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

Exam Multivariate Data Analysis (CS4070)
23 January 2024, 13:30-16:30

Probability distributions formulas:

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

Write $Z \sim N(\mu, \Sigma)$ if Z has the multivariate normal distribution with mean vector $\mu \in \mathbb{R}^d$ and $d \times d$ covariance matrix Σ. That is, its density is given by

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

Start of questions:

1. Consider the linear model $y = X\beta + \varepsilon$ with $\varepsilon \sim N(0, \sigma^2 I_n)$ (where $\beta \in \mathbb{R}^p$ is unknown, $\sigma^2 > 0$ is known, I_n is the $n \times n$ identity matrix and X is an $n \times p$ matrix containing explanatory variables, considered fixed and known).

   (a) [1 pt]. Write down the least squares criterion for estimating β.

   (b) [2 pt]. Write down the likelihood for β and give the definition of the maximum likelihood estimator (MLE) of β in terms of a maximisation problem.

   (c) [1 pt]. Relate the MLE to the least squares estimator.

   (d) [3 pt]. Assume the prior distribution $\beta \sim N(0, \gamma I_p)$ and that $\gamma > 0$ is known. Derive the posterior for β.

2. Suppose $i \in \{1, \ldots, n\}$ and

$$\begin{aligned} y_i \mid \theta &\overset{\text{iid}}{\sim} Unif(0, \theta) \\ \theta \mid \lambda &\sim Par(\lambda) \\ \lambda &\sim Ga(\alpha, \beta), \end{aligned}$$

   where $\alpha, \beta > 0$ are known hyperparameters and $Unif$, Par and Ga denote the uniform, Pareto and Gamma distribution, respectively. The density of $Par(\lambda)$ is given by $p(\theta) = \frac{\lambda}{\theta^{\lambda+1}} \mathbf{1}_{[1,\infty)}(\theta)$ and for the density of $Ga(\alpha, \beta)$ we have $p(\lambda) \propto \lambda^{\alpha-1} e^{-\beta\lambda} \mathbf{1}_{(0,\infty)}(\lambda)$.

   (a) [1 pt]. Given you can sample from the required conditional distributions, how does Gibbs sampling work in this model?

   (b) [2 pt]. Derive the conditional distribution of $\lambda \mid \theta, y_1, \ldots, y_n$.

   (c) [1 pt]. Based on (b), how is such a prior on λ called?

3. [2 pt]. Suppose that for each $i \in \{1, 2\}$, the Markov transition $q_i(x, y)$, $x, y \in \mathbb{R}^d$, is invariant for the density p. That is, $p(y) = \int_{\mathbb{R}^d} p(x) q_i(x, y) \, \mathrm{d}\, x$.

   Show that the Markov transition $q(x, y) := \int_{\mathbb{R}^d} q_1(x, z) q_2(z, y) \, \mathrm{d}\, z$ is invariant for p.

   Hint: You may interchange the order of integration without justification.

4. Assume $Y_1, \ldots, Y_n$ are independent and follow a Binomial distribution $Y_i \sim \mathrm{B}(m, p_i)$ for a fixed positive integer m, so that $P(Y_i = k) = \binom{m}{k} p_i^k (1-p_i)^{m-k}$ with $\binom{m}{k} = \frac{m!}{k!(m-k)!}$ for $k = 0, \ldots, m$. Here $p_i = \psi(\theta^T x_i)$ for vectors $x_1, \ldots, x_n$ in $\mathbb{R}^p$ of predictor variables and $\theta \in \mathbb{R}^p$ is an unknown parameter vector. Furthermore, $\psi \colon \mathbb{R} \to [0, 1]$ is fixed and specified.

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

   Assume for the remaining part of the question that $\psi(z) = 1/(1 + e^{-z})$.

   (b) [1 pt]. Show that $\psi'(z) = \psi(z)(1 - \psi(z))$.

   (c) [3 pt]. Derive an expression for $\frac{\partial \ell(\theta)}{\partial \theta_j}$. Show that this expression simplifies to $\sum_{i=1}^n x_{ij}(y_i - mp_i)$, where x_{ij} denotes the j-th element in the vector x_i.

   (d) [2 pt]. Derive an expression for the elements $\frac{\partial^2 \ell(\theta)}{\partial \theta_k \partial \theta_j}$ of the Hessian matrix $H(\theta)$ in terms of m, p_i and the elements of the vectors x_i.

   (e) [1 pt]. Give one step of Newton's algorithm for optimising the loglikelihood.

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

5. [3 pt]. Suppose

$$
\begin{aligned}
x_1, \ldots, x_n \mid \lambda &\overset{\text{iid}}{\sim} N(0, \lambda) \\
\lambda &\sim Exp(1).
\end{aligned}
$$

   Suppose we want to use the Metropolis–Hastings algorithm to draw from the posterior of λ. Give the details necessary for implementing the update step for λ. Also specify a proposal distribution for the Metropolis–Hastings algorithm.

6. Consider Gaussian process regression, where η denotes the parameter vector of the kernel K and σ^2 the variance of the noise.

   (a) [2 pt]. What is the computationally most expensive step in Gaussian process regression? What is the order of the computational complexity in terms of the number of observations n?

   (b) [1 pt]. How can (σ^2, η) be determined by the empirical Bayes method?

   (c) [1 pt]. Name one algorithm that can be used to compute the empirical Bayes choice for (σ^2, η).

