UNIVERSITETET I OSLO

Matematisk Institutt

EXAM IN:	**STK 4160/9160:**
	Model Selection and Model Averaging
WITH:	**Nils Lid Hjort**
TIME FOR EXAM:	**26 June 2009 at 11:55 – 8 June s.y. at 14:00**

This is the exam set for STK 4160, spring semester 2009. It is made available on the course website as of *Tuesday 26 May 12:00*, and candidates must submit their written reports by *Monday 8 June 14:00* (or earlier), to the reception office at the Department of Mathematics, *in duplicate*. The supplementary oral examinations take place June 11 and 12 (practical details for these will be provided later).

Reports may be written in nynorsk, bokmål, riksmål, English or Latin, and should preferably be text-processed (TeX, LaTeX, `Word`), but may also be hand-processed. Give your name on the first page. Write concisely (in der Beschränkung zeigt sich erst der Meister; brevity is the soul of wit; краткость – сестра таланта). Relevant figures need to be included in the report. Copies of machine programmes used (in `R`, or `matlab`, or similar) are also to be included, perhaps as an Appendix to the report. The full exam set *is* (admittedly) labourious, and candidates are graciously allowed not to despair if they do not manage to answer all questions well.

Importantly, each student needs to submit *two special extra pages* with her or his report. *The first* (page A) is the 'erklæring' (self-declaration form), properly signed; it is available at the webpage as 'Exam Project, page A, declaration form'. *The second* (page B) is the student's one-page summary of the exam project report, which should also contain a brief self-assessment of its quality.

This exam set contains three exercises and comprises nine pages (including a two-page Appendix with useful details for `R` work).

Exercise 1

START WIDE, EXPAND FURTHER, and never look back. We shall at least partly be following Arnold Schwarzenegger's advice in this exercise, in that we aim at expanding upon the most widely-used of all models for a probability density, namely the normal one; we do reserve the right to look back, however. The normal model is

$$f(y,\xi,\sigma) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\exp\Big\{-\tfrac{1}{2}\frac{(y-\xi)^2}{\sigma^2}\Big\} = \phi\Big(\frac{y-\xi}{\sigma}\Big)\frac{1}{\sigma} \quad \text{for } y \in I\!R,$$

in terms of a location parameter ξ and scale parameter σ; here $\phi(x) = (2\pi)^{-1/2}\exp(-\frac{1}{2}x^2)$ is the standard normal density. The model is often effective even when it does not perfectly fit the data, but it is potentially fruitful to construct wider families with more modelling flexibility. This exercise is about one such type of extension, in terms of a log-linear expansion in certain basis functions.

Define the functions

$$\psi_j(u) = \sqrt{2}\cos(j\pi u) \quad \text{for } u \in [0,1],$$

for $j = 1, 2, 3, \ldots$, supplemented by the unit function $\psi_0(u) = 1$. These are so-called orthonormal, with respect to the uniform distribution, in the sense that

$$\int_0^1 \psi_j(u)^2 \, \mathrm{d}u = 1 \quad \text{and} \quad \int_0^1 \psi_j(u)\psi_k(u) \, \mathrm{d}u = 0 \quad \text{for } j \neq k.$$

That these properties hold follow via traditional integration exercises, but you do not need to prove them here. The construction below utilises only $\psi_1, \psi_2, \ldots$, i.e. not ψ_0 as such, but the fact that the orthonormality property holds also with respect to ψ_0 is useful.

(a) Our extended model, or order m, takes the form

$$f_m(y, \xi, \sigma, a) = \phi\Big(\frac{y-\xi}{\sigma}\Big)\frac{1}{\sigma}\exp\Big\{\sum_{j=1}^m a_j \psi_j\Big(\Phi\Big(\frac{y-\xi}{\sigma}\Big)\Big)\Big\}\Big/ k_m(a),$$

with $a \in I\!R^m$ consisting of additional parameters $a_1, \ldots, a_m \in I\!R$. Here $\Phi(\cdot)$ is the cumulative standard normal distribution function, as usual (the integral of ϕ). Show that this actually defines a probability density on $(-\infty, \infty)$, with integration constant

$$k_m(a) = k_m(a_1, \ldots, a_m) = \int_0^1 \exp\Big\{\sum_{j=1}^m a_j \psi_j(u)\Big\} \, \mathrm{d}u.$$

Below we shall for concreteness of illustration focus on the second order model $m = 2$, though results easily extend to the general case.

