# UNIVERSITETET I OSLO

*Matematisk Institutt*

| | |
|---|---|
| EXAM IN: | **STK 4160/9160:** |
| | **Model Selection and Model Averaging** |
| | **Part I of two parts** |
| WITH: | **Nils Lid Hjort** |
| AUXILIA: | **Calculator, plus one single sheet of paper** |
| | **with the candidate's own personal notes** |
| TIME FOR EXAM: | **Part I: Tuesday 4/vi/2019, 9:00-13:00, written exam;** |
| | **Part II: The Project, 5–17/vi s.y.** |

This exam set contains four exercises and comprises five pages, including an informative one-page Appendix which may be helpful for a few of the exercise points.

## Exercise 1

After seeing three binomial outcomes, do we think the three probabilities are about the same, or different?

(a) Consider $X$ from a binomial $(n, p)$, with $n$ given and $p$ unknown. Find the maximum likelihood estimator $\widehat{p}$, and show that the maximised log-likelihood function may be expressed as

$$\ell_{\max} = n\{\widehat{p}\log\widehat{p} + (1-\widehat{p})\log(1-\widehat{p})\} + \log\binom{n}{x}.$$

(b) Suppose you observe $(X, Y, Z) = (33, 40, 47)$ from three independent binomials $(n, p)$, $(n, q)$, $(n, r)$, with $n = 100$ and three at the outset unknown probabilities $p, q, r$. Two models are considered, first $M_0$, which assumes $p = q = r$, and then $M_1$, which takes the three parameters as not related. Give expressions for the maximised log-likelihood functions $\ell_{\max,0}$ and $\ell_{\max,1}$.

(c) Which model does AIC prefer? And which model does BIC prefer? Comment briefly on what you find.

## Exercise 2

Here we look at the business of least false parameter values and sandwich variance matrices etc., in a simple setup with only one parameter. Suppose $Y_1, \ldots, Y_n$ are independent observations on the positive halfline, with common but unknown density $g(y)$. The parametric model to be used is the common exponential, with density $f(y, \theta) = \theta\exp(-\theta y)$.

(a) Find the score function $u(y, \theta)$ for the model, and find a formula for the maximum likelilhood (ML) estimator $\widehat{\theta}$.

(b) Suppose first that the parametric model is actually correct, so that the density behind the data is $f(y, \theta_0)$ for an appropriate $\theta_0$. Using general results from the curriculum, find the limit distribution of

$$Z_n = \sqrt{n}(\widehat{\theta} - \theta_0).$$

(c) Assume now that the exponential model is not necessarily perfect, but that the mean and variance of the $g(y)$ density, i.e.

$$\xi_0 = \mathrm{E}_g\, Y \quad \text{and} \quad \sigma_0^2 = \mathrm{Var}_g\, Y,$$

are finite. Find the parameter value $\theta_0$ that minimises the Kullback–Leibler distance

$$\mathrm{KL}(g, f(\cdot, \theta)) = \int_0^\infty g(y) \log \frac{g(y)}{f(y, \theta)}\, \mathrm{d}y.$$

Explain that the ML estimator converges to this $\theta_0$ as $n$ increases.

(d) Again using results from the curriculum, find the limit distribution of $Z_n = \sqrt{n}(\widehat{\theta} - \theta_0)$, but now outside model conditions. Check that the formula you find for the variance of the limit distribution agrees with what you find under point (b), if the model happens to be correct.

(e) Give an approximate 95% confidence interval for the least false parameter $\theta_0$.

**Exercise 3**

This exercise looks into a particular way of stretching a given parametric model. I let the start model be the simple exponential one, with $f_0(y, \theta) = \theta \exp(-\theta y)$, though most of the theory goes through with any parametric start.

(a) Consider the function $T(u) = \sqrt{12}(u - \frac{1}{2})$ on $[0, 1]$. Show that $\int_0^1 T(u)\, \mathrm{d}u = 0$ and that $\int_0^1 T(u)^2\, \mathrm{d}u = 1$.

(b) The stretched model we shall work with here has density of the form

$$f(y, \theta, \gamma) = f_0(y, \theta) \exp\{\gamma T(F_0(y, \theta))\}/c(\gamma),$$

where $F_0(y, \theta) = 1 - \exp(-\theta y)$ is the cumulative distribution function. Show that

$$c(\gamma) = \int_0^1 \exp\{\gamma T(u)\}\, \mathrm{d}u.$$

Then use $e^x = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \cdots$ to show that for small $|\gamma|$,

$$c(\gamma) = 1 + \tfrac{1}{2}\gamma^2 + O(|\gamma|^3).$$

(c) Give expressions for $u(y, \theta_0)$ and $v(y, \theta_0)$, the score function components for the two-parameter model, computed at the position $(\theta_0, 0)$ in the parameter space. Show that the Fisher information matrix, computed at this null position, can be expressed as

$$J(\theta_0, 1) = \begin{pmatrix} 1/\theta_0^2 & k/\theta_0 \\ k/\theta_0 & 1 \end{pmatrix},$$

where

$$k = \int_0^\infty (1-w)T(1-\exp(-w))\exp(-w)\,\mathrm{d}w.$$

Some calculations, not required for you to go through during exam hours, show that $k = -\sqrt{3/4}$.

