STATISTICAL MODELLING Part IIC, Michaelmas 2024

Practical 5: Exponential Families

This practical we will examine the finite sample performance of confidence intervals for the MLE of parameters of samples from exponential families, in particular the family of gamma distributions. Recall that Exercise 3.6 in the course notes asks you to to show this and to find this family's natural parameter(s) and cumulant function (see also Example Sheet 3).

Exercises

1. Consider sampling independent $X_i \sim \Gamma(\alpha, \beta)$, which each have the density function:

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) \tag{1}$$

   (with a slight abuse of notation, here the Γ means the gamma function). In the next few exercises we will hold the shape parameter fixed at $\alpha = 2$. By showing that this one parameter family of distributions $\{\Gamma(2, \beta)\}_{\beta \in (0,\infty)}$ is an exponential family or otherwise, write a function that fits the MLE of the rate parameter β from a sample $\{X_i\}_{i=1}^n$.

2. Using the R function `rgamma`, sample a $n \times B$ matrix of $X_{i,j} \overset{iid}{\sim} \Gamma(2, 2)$ for $n = 30$ and $B = 1000$. Using the R function `apply`, find the MLEs for the rate parameter given the samples $\{X_{i,j}\}_{i=1}^n$. Plot these parameter estimates as a histogram.

3. Recall that in class we showed that there is an approximation for the distribution of the MLE in an exponential family given by:

$$\hat{\theta} \dot{\sim} N(\theta, (i^{(n)}(\theta))^{-1}) \tag{2}$$

   where $i^{(n)}(\theta)$ is the information matrix for the n samples under the true distribution with natural parameter θ. Since we do not know $i^{(n)}(\theta)$, to obtain confidence intervals we must replace this covariance matrix with the inverse of the observed information matrix, where θ is replaced by the MLE $\hat{\theta}$.

   Using this approximate distribution, obtain a 95% confidence interval for the rate parameters β for each of the datasets $\{X_{i,j}\}_{i=1}^{30}$. Calculate the proportion of the 1000 simulations for which these intervals contain the true parameter.

4. Find the exact distribution of the MLE $\hat{\beta}$. Compare your MLEs with the theoretical distribution by constructing a Q-Q plot as in Practical 2. Also compare the empirical sample of MLEs with the asymptotic distribution in equation (2). Repeat this but for $n = 300$ or more instead.

5. Now consider the joint estimation of (α, β) when they are both unknown. Write a function that estimates the MLE for these parameters using the R function `optim`. Read the documentation `?optim` for guidance on how to do this. Apply this to your simulations from earlier and plot these as pairs of points $(\hat{\alpha}, \hat{\beta})$.

6. The `optim` function has an option to return the observed information matrix by setting `Hessian = TRUE`. Use this to write a function that gives a joint confidence ellipse for the parameters and test their empirical size as in question 3. *Hint: the R function `eigen` may be helpful for this.*