(b) To get a feel for the flexibility of this family, draw some of these densities in the same diagram, for order $m = 2$, for some values of a_1, a_2 in the vicinity of zero. Use $\xi = 0$ and $\sigma = 1$, and include the null model among those plotted (i.e. where the a_j are equal to zero). Note that the k_2 function may be computed in `R` via the numerical integration routine given in the Appendix.

(c) Use the Taylor expansion $\exp(v) \doteq 1 + v + \frac{1}{2}v^2$ of order two, which provides an acceptable approximation for v close to zero, to show that

$$k_2(a_1, a_2) \doteq 1 + \tfrac{1}{2}(a_1^2 + a_2^2)$$

for a_j values close to zero. [The formal mathematical statement is that $k_m(a) = 1 + \frac{1}{2}\|a\|^2 + o(\|a\|^2)$.]

(d) Show that the score function, with respect to parameters ξ, σ, a_1, a_2, computed at the normal null model, takes the form

$$\begin{pmatrix} (1/\sigma)\varepsilon \\ (1/\sigma)(\varepsilon^2 - 1) \\ \psi_1(\Phi(\varepsilon)) \\ \psi_2(\Phi(\varepsilon)) \end{pmatrix},$$

where $\varepsilon = (y - \xi)/\sigma$.

(e) Use this result to show that the model's information matrix, computed at the null model, is equal to

$$J = \begin{pmatrix} 1/\sigma^2 & 0 & c_1/\sigma & c_2/\sigma \\ 0 & 2/\sigma^2 & d_1/\sigma & d_2/\sigma \\ c_1/\sigma & d_1/\sigma & 1 & 0 \\ c_2/\sigma & d_2/\sigma & 0 & 1 \end{pmatrix},$$

where

$$c_j = \text{cov}\{\varepsilon, \psi_j(\Phi(\varepsilon))\} \quad \text{and} \quad d_j = \text{cov}\{\varepsilon^2 - 1, \psi_j(\Phi(\varepsilon))\}$$

for $j = 1, 2, \ldots$. Find also numerical values for those constants that are needed to analyse the second order model:

$$(c_1, c_2) = (-0.9484, 0) \quad \text{and} \quad (d_1, d_2) = (0, 1.0496).$$

(f) Use these results to exhibit 'the Q matrix', the lower right-hand block of the J^{-1} matrix (computed at the null model):

$$Q = \begin{pmatrix} 9.9521 & 0 \\ 0 & 2.2261 \end{pmatrix}.$$

(g) For a smooth focus parameter $\mu = \mu(\xi, \sigma, a_1, a_2)$, derive a suitable expression for the usual vector

$$\omega = J_{10} J_{00}^{-1} \frac{\partial \mu}{\partial \theta} - \frac{\partial \mu}{\partial a}$$

(where θ in this case is (ξ, σ), and where formulae again are to be derived under null model conditions). Suppose in particular that one wishes to estimate the log-density at some given position y_0, i.e.

$$\mu = \mu(y_0) = \log f(y_0, \xi, \sigma, a_1, a_2).$$

Show that ω then takes the form

$$\omega = \omega(y_0) = \begin{pmatrix} c_1 x_0 + \frac{1}{2} d_1 (x_0^2 - 1) + \psi_1(\Phi(x_0)) \\ c_2 x_0 + \frac{1}{2} d_2 (x_0^2 - 1) + \psi_2(\Phi(x_0)) \end{pmatrix},$$

where $x_0 = (y_0 - \xi)/\sigma$. Compute ω for $y_0 = \xi$ and for $y_0 = \xi + 2\sigma$.

(h) Assume independent observations $y_1, \ldots, y_n$ stem from the density

$$f_{\text{true}}(y) = f(y, \xi, \sigma, \delta_1/\sqrt{n}, \delta_2/\sqrt{n})$$

and that n is at least moderately large. Give mathematical descriptions of (i) the region A of (δ_1, δ_2) where estimation based on the two-parameter normal model is better than using the four-parameter expanded model f_2, for all smooth estimands μ; (ii) the region $B(y_0)$ of (δ_1, δ_2) where estimation of the log-density at position y_0 is better using the normal model than using the four-parameter expanded model. Try to display regions A, along with $B(y_0)$ for $y_0 = \xi$ and for $y_0 = \xi + 2\sigma$, in the same diagram. Comment briefly of what you find.

