MATH562 – Fall 2025

Problem Set: Week 9

1. Data $Y_1, \ldots, Y_n$ treated as a random sample from the geometric density f with support $\mathcal{Y} = \{0, 1, \ldots\}$ and parameter $\theta \in (0, 1)$ are in fact from the Poisson density g with mean $\lambda > 0$.

   (a) Show that $\mathrm{E}_g\{\log f(Y; \theta)\}$ is maximised by $\theta_g = (1 + \lambda)^{-1}$, and check that this matches the means of the models. Is this a general feature of misspecified exponential family models? **Solution**: The geometric density on $\mathcal{Y}$ is $\log f(Y; \theta) = Y \log(1 - \theta) + \log \theta$, and as $\mathrm{E}_g(Y) = \lambda$ we have $\mathrm{E}_g\{\log f(Y; \theta)\} = \lambda \log(1 - \theta) + \log \theta$. Differentiation of this with respect to θ yields $-\lambda/(1 - \theta) + 1/\theta$, and setting this equal to zero yields $\theta_g = (1 + \lambda)^{-1}$; note that the second derivative is negative for all θ. The mean of the geometric distribution with support $\mathcal{Y}$ is $1/\theta_g - 1 = \lambda$, so the misspecified geometric model matches the mean of g.

   The log likelihood for an exponential family is $s(y)^\mathsf{T}\varphi - k(\varphi) + \log m(y)$, and its expected value under another exponential family model would be

   $$\mathrm{E}_g\{s(Y)\}^\mathsf{T}\varphi - k(\varphi) + \mathrm{E}_g\{\log m(Y)\},$$

   which is to be maximised with respect to φ. This implies that $\mathrm{E}_g\{s(Y)\} = \nabla k(\varphi)$, so φ is chosen so that the mean $\nabla k(\varphi)$ under the candidate model equals the mean of $s(Y)$ under the true model. So this is a general property.

   (b) Compute $\imath_1(\theta_g) = \mathrm{Var}_g\{\frac{\partial}{\partial \theta_g} \log f(Y; \theta_g)\}$ and $\jmath_1(\theta_g) = -\mathrm{E}_g\{\frac{\partial^2}{\partial \theta_g^2} \log f(Y; \theta_g)\}$ and use them to determine the asymptotic variance $\mathrm{Var}(\hat{\theta}_g) \doteq \theta_g^3(1 - \theta_g)/n$. **Solution**: The log likelihood derivatives for a single observation are $\theta^{-1} - Y/(1 - \theta)$ and $-\{1/\theta^2 + Y/(1 - \theta)^2\}$, and these give $\imath_1(\theta_g) = 1/\{\theta_g(1 - \theta_g)\}$ and $\jmath_1(\theta_g) = 1/\{\theta_g^2(1 - \theta_g)\}$, so the "sandwich variance" – as given by the theorem on the asymptotic distribution of the MLE – for a sample of size n is $\imath_1(\theta_g)/\{n\jmath_1(\theta_g)^2\} = \theta_g^3(1 - \theta_g)/n$, or $\lambda/\{(1 + \lambda)^4 n\}$.

   (c) Show that the maximum likelihood estimator of θ based on $Y_1, \ldots, Y_n$ is $\hat{\theta} = 1/(1 + \bar{Y})$ and use the delta method to find its asymptotic variance. Is this a surprise? **Solution**: The log likelihood is $\ell(\theta) = n\bar{Y} \log(1 - \theta) + n \log \theta$, which gives $\hat{\theta} = h(\bar{Y})$, where $h(x) = 1/(1 + x)$ for $x > 0$. The delta method variance is $h'\{\mathrm{E}(Y)\}^2 \mathrm{Var}_g(\bar{Y}) = \{-1/(1 + \lambda)^2\}^2 \lambda/n = \theta_g^4(1 - \theta_g/(\theta_g n) = \theta_g^3(1 - \theta_g)/n$.

   This is not really a surprise, but it is reassuring that the sandwich formula gives the natural variance computed directly by applying the delta method to the formula for the MLE.

2. When the generalized Pareto distribution is written as

   $$\Pr(Y > y) = (1 - y/\psi)_+^\lambda, \quad 0 < y < \psi, \quad \psi, \lambda > 0,$$

   where $a_+ = \max(a, 0)$, the parameter ψ represents the upper support point for Y.