(d) Assume independent observations $Y_1, \ldots, Y_n$ are available. For a parameter of interest $\mu = \mu(\theta, \gamma)$, like the mean or median, one may use either

$$\widehat{\mu}_{\rm narr} = \mu(\widehat{\theta}_{\rm narr}, 0) \quad \text{or} \quad \widehat{\mu}_{\rm wide} = \mu(\widehat{\theta}_{\rm wide}, \widehat{\gamma}_{\rm wide}),$$

using ML estimation under either narrow model or wide model circumstances. For how small values of $|\gamma|$ can the narrow based estimator be expected to be better than the wide one?

(f) Explain briefly how FIC can be used to choose between the two estimators $\widehat{\mu}_{\rm narr}$ and $\widehat{\mu}_{\rm wide}$. How would this compare with using the AIC?

## Exercise 4

Though details and results from this exercise can be extended in several directions, we limit for simplicity discussion to the setup described in the Appendix, with independent observations $y_1, \ldots, y_n$ coming from the model

$$f_n(y) = f(y, \theta_0, \gamma_0 + \delta/\sqrt{n}),$$

where $\theta$ has dimension $p$ and $\gamma$ is one-dimensional. The two models to study are the narrow model, where $\gamma = \gamma_0$ is known, leading to the ML estimator $\widehat{\theta}_{\rm narr}$, and the wide model, leading to the ML estimator $(\widehat{\theta}_{\rm wide}, \widehat{\gamma}_{\rm wide})$. As briefly summarised in the Appendix, we know from Claeskens and Hjort (2008, Chs. 5, 6, 7) that

$$\sqrt{n}(\widehat{\mu}_{\rm narr} - \mu_{\rm true}) \to_{\rm d} \Lambda_0 + \omega\delta \quad \text{and} \quad \sqrt{n}(\widehat{\mu}_{\rm wide} - \mu_{\rm true}) \to_d \Lambda_0 + \omega(\delta - D),$$

where $\mu_{\rm true} = \mu(\theta_0, \gamma_0 + \delta/\sqrt{n})$. Here $D \sim {\rm N}(\delta, \kappa^2)$, with $\kappa^2 = J^{11}$, independent of $\Lambda_0 \sim {\rm N}(0, \tau_0^2)$, and with the required quantities defined in the Appendix.

(a) Consider first the most familiar loss function, that of squared error loss, where one measures precision via

$$L_0(\mu_{\rm true}, \widehat{\mu}) = {\rm err}_n^2 = n(\widehat{\mu} - \mu_{\rm true})^2,$$

with ${\rm err}_n = \sqrt{n}(\widehat{\mu} - \mu_{\rm true})$. Explain how the results noted above lead to

$$\begin{aligned} {\rm E}\, L_0(\mu_{\rm true}, \widehat{\mu}_{\rm narr}) &\to {\rm risk}_{\rm narr} = \tau_0^2 + \omega^2\delta^2, \\ {\rm E}\, L_0(\mu_{\rm true}, \widehat{\mu}_{\rm wide}) &\to {\rm risk}_{\rm wide} = \tau_0^2 + \omega^2\kappa^2. \end{aligned}$$

When will the narrow model lead to better inference for $\mu$ than the wide model? Give your answer in terms of the distance $|\gamma - \gamma_0|$.

(b) Explain briefly how the limiting risk results above, for the $L_0$ squared error loss function, lead to FIC methods for choosing between the narrow and the wide models.

(c) We shall now study a different loss function, namely

$$\begin{aligned} L(\mu_{\rm true}, \widehat{\mu}) &= \frac{1}{a^2}\{\exp(a\,{\rm err}_n) - 1 - a\,{\rm err}_n\} \\ &= \frac{1}{a^2}\big[\exp\{a\sqrt{n}(\widehat{\mu} - \mu_{\rm true})\} - 1 - a\sqrt{n}(\widehat{\mu} - \mu_{\rm true})\big]. \end{aligned}$$

The $a$ is seen as a fine-tuning parameter given by the statistician, in view of the context and analysis of the consequences. Using $e^x = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \cdots$, show that for small $a$, the new loss function is close to the $L_0$ loss function (modulo a constant). The point is that a positive or negative $a$ can 'skew up' the loss function, e.g. in situations where overestimation is more serious than underestimation, etc.