(i) Let y^* be the mode of the distribution, i.e. the position at which $f_2(y)$ is maximal. When data $y_1, \ldots, y_n$ have been observed, explain briefly how the Focussed Information Criterion may be used to choose among the three models corresponding to order zero (i.e. the normal), one, two, when the purpose is to estimate y^* well. Exhibit the necessary quantities for your formulae, and make clear what your assumptions are for your FIC scheme to be valid.

Exercise 2

The 2009 Oscar Award goes to ... Shani Davis! This was announced from Oslo a week or so ago (19/v/9), and no, this is not the Academy Award, even though the history of its informal name arguably and fittingly can be traced to Bette Davis; but rather *the Oscar Mathisen Award*, the most prestigious award of speedskating. As we recall, Davis won the Oscar also in 2005, for his World Allround Championship triumph in Moskva, and this season he managed to also become the World Sprint Champion (also in Moskva). The official citation points specifically to his two world records set at the Olympic Oval in Salt Lake City in March: 1:06.42 on the 1000-m and 1:41.80 on the 1500-m.

The present exercise relates to estimating and assessing the distribution of personal best times for the 1500-m, via model selection mechanisms (admittedly using data from the Adelskalenderen as of April 2006, for exam project convenience reasons, rather than from April 2009). As the figure indicates, Davis's 1:41.80 is spectacularly impressive.

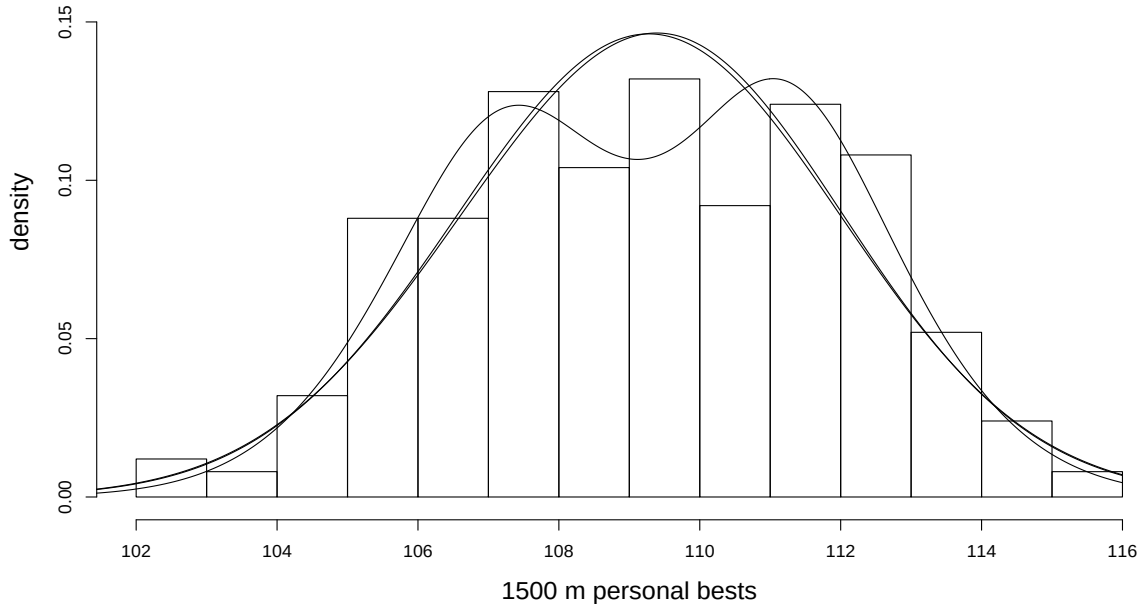

Figure A: Histogram over the personal bests of the best 250 skaters in the world (as per the Adelskalenderen as of post-season 2006), along with three density estimates. Shani Davis's new world record is 1:41.80, and Håvard Bøkko's Norwegian record is 1:42.67.

