UNIVERSITETET I OSLO

Matematisk Institutt

EXAM IN:	**STK 4160/9160 – Model Selection and Model Averaging**
	Part II 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: The Project, 3–15/vi/2015;**
	Part II: Thursday 11/vi s.y., 14:30–18:30, written exam

This exam set contains four exercises and comprises four pages, including an informative Appendix which may be helpful for Exercises 3 and 4.

Exercise 1

Suppose independent observations $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ stem from two samples, the first from the $\mathrm{N}(a,1)$ and the second from the $\mathrm{N}(b,1)$ distribution. We shall use this simple setup to investigate certain questions related to the AIC and BIC. The sample means are as usual denoted $\bar{X}$ and $\bar{Y}$.

(a) Show that the log-likelihood functions for the x-sample may be expressed as

$$\ell_x(a) = -\tfrac{1}{2}Q_x - \tfrac{1}{2}n(\bar{X}-a)^2 - \tfrac{1}{2}n\log(2\pi),$$

with $Q_x = \sum_{i=1}^n (X_i - \bar{X})^2$. Set up the full log-likelihood $\ell(a,b)$ for the observed data. Identify the maximum likelihood estimator $\widehat{a}$ for a and $\widehat{b}$ for b and give their distributions (without proof).

(b) We now consider two models. Model M_0 takes $a = b$, so that all the $2n$ observations come from the same population, whereas model M_1 does not assume anything regarding a or b, hence left as two free parameters. Find explicit expressions for

$$\ell_{0,\max} = \max\{\ell(a,b)\colon \text{model } M_0\} \quad \text{and} \quad \ell_{1,\max} = \max\{\ell(a,b)\colon \text{model } M_1\},$$

and use these to show that

$$\Delta_n = 2(\ell_{1,\max} - \ell_{0,\max}) = \tfrac{1}{2}n(\bar{X}-\bar{Y})^2.$$

(c) Identify precisely when AIC selects model M_1 over M_0, and also precisely when BIC selects model M_1 over M_0, in terms of $\bar{X}$ and $\bar{Y}$.

(d) Assume for this point that model M_0 is actually correct. Find expressions for

$$p_{n,0} = \Pr\{\text{AIC selects } M_1\} \quad \text{and} \quad q_{n,0} = \Pr\{\text{BIC selects } M_1\}.$$

Determine also the limits of $p_{n,0}$ and $q_{n,0}$ as n increases. I mention here, in case this might be of use for your analysis, that

$$\Pr\{\chi^2_1 \le x\} = 0.683,\ 0.843,\ 0.917,\ 0.954,\ 0.975 \quad \text{for } x = 1,2,3,4,5.$$

(e) Let now $\delta = a - b$, where the previous point concerned the case $\delta = 0$. Assume for this point that $a \neq b$, and consider

$$p_n(\delta) = \Pr_\delta\{\text{AIC selects } M_1\} \quad \text{and} \quad q_n(\delta) = \Pr_\delta\{\text{BIC selects } M_1\}.$$

Find the limits of $p_n(\delta)$ and $q_n(\delta)$ as n increases. Comment briefly on these findings.

Exercise 2

Consider again the set-up of Exercise 1, with a sample from $\mathrm{N}(a, 1)$ and another sample from $\mathrm{N}(b, 1)$.

(a) For estimating the parameter $\delta = a - b$, determine $\widehat{\delta}_0$ and $\widehat{\delta}_1$, the maximum likelihood estimators under models M_0 and M_1, respectively.

(b) Determine the mean squared errors of these two estimators. When is the M_0 model based estimator better than the M_1 based estimator?

(c) Define risk functions

$$r_{n,\mathrm{AIC}}(\delta) = n\,\mathrm{E}(\widehat{\delta}_{\mathrm{AIC}} - \delta)^2 \quad \text{and} \quad r_{n,\mathrm{BIC}}(\delta) = n\,\mathrm{E}(\widehat{\delta}_{\mathrm{BIC}} - \delta)^2,$$

where

$$\widehat{\delta}_{\mathrm{AIC}} = \begin{cases} \widehat{\delta}_0 & \text{if AIC chooses } M_0, \\ \widehat{\delta}_1 & \text{if AIC chooses } M_1, \end{cases} \qquad \widehat{\delta}_{\mathrm{BIC}} = \begin{cases} \widehat{\delta}_0 & \text{if BIC chooses } M_0, \\ \widehat{\delta}_1 & \text{if BIC chooses } M_1. \end{cases}$$

It takes a bit of time to find clear mathematical expressions for these two post-selection risk functions and you are not required to do so during today's examination hours. You may attempt to compute the risk functions at zero, however (as this is easier than for $\delta \neq 0$), and you should also indicate how you believe the two risk functions will look like, based on similar studies from the curriculum.

