UNIVERSITETET I OSLO

Matematisk Institutt

EXAM IN:	**STK 9051SP, special curriculum: Statistical Inference via Minimum Divergence Methods**
FOR:	**Sam-Erik Walker**
WITH:	**Nils Lid Hjort**
TIME FOR EXAM:	**23.v.–6.vi.2016**

This is the exam project set for STK 9051SP, special curriculum on statistical inference via minimum divergence methods, spring semester 2016. It is made available for the candidate as of *Monday 23 May 12:00*, and he should submit his written report by *Monday 6 June 13:00* (or earlier), electronically as a pdf, to Nils Lid Hjort. There will also be a *conversation* with the candidate, Hjort and a colleague, with a blackboard nearby, on Tuesday 7 June; this conversation might touch both the exam project report and other aspects of the special curriculum.

The candidate is required to work by himself, i.e. independently of others. Importantly, in his report the candidate should also include a one-page summary of the work carried out, and this should also contain a brief self-assessment of its quality.

This exam set contains five exercises and comprises five pages.

Exercise 1

This exercise concerns using the BHHJ inference approach (for Basu, Harris, Hjort, Jones) for the gamma distribution, i.e. when the working model has density of the form

$$f(y,a,b) = \frac{b^a}{\Gamma(a)} y^{a-1} \exp(-by) \quad \text{for } y > 0.$$

(a) We start out in traditional likelihood modus. Write down formulae for the score functions

$$u_1(y,a,b) = \partial \log f(y,a,b)/\partial a,$$
$$u_2(y,a,b) = \partial \log f(y,a,b)/\partial b.$$

Also find explicit expressions for the elements of the Fisher information matrix $I = I(a,b)$. Identify the limit distribution for

$$\begin{pmatrix} \sqrt{n}(\widehat{a}_{\rm ml} - a) \\ \sqrt{n}(\widehat{b}_{\rm ml} - b) \end{pmatrix},$$

where $(\widehat{a}_{\rm ml}, \widehat{b}_{\rm ml})$ are the maximum likelihood estimators. What is the limiting correlation between $\widehat{a}_{\rm ml}$ and $\widehat{b}_{\rm ml}$?

(b) With a fixed positive α, consider the BHHJ method associated with the divergence

$$d_\alpha(g, f_\theta) = \int \{f_\theta^{1+\alpha} - (1+1/\alpha) g f_\theta^\alpha + (1/\alpha) g^{1+\alpha}\} \, \mathrm{d}y$$

(here with $\theta = (a, b)$). For a dataset $y_1, \ldots, y_n$, write down the explicit data-based function $H_n(a, b)$ to be minimised in order to produce the BHHJ estimates $(\widehat{a}, \widehat{b})$ – when explicit dependency upon α is helpful for the notation, we might write $(\widehat{a}_\alpha, \widehat{b}_\alpha)$.

(c) General theory, from the BHHJ paper and elsewhere, implies that

$$\begin{pmatrix} \sqrt{n}(\widehat{a} - a) \\ \sqrt{n}(\widehat{b} - b) \end{pmatrix} \to_d \mathrm{N}(0, \Sigma),$$

for a certain matrix $\Sigma = \Sigma_\alpha$. Give a formula for this matrix, and explain clearly how it may be estimated from data, when these are assumed to form an i.i.d. set from the same underlying density $g(y)$.

(d) Assume then that the model is actually correct, for certain (a_0, b_0). Give suitable formulae for Σ. For the special case $(a_0, b_0) = (5.555, 2.222)$, compute Σ_α, for α ranging from 0 to say 1.5, so that you can display the two curves

$$r_1(\alpha) = \big[\Sigma_{\alpha,1,1} / \{I(a_0, b_0)^{-1}\}_{1,1}\big]^{1/2},$$
$$r_2(\alpha) = \big[\Sigma_{\alpha,2,2} / \{I(a_0, b_0)^{-1}\}_{2,2}\big]^{1/2}.$$

Explain how these curves may be relevantly interpreted, and comment on what you find.

(e) Consider the mean parameter $\mu = \mathrm{E}_g Y$, which under model conditions is equal to a/b. Assuming that the model is correct, find the limit distribution

$$\sqrt{n}(\widehat{a}_{\mathrm{ml}}/\widehat{b}_{\mathrm{ml}} - a/b) \to_d \mathrm{N}(0, \kappa_0^2),$$

with a formula for κ_0.