(a) Access the data from the Claeskens & Hjort book's website. Fit the three log-linear expansion models corresponding to order zero, one, two (cf. Exercise 1), where order zero corresponds to the ordinary Gaußian curve. For each model, give parameter estimates and standard errors (estimates of standard deviation). If you manage, attempt to duplicate Figure A. [You may use programming tools as indicated in the Appendix, involving (i) a way of programming the $k_m(a)$ function and (ii) advice regarding the start value for the numerical optimisation algorithms.]

(b) Compute AIC and BIC scores for the three models. Which model is best, as judged by respectively the AIC and the BIC? Give also numerical approximations to the Bayesian posterior probabilities for the three models. Comment also on which model you would prefer yourself.

(c) Are there other models that you think could be useful for describing and assessing the distribution of 1500-m personal best times? How would you compare such an alternative model to the three dealt with above?

Exercise 3

IF A MAN WHO CANNOT COUNT finds a four-leaf clover, does he have the right to be lucky?, asks Stanisław Jerzy Lec ("Czy człowiek, który znalazł czterolistną koniczynę, a nie umie liczyć, ma też prawo do szczęścia?"). This exercise at any rate attempts to combine counting with luck, mathematics and artistic skills; specifically, we shall count the number y of days of absence from a certain American junior high school and see how it is related to x (1 if boy, 0 if girl), z_1 (average grade for mathematics tests), and z_2 (average grade for artistic and language tests). Both scores z_1 and z_2 are normalised to the range $[0, 100]$.

The datafile `attendance1-data` (available at the course website) gives a table of seven columns pertaining to such attendance-and-absence data for the school in question, for a total of $n = 159$ pupils. The columns are respectively id-number; school = 1; boy-girl indicator x; mathematics score z_1; arts and languages score z_2; days attended in a semester; and days y absent in the same semester. We ignore some of the information here and concentrate on y, as influenced by x, z_1, z_2. We take x as protected but question whether z_1 and z_2 need to be included for explaining or predicting y. As far as points (a)–(e) below are concerned, our machinery is going to be that of regression models with independent Poisson driven counts, but we go further in points (f)–(g).

(a) Consider the four natural candidate models 0 (none of z_1 and z_2 included), 1 (z_1 but not z_2 included), 2 (z_2 but not z_1 included), 12 (both of z_1 and z_2 included), each a special case of the widest Poisson regression models

$$y_i \sim \text{Pois}(\xi_i) \quad \text{with } \xi_i = \exp(\beta_0 + \beta_1 x_i + \gamma_1 z_{1,i} + \gamma_2 z_{2,i}) \quad \text{for } i = 1, \ldots, n.$$

Fit each model using maximum likelihood. Provide a table giving the AIC and BIC scores. Comment on your findings, and briefly discuss the underlying assumptions.

(b) For the parameter $\mu = \mathrm{E}(y \,|\, x_0, z_{1,0}, z_{2,0})$, with covariate values set to $(1, 75, 75)$ (a boy, with relatively good marks), compute for each of the four candidate models the point estimate and an approximate 95% confidence interval (the latter computed in standard fashion, using standard errors and normal approximation, and trusting that the model in question is adequate).

(c) For this focus parameter μ, carry out FIC analysis with the four candidate models. Comment on your findings and assumptions.

(d) In this situation, compute also the 'weighted AIC' and 'weighted FIC' estimate of μ. What are the distributions of these estimators like?

(e) At this school one is concerned with the apparent fact that the girls tend to be more frequently absent than the boys. One wishes to learn more about how school competence and performance in the two fields may relate to absence rates, and in particular to the probability that the absence level exceeds the threshold $y_0 = 10$ days. Define therefore

$$\mu = \mu(z_0) = \Pr\{y > y_0 \,|\, x_0 = 0, z_{1,0} = z_0, z_{2,0} = z_0\} = 1 - G(y_0, \xi(0, z_0, z_0)),$$

where $G(y, \xi)$ is the cumulative Poisson distribution function with parameter ξ, and $\xi(x, z_1, z_2)$ is of the exponential form above. Carry out an appropriate AFIC analysis (averaged or weighted FIC), again comparing the four models 0, 1, 2, 12, across nineteen equally spaced and equally important positions $(5, 5), \ldots, (95, 95)$ in the space of (z_1, z_2) marks, for girls. Comment on the results.