Exercise 3

For this exercise you should freely use material, notation and results from the exam set's Appendix. The material there uses the usual $f(y, \theta, \gamma)$ notation with θ of dimension p and γ in general being a parameter of length q, but in the present exercise we are content to study the one-dimensional case of $q = 1$. It is then also convenient to write κ^2 for the lower right-hand corner element J^{11} of the $(p+1) \times (p+1)$ inverse Fisher information matrix. As also explained and summarised in the Appendix, there is a focus parameter $\mu = \mu(\theta, \gamma)$ at play, with narrow model and wide model estimators $\widehat{\mu}_{\mathrm{narr}}$ and $\widehat{\mu}_{\mathrm{wide}}$, respectively.

(a) Consider first the usual loss function for estimation, namely squared error, scaled here with n to have proper limits:

$$L(\mu, \widehat{\mu}) = n(\widehat{\mu} - \mu)^2.$$

The risk function is the expected loss, as a function of the parameters. Explain that the limiting risk functions for the two estimators above become

$$r_{\rm narr}(\delta) = \tau_0^2 + \omega^2\delta^2 \quad \text{and} \quad r_{\rm wide}(\delta) = \tau_0^2 + \omega^2\kappa^2.$$

(b) For which range of the parameter δ will the narrow model lead to more precise estimators than with the wide model, under this squared error loss function, for large n? Translate this result to the original parameter scale γ.

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

$$L^*(\mu, \widehat{\mu}) = \begin{cases} 1 & \text{if } |\sqrt{n}(\widehat{\mu} - \mu)| > \varepsilon, \\ 0 & \text{if } |\sqrt{n}(\widehat{\mu} - \mu)| \le \varepsilon, \end{cases}$$

for ε a small number. Show that the limiting risk functions for the two estimators above may be represented as

$$r_{\rm narr}(\delta) \doteq 1 - h_{\rm narr}(0)\, 2\varepsilon \quad \text{and} \quad r_{\rm wide}(\delta) \doteq 1 - h_{\rm wide}(0)\, 2\varepsilon,$$

where $h_{\rm narr}(x)$ and $h_{\rm wide}(x)$ are the densities of respectively $\Lambda_0 + \omega\delta$ and $\Lambda_0 + \omega(\delta - D)$. (The '$a(\varepsilon) \doteq c\varepsilon$' notation indicates that $a(\varepsilon)/\varepsilon \to c$ as $\varepsilon \to 0$.)

(d) Show that $r_{\rm narr}(\delta) < r_{\rm wide}(\delta)$ corresponds to $s_{\rm narr}(\delta) > s_{\rm wide}(\delta)$, where

$$s_{\rm narr}(\delta) = \phi\Big(\frac{\omega\delta}{\tau_0}\Big)\frac{1}{\tau_0} \quad \text{and} \quad s_{\rm wide}(\delta) = \phi(0)\frac{1}{(\tau_0^2 + \omega^2\kappa^2)^{1/2}},$$

with $\phi(x) = (2\pi)^{-1/2}\exp(-\frac{1}{2}x^2)$ the standard normal density.

(e) Going to the log-scale of things, explain how the above study of the loss function L^* leads to transformed risk functions

$$r^*_{\rm narr}(\delta) = \log\tau_0^2 + \omega^2\delta^2/\tau_0^2 \quad \text{and} \quad r^*_{\rm wide}(\delta) = \log(\tau_0^2 + \omega^2\kappa^2),$$

with low value of $r^*(\delta)$ indicating better performance. Find the tolerance radius c around the narrow model, with the property that when $|\delta| \le c$, then the narrow model leads to better estimation of μ than the wide model does.

(f) Put up natural estimators of $r^*_{\rm narr}(\delta)$ and $r^*_{\rm wide}(\delta)$, with the property that these become unbiased in the limit situation. Explain, but briefly, how this leads to a new focused information criterion, say FIC*.

(g) Give also a FIC* = FIC$^*(S)$ formula for the general case of a q-dimensional γ parameter, for the 2^q candidate models corresponding to subsets S of $\{1, \ldots, q\}$.

Exercise 4

Write up the essence of the AFIC method (the average weighted focused information criterion), using a maximum of two pages. You are again free to use material, notation and results summarised in the Appendix below.

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 θ of dimension p and γ of dimension q. With γ_0 a suitable null point in the parameter range for γ, 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

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

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 π_S, where $\pi_S v$ maps $v = (v_1, \ldots, v_q)$ to the subvector v_S with v_j for $j \in S$.