Paper 4, Section II

28J Principles of Statistics

We consider a statistical model $\{f(\cdot, \theta) : \theta \in \Theta\}$.

(a) Define the maximum likelihood estimator (MLE) and the Fisher information $I(\theta)$.

(b) Let $\Theta = \mathbb{R}$ and assume there exist a continuous one-to-one function $\mu : \mathbb{R} \to \mathbb{R}$ and a real-valued function h such that

$$\mathbb{E}_\theta[h(X)] = \mu(\theta) \qquad \forall \theta \in \mathbb{R}.$$

(i) For $X_1, \ldots, X_n$ i.i.d. from the model for some $\theta_0 \in \mathbb{R}$, give the limit in almost sure sense of

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} h(X_i)\,.$$

Give a consistent estimator $\hat{\theta}_n$ of θ_0 in terms of $\hat{\mu}_n$.

(ii) Assume further that $\mathbb{E}_{\theta_0}[h(X)^2] < \infty$ and that μ is continuously differentiable and strictly monotone. What is the limit in distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$? Assume too that the statistical model satisfies the usual regularity assumptions. Do you necessarily expect $\mathrm{Var}(\hat{\theta}_n) \geqslant (nI(\theta_0))^{-1}$ for all n? Why?

(iii) Propose an alternative estimator for θ_0 with smaller bias than $\hat{\theta}_n$ if $B_n(\theta_0) = \mathbb{E}_{\theta_0}[\hat{\theta}_n] - \theta_0 = \frac{a}{n} + \frac{b}{n^2} + O(\frac{1}{n^3})$ for some $a, b \in \mathbb{R}$ with $a \neq 0$.

(iv) Further to all the assumptions in iii), assume that the MLE for θ_0 is of the form

$$\hat{\theta}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} h(X_i).$$

What is the link between the Fisher information at θ_0 and the variance of $h(X)$? What does this mean in terms of the precision of the estimator and why?

[*You may use results from the course, provided you state them clearly.*]

Paper 3, Section II

28J Principles of Statistics

We consider the exponential model $\{f(\cdot, \theta) : \theta \in (0, \infty)\}$, where

$$f(x, \theta) = \theta e^{-\theta x} \quad \text{for } x \geqslant 0 .$$

We observe an i.i.d. sample $X_1, \ldots, X_n$ from the model.

(a) Compute the maximum likelihood estimator $\hat{\theta}_{MLE}$ for θ. What is the limit in distribution of $\sqrt{n}(\hat{\theta}_{MLE} - \theta)$?

(b) Consider the Bayesian setting and place a Gamma$(\alpha, \beta), \alpha, \beta > 0$, prior for θ with density

$$\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha - 1} \exp(-\beta\theta) \quad \text{for } \theta > 0 ,$$

where Γ is the Gamma function satisfying $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$ for all $\alpha > 0$. What is the posterior distribution for θ? What is the Bayes estimator $\hat{\theta}_\pi$ for the squared loss?

(c) Show that the Bayes estimator is consistent. What is the limiting distribution of $\sqrt{n}(\hat{\theta}_\pi - \theta)$?

[*You may use results from the course, provided you state them clearly.*]

Paper 2, Section II

28J Principles of Statistics

(a) We consider the model $\{Poisson(\theta) : \theta \in (0,\infty)\}$ and an i.i.d. sample $X_1, \ldots, X_n$ from it. Compute the expectation and variance of X_1 and check they are equal. Find the maximum likelihood estimator $\hat{\theta}_{MLE}$ for θ and, using its form, derive the limit in distribution of $\sqrt{n}(\hat{\theta}_{MLE} - \theta)$.

(b) In practice, Poisson-looking data show overdispersion, i.e., the sample variance is larger than the sample expectation. For $\pi \in [0,1]$ and $\lambda \in (0,\infty)$, let $p_{\pi,\lambda} : \mathbb{N}_0 \to [0,1]$,

$$k \mapsto p_{\pi,\lambda}(k) = \begin{cases} \pi e^{-\lambda} \frac{\lambda^k}{k!} & \text{for } k \geqslant 1 \\ \\ (1-\pi) + \pi e^{-\lambda} & \text{for } k = 0. \end{cases}$$

Show that this defines a distribution. Does it model overdispersion? Justify your answer.

(c) Let $Y_1, \ldots, Y_n$ be an i.i.d. sample from $p_{\pi,\lambda}$. Assume λ is known. Find the maximum likelihood estimator $\hat{\pi}_{MLE}$ for π.

(d) Furthermore, assume that, for any $\pi \in [0,1]$, $\sqrt{n}(\hat{\pi}_{MLE} - \pi)$ converges in distribution to a random variable Z as $n \to \infty$. Suppose we wanted to test the null hypothesis that our data arises from the model in part (a). Before making any further computations, can we necessarily expect Z to follow a normal distribution under the null hypothesis? Explain. Check your answer by computing the appropriate distribution.

[*You may use results from the course, provided you state it clearly.*]

Paper 1, Section II

29J Principles of Statistics

In a regression problem, for a given $X \in \mathbb{R}^{n \times p}$ fixed, we observe $Y \in \mathbb{R}^n$ such that

$$Y = X\theta_0 + \varepsilon$$

for an unknown $\theta_0 \in \mathbb{R}^p$ and ε random such that $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ for some known $\sigma^2 > 0$.

(a) When $p \leqslant n$ and X has rank p, compute the maximum likelihood estimator $\hat{\theta}_{MLE}$ for θ_0. When $p > n$, what issue is there with the likelihood maximisation approach and how many maximisers of the likelihood are there (if any)?

(b) For any $\lambda > 0$ fixed, we consider $\hat{\theta}_\lambda$ minimising

$$\|Y - X\theta\|_2^2 + \lambda\|\theta\|_2^2$$

over $\mathbb{R}^p$. Derive an expression for $\hat{\theta}_\lambda$ and show it is well defined, i.e., there is a unique minimiser for every X, Y and λ.

Assume $p \leqslant n$ and that X has rank p. Let $\Sigma = X^\top X$ and note that $\Sigma = V\Lambda V^\top$ for some orthogonal matrix V and some diagonal matrix Λ whose diagonal entries satisfy $\Lambda_{1,1} \geqslant \Lambda_{2,2} \geqslant \ldots \geqslant \Lambda_{p,p}$. Assume that the columns of X have mean zero.

(c) Denote the columns of $U = XV$ by $u_1, \ldots, u_p$. Show that they are sample principal components, i.e., that their pairwise sample correlations are zero and that they have sample variances $n^{-1}\Lambda_{1,1}, \ldots, n^{-1}\Lambda_{p,p}$, respectively. [*Hint: the sample covariance between* u_i *and* u_j *is* $n^{-1}u_i^\top u_j$.]

(d) Show that

$$\hat{Y}_{MLE} = X\hat{\theta}_{MLE} = U\Lambda^{-1}U^\top Y.$$

Conclude that prediction $\hat{Y}_{MLE}$ is the closest point to Y within the subspace spanned by the normalised sample principal components of part (c).

(e) Show that

$$\hat{Y}_\lambda = X\hat{\theta}_\lambda = U(\Lambda + \lambda I_p)^{-1}U^\top Y.$$

Assume $\Lambda_{1,1}, \Lambda_{2,2}, \ldots, \Lambda_{q,q} >> \lambda >> \Lambda_{q+1,q+1}, \ldots, \Lambda_{p,p}$ for some $1 \leqslant q < p$. Conclude that prediction $\hat{Y}_\lambda$ is approximately the closest point to Y within the subspace spanned by the q normalised sample principal components of part (c) with the greatest variance.