain the computationally most expensive step.

Answers

1. (a) If $\psi_i := \psi(\theta^T x_i)$, then

$$L(\theta) = \prod_i \psi_i e^{-\psi_i y_i}$$

Hence

$$\ell(\theta) = \sum_i \left(\log \psi_i - \psi_i y_i\right).$$

(b)

$$\frac{\partial \ell(\theta)}{\partial \theta_j} = \sum_i \frac{\psi'(\theta^T x_i) x_{ij}}{\psi(\theta^T x_i)} - \psi'(\theta^T x_i) x_{ij} y_i.$$

If $\psi(x) = e^x$ then this simplifies to

$$\frac{\partial \ell(\theta)}{\partial \theta_j} = \sum_i x_{ij} - \psi_i x_{ij} y_i.$$

(c) It follows that

$$\frac{\partial^2 \ell(\theta)}{\partial \theta_j \partial \theta_k} = \sum_i -x_{ij} y_i x_{ik} \psi_i.$$

Therefore,

$$H(\theta) = -X^T \mathrm{diag}(y_i \psi_i) X$$

(d)

$$\theta := \theta - H(\theta)^{-1} \nabla \ell(\theta).$$

(e) To the gradient and Hessian one should add

$$\nabla \left(-\frac{1}{2\alpha} \|\theta\|^2 \right) = -\alpha^{-1} \theta$$

and

$$-\alpha^{-1} I,$$

respectively.

2. (a) Note that the $\Theta \sim Ga(2, \lambda)$. Further, it is easily seen that $\Theta_1, \ldots, \Theta_p$ are aposteriori independent and that the posterior distribution for Θ_i only depends on X_i. To find that distribution, we drop the index i from the notation and check that.

$$\begin{aligned} p(\theta \mid x) \propto p(x \mid \theta) p(\theta) &= \frac{1}{\theta} \mathbf{1}_{(0,\theta)}(x) \theta \lambda^2 e^{-\lambda\theta} \mathbf{1}_{(0,\infty)}(\theta) \\ &= \lambda^2 e^{-\lambda\theta} \mathbf{1}_{(x,\infty)}(\theta) \propto e^{-\lambda\theta} \mathbf{1}_{(x,\infty)}(\theta). \end{aligned}$$

Therefore,

$$p(\theta \mid x) = \frac{e^{-\lambda\theta} \mathbf{1}_{(x,\infty)}(\theta)}{\int_0^\infty e^{-\lambda\theta} \mathbf{1}_{(x,\infty)}(\theta) d\theta} = \lambda e^{-\lambda(\theta - x)} \mathbf{1}_{(x,\infty)}(\theta).$$

So each Θ_i is distributed as $X_i + Z_i$, where $\{Z_i,\, 1 \le i \le p\}$ is a sequence of IID Exp(λ)-distributed random variables, independent of all X_i. The posterior mean henceforth equals

$$\mathbb{E}[\Theta_i \mid X_i] = \frac{1}{\lambda} + X_i.$$

(b). Use the law of repeated expectation:

$$\mathbb{E}[X_i] = \mathbb{E}\mathbb{E}[X_i \mid \Theta_i] = \mathbb{E}[\Theta_i/2] = \frac{1}{2}\frac{2}{\lambda} = \frac{1}{\lambda}.$$

(b) Note that

$$p(x_1, \ldots, x_p; \lambda) = \int \prod_{i=1}^{p} p(x_i \mid \theta_i) p(\theta_i; \lambda) \mathrm{d}\, x_i = \prod_{i=1}^{p} \int p(x_i \mid \theta_i) p(\theta_i; \lambda) \mathrm{d}\, x_i.$$

(c) Verify that $p(x_1, \ldots, x_p; \lambda) = \lambda^p e^{-\lambda \sum_{i=1}^{p} x_i}$.

(d) Take the log, differentiate wrt λ, equate to zero and then check that $\hat{\lambda} = 1/\bar{x}_p$.

(e) Combining (a) and (b) gives $\hat{\theta}_i^{EB} = \bar{X}_p + X_i$ (just plug-in the emp.Bayes estimator for λ).

3. (a)

$$p(z \mid y) \propto p(y \mid z) p(z) \propto \exp\left(-\frac{1}{2\lambda} \|y - Xz\|^2 - \frac{1}{2\alpha} \|z\|^2 \right).$$

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

$$p(z \mid y) \propto \exp\left(-\frac{1}{2} z^T (\alpha^{-1} I + \lambda^{-1} X^T X) z + \lambda^{-1} z^T X^T y \right).$$

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

$$\mu = \Sigma \lambda^{-1} X^T y = (\alpha^{-1} I + \lambda^{-1} X^T X)^{-1} \lambda^{-1} X^T y = (\lambda \alpha^{-1} I + X^T X)^{-1} X^T y.$$

(b)

$$\begin{aligned} \operatorname{Cov} y &= \operatorname{Cov} \mathrm{E}(y \mid z) + \mathrm{E} \operatorname{Cov}(y \mid z) \\ &= \operatorname{Cov} Xz + \mathrm{E}\, \lambda I_n = X \operatorname{Cov} z X^T + \lambda I_n = \alpha X X^T + \lambda I_n \end{aligned}$$

4. Note that the update step for η only depends on $\theta \equiv (\theta_1, \ldots, \theta_n)$. So

$$p(\eta \mid \theta) \propto p(\eta) \prod_i p(\theta_i \mid \eta) \propto e^{-\eta} \eta^{-n/2} \exp\left(-\frac{1}{2\eta} \sum_i \theta_i^2 \right) \mathbf{1}_{(0,\infty)}(\eta).$$

So if $q(\eta, \eta^\circ)$ specifies a proposal density, then η° is accepted with probability $1 \wedge A$ where

$$A = \frac{p(\eta^\circ \mid \theta)}{p(\eta \mid \theta)} \frac{q(\eta^\circ, \eta)}{q(\eta, \eta^\circ)}.$$

One can for example use random walk proposals where

$$\eta^\circ := \eta + hZ$$

and $Z \sim N(0,1)$. Note however that then we propose many values for η that may be negative that get rejected. Better would be (not necessary for full points)

$$\log \eta^\circ := \log \eta + hZ$$

5.

$$\begin{bmatrix} f(1) \\ f(3) \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K(1,1) & K(1,3) \\ K(3,1) & K(3,3) \end{bmatrix} \right)$$

Now $K(1,1) = K(3,3) = 1$ and $K(1,3) = K(3,1) = e^{-2}$.

6. One needs to invert the $n \times n$ matrix with elements $K(x_i, x_j)$.