Answers

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

   (b) $L(\beta) = (2\pi)^{-n/2}\sigma^{-n}\exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|^2\right)$ and $\hat{\beta}_{\text{MLE}} = \text{argmax}_\beta L(\beta)$

   (c) We have

$$\begin{aligned}\underset{\beta}{\text{argmax}}\, L(\beta) &= \underset{\beta}{\text{argmax}}\exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|^2\right)\\ &= \underset{\beta}{\text{argmin}}\,\|y - X\beta\|^2 = \underset{\beta}{\text{argmin}}\, S(\beta).\end{aligned}$$

   They are the same.

   (d)

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

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

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

   Hence, the cov-matrix of the posterior is $\Sigma = (\gamma^{-1}I + \sigma^{-2}X^TX)^{-1}$ and its mean is

$$\mu = \Sigma\sigma^{-2}X^Ty = (\gamma^{-1}I + \sigma^{-2}X^TX)^{-1}\sigma^{-2}X^Ty = (\sigma^2\gamma^{-1}I + X^TX)^{-1}X^Ty.$$

2. (a) Iteratively sample from the full conditionals of $\theta \mid \lambda, y_1, \ldots, y_n$ and $\lambda \mid \theta, y_1, \ldots, y_n$.

   (b) We have

$$\begin{aligned}p(\lambda \mid \theta, y_1, \ldots, y_n) &\propto p(y_1, \ldots, y_n \mid \theta)p(\theta \mid \lambda)p(\lambda) \propto p(\theta \mid \lambda)p(\lambda)\\ &\propto \frac{\lambda}{\theta^{\lambda+1}}\mathbf{1}_{[1,\infty)}(\theta)\lambda^{\alpha-1}e^{-\beta\lambda}\mathbf{1}_{(0,\infty)}(\lambda)\\ &\propto \lambda^{(\alpha+1)-1}e^{-(\beta+\log(\theta))\lambda}\mathbf{1}_{(0,\infty)}(\lambda)\\ &\propto Ga\left(\alpha+1, \beta+\log(\theta)\right).\end{aligned}$$

   (c) Such a prior is called partially conjugate.

3.

$$\begin{aligned}\int_{\mathbb{R}^d} p(x)q(x,y)\,\mathrm{d}\,x &= \int_{\mathbb{R}^d} p(x)\int_{\mathbb{R}^d} q_1(x,z)q_2(z,y)\,\mathrm{d}\,z\,\mathrm{d}\,x\\ &= \int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d} p(x)q_1(x,z)\,\mathrm{d}\,x\right)q_2(z,y)\,\mathrm{d}\,z\\ &= \int_{\mathbb{R}^d} p(z)q_2(z,y)\,\mathrm{d}\,z = p(y)\end{aligned}$$

4. (a)

$$L(\theta) = \prod_i \binom{m}{y_i} (\psi(\theta^T x_i))^{y_i} (1 - \psi(\theta^T x_i))^{m-y_i}$$

and

$$\ell(\theta) = \sum_i \log\left(\binom{m}{y_i}\right) + \sum_i y_i \log(\psi(\theta^T x_i)) + \sum_i (m - y_i)\log(1 - \psi(\theta^T x_i))$$

(b)

$$\psi'(z) = -\frac{1}{(1+e^{-z})^2}(-e^{-z}) = \frac{e^{-z}}{(1+e^{-z})^2} = \psi(z)(1-\psi(z))$$

(c)

$$\begin{aligned}
\frac{\partial \ell(\theta)}{\partial \theta_j} &= \sum_i \frac{y_i}{\psi(\theta^T x_i)} \psi'(\theta^T x_i) x_{ij} + \sum_i \frac{m - y_i}{1 - \psi(\theta^T x_i)} (-\psi'(\theta^T x_i)) x_{ij} \\
&= \sum_i y_i (1 - \psi(\theta^T x_i)) x_{ij} - \sum_i (m - y_i) \psi(\theta^T x_i) x_{ij} \\
&= \sum_i y_i x_{ij} - \sum_i m \psi(\theta^T x_i) x_{ij} \\
&= \sum_i (y_i - m\psi(\theta^T x_i)) x_{ij} = \sum_i x_{ij} (y_i - m p_i)
\end{aligned}$$

(d) It follows that

$$\begin{aligned}
(H(\theta))_{kj} = \frac{\partial^2 \ell(\theta)}{\partial \theta_k \partial \theta_j} &= \frac{\partial}{\partial \theta_k} \sum_i (y_i - m\psi(\theta^T x_i)) x_{ij} \\
&= -m \sum_i \psi'(\theta^T x_i) x_{ik} x_{ij} \\
&= -m \sum_i p_i (1 - p_i) x_{ik} x_{ij}.
\end{aligned}$$

(e)

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

(f) 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.

5. Denoting $x \equiv (x_1, \ldots, x_n)$ we have

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

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

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

One can for example use random walk proposals, where

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

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

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

6. (a) The computationally most expensive step in Gaussian process regression is the inversion of the $n \times n$ matrix $\mathcal{K}$ or $\mathcal{K} + \sigma^2$. The order of the computational complexity in terms of the number of observations n is $\mathcal{O}(n^3)$.
   (b) Find the parameters (σ^2, η) for which the marginal likelihood $p(y; \sigma^2, \eta)$ or the marginal loglikelihood $\log(p(y; \sigma^2, \eta))$ is maximal.
   (c) Gradient methods can be used to optimise with respect to (σ^2, η), for example, gradient descent, stochastic gradient descent or the Newton method.