(f) The above analyses rest on the assumption of Poisson distributed absence counts. Some tests reveal however that this assumption is rather far from being satisfied; in particular, the fitted Poisson distributions do not quite produce zeros often enough, compared with the $N_0 = 19$ cases of $n = 159$ pupils with no absence at all. There are various strategies aiming at repairing for such underprediction of zeros, and the present point aims at pushing you towards fitting and evaluating one such model. The idea is simply to operate with a single p_0 probability for $y_i = 0$, across pupils, and then use Poisson regression for the other data, but appropriately conditional on having counts $y_i \geq 1$. Show that this leads to the model likelihood function

$$L_n(p_0, \beta, \gamma) = p_0^{N_0}(1 - p_0)^{n - N_0} \prod_{i: y_i \geq 1} \frac{f(y_i, \xi_i)}{1 - \exp(-\xi_i)},$$

where $f(y, \xi)$ is the Poisson density at y with parameter ξ, and ξ_i is as in point (a) above (i.e. including each of x, z_1, z_2 as covariates). Fit the model by maximum likelihood; exhibit parameter estimates and estimated standard deviations; evaluate the AIC score; compute the model based estimate of the parameter μ worked with in points (b) and (c); and comment on your findings.

(g) The final model I wish you to try out is the following. Instead of taking y_i to be Poission with fixed parameter ξ_i, assume rather that $y_i \,|\, \xi_i \sim \text{Pois}(\xi_i)$ but that $\xi_i \sim \text{Gamma}(c \exp(x_i^{\rm t}\beta), c)$. Estimate once more the focus parameter μ of points (b)–(c), along with a model-based confidence interval. Compute again the AIC score, and finally also the Takeuchi model-robust TIC scores for each of the six models worked with in this exercise. How do you conclude – which of the six models appears to be best?

Appendix: Useful details for R programming

1. The following is one way to programme the integration constant $k_2(a)$, met in Exercises 1 and 2, using R. One first defines functions ψ_1, ψ_2 (and yet further ψ_j functions, if required) via

```
psi1 <- function(u)
{sqrt(2)*cos(pi*u)}
```

etc., and then uses

```
k2 <- function(a)
{
inte2 <- function(u)
{exp(a[1]*psi1(u) + a[2]*psi2(u))}
integrate(inte2,0,1)$value
}
```

Functions $k_m(a)$ for other values of m may be programmed similarly. This is also useful for maximum likelihood estimation inside these extended normal models. When using iteration based algorithms like `nlm`, it is often useful to use $(\bar{y}, \text{sd}(y), 0, \ldots, 0)$ as start value, where $\bar{y}$ and $\text{sd}(y)$ are mean and standard deviation for the data (i.e. more or less the maximum likelihood estimates for the narrow normal model, before expansion).

2. To work properly with the absence from school data, you may start out as follows, creating a matrix `data` of dimension 159×7:

```
data <- matrix(scan("attendance1-data", skip=6), byrow=T, ncol=7)
yy <- data[ ,7]
xx <- data[ ,3] # boy = 1, girl = 0
z1 <- data[ ,4] # mathematics score
z2 <- data[ ,5] # languages and arts score
nn <- length(yy)
```

3. Estimation in the Poisson regression model may be performed using

```
glm(yy ~ xx + z1 + z2, family=poisson)
```

*&*cetera.

4. To carry out inference in the zero-inflated model of Exercise 3(e), it is useful to work with the subset of `data` corresponding precisely to the 140 cases where $y_i \geq 1$. This may be organised in various ways, e.g. as follows:

```
check <- 1*(yy >= 1)
indexplus <- (1:nn)[check == 1]
nplus <- length(indexplus)
dataplus <- 0*(1:nplus) %*% t(1:7)
for (j in 1:nplus)
{ dataplus[j, ] <- data[indexplus[j], ] }
```

Figure B: Shani Davis, the inspiration for Exercise 2. His personal bests are 34.78, 1:06.42, 1:41.80, 6:10.49, 13:05.94, he holds three world records, and won an Olympic gold and a silver in the 2006 Torino games. He is also the current Adelskalenderen Leader (since March 6, 2009). – Photo: Kirsti Biseth.