   (a) Find the profile log likelihood for ψ based on a random sample $Y_1, \ldots, Y_n$. **Solution**: The density function is $(\lambda/\psi)(1 - y/\psi)_+^{\lambda - 1}$, so the log likelihood for a random sample can be written as

   $$\ell(\psi, \lambda) = n \log \lambda - n \log \psi + (\lambda - 1) \sum_{j=1}^{n} \log(1 - y_j/\psi)_+,$$

   and this implies that $\hat{\lambda}_\psi = \arg\max_\lambda \ell(\psi, \lambda) = n/s_\psi$. Hence the profile log likelihood is

   $$\ell_p(\psi) \equiv s_\psi - n \log \psi - n \log s_\psi, \quad \psi > \max(y_1, \ldots, y_n),$$

   because clearly the upper bound to the support of the density, ψ, exceeds all the y_j.

(b) Show that if ψ is regarded as fixed, then the minimal sufficient statistic for λ is $S_\psi = \sum_j Z_j$, where $Z_j = -\log(1 - Y_j/\psi)$. By considering $\Pr(Z_j > z)$ or otherwise, show that S_ψ has a gamma distribution and deduce that the conditional density of the data given $S_\psi = s_\psi$ is

$$f(y_1, \ldots, y_n \mid s_\psi; \psi) = \frac{\Gamma(n) e^{s_\psi}}{\psi^n s_\psi^{n-1}}, \quad 0 < y_1, \ldots, y_n < \psi, \sum_{j=1}^{n} \log(1 - y_j/\psi) = s_\psi.$$

Solution: We write the joint density as

$$f(y_1, \ldots, y_n; \psi, \lambda) = (\lambda/\psi)^n \exp\{-(\lambda - 1)s_\psi\}, \quad \lambda > 0,$$

and note that if ψ is fixed then this is a (1,1)-exponential family with canonical parameter $-\lambda$ and canonical statistic s_ψ, which must therefore be minimal sufficient for λ. Now

$$\Pr(Z > z) = \Pr\{-\log(1 - Y/\psi) > z\} = \Pr(1 - Y/\psi < e^{-z}) = \Pr\{Y > \psi(1 - e^{-z})\} = e^{-\lambda z}, \quad z > 0,$$

so $S_\psi = \sum_{j=1}^{n} Z_j$ is the sum of n independent exponential variables and therefore has a gamma (n, λ) distribution, and density

$$f_{S_\psi}(s_\psi; \psi, \lambda) = \frac{\lambda^n s_\psi^{n-1}}{\Gamma(n)} e^{-\lambda s_\psi}, \quad s_\psi > 0.$$

Thus the conditional density of $Y_1, \ldots, Y_n$ given $S_\psi = s_\psi$ is of the stated form.

(c) Compare the profile log likelihood with the log likelihood obtained from (b). Which is preferable? **Solution**: The conditional log likelihood from (b) is

$$\ell_c(\psi) = \log f(y \mid s_\psi; \psi) \equiv s_\psi - n \log \psi - (n - 1) \log s_\psi, \quad \psi > \max(y_1, \ldots, y_n),$$

so the only difference is the replacement of n by $n - 1$ in the last term. The latter is preferable, because it is based on a true density, so it should have slightly better behaviour in small samples.

*3. The generalized Pareto distribution was given in an earlier question. $\lambda > 0$.

(a) Show that the derivatives with respect to ψ satisfy the first two Bartlett identities only if $\lambda > 2$.
Hint: Besides appearing in the MLE theory, the Bartlett identities were explicitly listed in the first set of slides. Use that $\int f(y; \psi)\, dy = 1$, for a density f independent of the parameter ψ, and the Leibnitz integration rule.

(b) If M_n denotes the sample maximum, show that $n^{1/\lambda}(\psi - M_n) \xrightarrow{d} W$ as $n \to \infty$, where $\Pr(W > w) = \exp\{-(w/\psi)^\lambda\}$ for $w > 0$. Deduce that when $\lambda \leq 2$ convergence to a limiting distribution for inference on ψ occurs more rapidly than with maximum likelihood estimation.

4. Suppose that the parameter θ consists of a $p \times 1$ parameter of interest ψ and a $q \times 1$ nuisance parameter λ, and that the maximum likelihood estimator $\hat{\theta}$ has approximate distribution

$$\hat{\theta} = \begin{pmatrix} \hat{\psi} \\ \hat{\lambda} \end{pmatrix} \sim \mathbb{N}_{p+q} \left\{ \begin{pmatrix} \psi \\ \lambda \end{pmatrix}, \begin{pmatrix} \hat{\jmath}_{\psi\psi} & \hat{\jmath}_{\psi\lambda} \\ \hat{\jmath}_{\lambda\psi} & \hat{\jmath}_{\lambda\lambda} \end{pmatrix}^{-1} \right\},$$

where the circumflex denotes a quantity evaluated at the overall maximum likelihood estimate.