(d) For the following, you may use the fact about moment-generating functions that if $X \sim {\rm N}(\xi, \sigma^2)$, then

$$\mathrm{E}\,\exp(tX) = \exp(t\xi + \tfrac{1}{2}t^2\sigma^2).$$

Show that, under mild regularity conditions,

$$\begin{aligned} \mathrm{E}\,L(\mu_{\rm true}, \widehat{\mu}_{\rm narr}) &\to {\rm risk}_{\rm narr} = \frac{1}{a^2}\big[\exp(a\,\omega\delta + \tfrac{1}{2}a^2\tau_0^2) - 1 - a\,\omega\delta\big], \\ \mathrm{E}\,L(\mu_{\rm true}, \widehat{\mu}_{\rm wide}) &\to {\rm risk}_{\rm wide} = \frac{1}{a^2}\big[\exp\{\tfrac{1}{2}a^2(\tau_0^2 + \omega^2\kappa^2)\} - 1\big]. \end{aligned}$$

(e) Show that there is an interval around zero, for values of $\delta = \sqrt{n}(\gamma - \gamma_0)$, where the narrow risk function is smaller than the wide risk function.

(f) Again, we take the loss function skewness tuning parameter $a$ as given. With a given dataset, explain how one can estimate the two risks.

(g) Also explain how this leads to a more general machinery for FIC.

## Appendix

The following is a mini-summary of some of the core material from Chapters 6–7 in Claeskens and Hjort (2008), pertaining to a certain local neighbourhood large-sample framework, set here in the simpler framework of i.i.d. data. Notation and results given here may be used in Exercises 3 and 4 without proofs or further detailed discussion.

Assume independent observations $Y_1, \ldots, Y_n$ stem from a distribution with density function $f(y, \theta, \gamma)$, with $\theta$ of dimension $p$ and $\gamma$ of dimension $q$. With $\gamma_0$ a suitable null point in the parameter range for $\gamma$, corresponding to $f(y, \theta, \gamma_0)$ being a $p$-dimensional 'narrow model', assume that $\gamma = \gamma_0 + \delta/\sqrt{n}$, i.e. that the data follow the density

$$f_n(y) = f(y, \theta_0, \gamma_0 + \delta/\sqrt{n}).$$

As in the core material of the chapters mentioned, we shall work with a focus parameter $\mu = \mu(\theta, \gamma)$, with true value $\mu_{\rm true} = \mu(\theta_0, \gamma_0 + \delta/\sqrt{n})$. The necessary notation includes the Fisher information matrix $J$ of dimension $(p+q) \times (p+q)$, computed at the null model, with inverse $J^{-1}$. These matrices have blocks according to the usual notation

$$J = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix} \quad \text{and} \quad J^{-1} = \begin{pmatrix} J^{00} & J^{01} \\ J^{10} & J^{11} \end{pmatrix},$$

with $J_{00}$ being of size $p \times p$, $Q = J^{11}$ of size $q \times q$, etc. Also, let

$$\tau_0^2 = (\tfrac{\partial \mu}{\partial \theta})^{\rm t} J_{00}^{-1} \tfrac{\partial \mu}{\partial \theta} \quad \text{and} \quad \omega = J_{10} J_{00}^{-1} \tfrac{\partial \mu}{\partial \theta} - \tfrac{\partial \mu}{\partial \gamma},$$

with derivatives computed at the null model. With maximum likelihood estimators $\widehat{\mu}_{\rm narr} = \mu(\widehat{\theta}_{\rm narr}, \gamma_0)$ and $\widehat{\mu}_{\rm wide} = \mu(\widehat{\theta}_{\rm wide}, \widehat{\gamma}_{\rm wide})$, under respectively the narrow and the wide models, we have

$$\sqrt{n}(\widehat{\mu}_{\rm narr} - \mu_{\rm true}) \to_d \Lambda_0 + \omega^{\rm t}\delta \quad \text{and} \quad \sqrt{n}(\widehat{\mu}_{\rm wide} - \mu_{\rm true}) \to_d \Lambda_0 + \omega^{\rm t}(\delta - D),$$

where $\Lambda_0 \sim {\rm N}(0, \tau_0^2)$ and $D \sim {\rm N}_q(\delta, Q)$ are independent. More generally, with $\widehat{\mu}_S$ the maximum likelihood estimator under submodel $S$, where $S$ is a subset of $\{1, \ldots, q\}$,

$$\sqrt{n}(\widehat{\mu}_S - \mu_{\rm true}) \to_d \Lambda_0 + \omega^{\rm t}(\delta - G_S D),$$

where

$$G_S = \pi_S^{\rm t} Q_S \pi_S Q^{-1} = \pi_S^{\rm t}(\pi_S Q^{-1} \pi_S^{\rm t})^{-1} \pi_S Q^{-1}$$

is a $q \times q$ matrix determined by this $S$, involving the projection matrices $\pi_S$, where $\pi_S v$ maps $v = (v_1, \ldots, v_q)$ to the subvector $v_S$ with $v_j$ for $j \in S$.