(a) Use the formula for the inverse of a partitioned matrix to show that $\mathrm{Var}(\hat{\psi}) \doteq (\hat{\jmath}_{\psi\psi} - \hat{\jmath}_{\psi\lambda}\hat{\jmath}_{\lambda\lambda}^{-1}\hat{\jmath}_{\lambda\psi})^{-1}$.

Solution: If the inverses exist and we write

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad A^{-1} = \begin{pmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{pmatrix},$$

then

$$\begin{aligned} A^{11} &= (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}, & A^{12} &= -A_{11}^{-1}A_{12}A^{22}, \\ A^{21} &= -A_{22}^{-1}A_{21}A^{11}, & A^{22} &= (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}, \end{aligned}$$

from which the given formula follows at once.

(b) Show that the profile log likelihood $\ell_p(\psi) = \ell(\psi, \hat{\lambda}_\psi)$ satisfies

$$\tilde{\jmath}_{\mathrm{p}} = -\frac{\partial^2 \ell_p(\psi)}{\partial\psi\partial\psi^{\mathsf{T}}} = \tilde{\jmath}_{\psi\psi} - \tilde{\jmath}_{\psi\lambda}\tilde{\jmath}_{\lambda\lambda}^{-1}\tilde{\jmath}_{\lambda\psi},$$

where a tilde denotes a quantity evaluated at $\hat{\theta}_\psi = (\psi, \hat{\lambda}_\psi)$. Deduce that $\hat{\psi} \sim \mathbb{N}_p(\psi, \hat{\jmath}_{\mathrm{p}}^{-1})$. **Solution**: Differentiation gives

$$\frac{\partial \ell_p(\psi)}{\partial\psi} = \ell_\psi(\psi, \hat{\lambda}_\psi) + \frac{\partial \hat{\lambda}_\psi^{\mathsf{T}}}{\partial\psi}\ell_\lambda(\psi, \hat{\lambda}_\psi) = \ell_\psi(\psi, \hat{\lambda}_\psi)$$

because $\hat{\lambda}_\psi$ maximises $\ell(\psi, \lambda)$ in the λ-direction and thus $\ell_\lambda(\psi, \hat{\lambda}_\psi) = 0$. Hence

$$\frac{\partial^2 \ell_p(\psi)}{\partial\psi\partial\psi^{\mathsf{T}}} = \ell_{\psi\psi}(\psi, \hat{\lambda}_\psi) + \ell_{\psi\lambda}(\psi, \hat{\lambda}_\psi)\frac{\partial \hat{\lambda}_\psi}{\partial\psi^{\mathsf{T}}}.$$

The final expression here is obtained by differentiation of the equation $\ell_\lambda(\psi, \hat{\lambda}_\psi) = 0$, giving

$$0 = \ell_{\lambda\psi}(\psi, \hat{\lambda}_\psi) + \ell_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)\frac{\partial \hat{\lambda}_\psi}{\partial\psi^{\mathsf{T}}},$$

and thus $\partial\hat{\lambda}_\psi/\partial\psi^{\mathsf{T}} = -\ell_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)^{-1}\ell_{\lambda\psi}(\psi, \hat{\lambda}_\psi)$, resulting in

$$\frac{\partial^2 \ell_p(\psi)}{\partial\psi\partial\psi^{\mathsf{T}}} = \ell_{\psi\psi}(\psi, \hat{\lambda}_\psi) - \ell_{\psi\lambda}(\psi, \hat{\lambda}_\psi)\ell_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)^{-1}\ell_{\lambda\psi}(\psi, \hat{\lambda}_\psi).$$

On multiplying by -1 and using the notation $-\ell_{\psi\psi}(\psi, \hat{\lambda}_\psi) = \tilde{\jmath}_{\psi\psi}$, etc., we obtain the expression for $\tilde{\jmath}_{\mathrm{p}}$, which becomes $\hat{\jmath}_{\mathrm{p}} = \hat{\jmath}_{\psi\psi} - \hat{\jmath}_{\psi\lambda}\hat{\jmath}_{\lambda\lambda}^{-1}\hat{\jmath}_{\lambda\psi}$ when $\psi = \hat{\psi}$, giving the required distributional result for $\hat{\psi}$.