(f) Then first without assuming the model to be correct, identify the limit distribution for

$$\sqrt{n}(\widehat{a}_\alpha/\widehat{b}_\alpha - a_{0,\alpha}/b_{0,\alpha}),$$

for the relevant $(a_{0,\alpha}, b_{0,\alpha})$. Assuming that the model is correct, again with $(a_0, b_0) = (5.555, 2.222)$, compute the limit standard deviation, say κ_α, and display the curve

$$r_3(\alpha) = \kappa_\alpha / \kappa_0.$$

Comment on your findings.

Exercise 2

Access the dataset on $n = 141$ lifelengths from Roman era Egypt (the last century b.C.), ranging from 1.5 yr to 96.0 yr. The object here is to fit the gamma distribution to these data, via the BHHJ method, along with inference for certain parameters of interest.

(a) Fit the data to the gamma distribution using maximum likelihood. I find (1.6091, 0.0524) for the estimates.

(b) Then apply the BHHJ method, for a grid of α values from 0 to say 1.5. Display the resulting curves $\widehat{a}_\alpha$ and $\widehat{b}_\alpha$, and comment. Also construct a figure with a histogram of the data points, with two estimated gamma density curves in the same diagram; that for the maximum likelihood and that for the BHHJ with $\alpha = 1$.

(c) Use these parameter estimates (the BHHJ method, for a grid of α) to also display curves for

$$\widehat{\mu}_\alpha = \widehat{a}_\alpha/\widehat{b}_\alpha \quad \text{and} \quad \widehat{\nu}_\alpha = F^{-1}(\tfrac{1}{2}, \widehat{a}_\alpha, \widehat{b}_\alpha),$$

where $F^{-1}(\cdot, a, b)$ is the quantile function for the gamma. Comment on these curves.

(d) Now focus on the mean parameter $\mu_\alpha = a_\alpha/b_\alpha$. Without assuming that the gamma model is correct, construct and compute an approximate 90% confidence for μ_α, and do this for each α on a grid of such values, say $[\text{low}_\alpha, \text{up}_\alpha]$. Display a graph of $[\text{low}_\alpha, \widehat{\mu}_\alpha, \text{up}_\alpha]$.

(e) Now, which of these (approximate) 90% confidence intervals would you recommend to your sufficiently interested reader?

(f) What is sometimes called the Pearson residual function (with this term invented many years after Karl's death) is the quantity

$$d(y) = g(y)/f(y, a_0, b_0) - 1,$$

where g is the ostensibly true data generating density and (a_0, b_0) the best fitting parameters of the model density (here the Gamma). Construct and display the estimated Pearson residual curve

$$\widehat{d}(y) = \widehat{g}(y)/f(y, \widehat{a}_{\text{ml}}, \widehat{b}_{\text{ml}}) - 1$$

for the Pearson dataset, where $\widehat{g}(y)$ is the nonparametric density estimator constructed by first using a traditional kernel density estimator for the $z_i = \log y_i$ data and then mapping back. (This is an easy and effective fix to overcome the 'problem of the boundary', in the present case.) Also add the nonparametric $\widehat{g}(y)$ curve to the two parametric gamma fits and the data histogram, and comment.

Exercise 3

This exercise concerns the ability of a few minimum discrepancy or disparity estimation methods to perform well under parametric model conditions. Consider the mixture model

$$f(y, \theta) = (1 - \theta) f_1(y) + \theta f_2(y), \quad \text{for } \theta \in [0, 1],$$

where we in this exercise for simplicity take f_1 and f_2 known, as respectively $\text{N}(0, 1)$ and $\text{N}(\delta, 1)$, with $\delta = 1.00$. Different estimation strategies may now be examined, not merely for their ability to estimate θ, but also for getting the estimated standard deviation right.

(a) The candidate is asked to put up and report on an appropriate simulation experiment to address some of these issues, say for $n = 50$ and $\theta = 0.666$. For each estimation strategy, say $\widehat{\theta}$, with associated estimator $\widehat{\kappa}$ for the quantity $\sqrt{n}\,\mathrm{sd}(\widehat{\theta})$,

(i) find the bias and standard deviation of $\widehat{\theta}$;

(ii) find the mean of $\widehat{\kappa}$ and compare it to its theoretical limit κ_0;

(iii) find the bias and standard deviation of $\widehat{\kappa}$.

Carry out such simulations, to address these questions, for (1) the maximum likelihood method; (2) the minimum Hellinger distance method; and (3) the Basu–Lindsay version of the minimum Hellinger distance method, involving the kernel smoothed model density (and use the normal kernel). What are your (tentative) findings?

(b) For the Basu–Lindsay method, $\sqrt{n}(\widehat{\theta}_h - \theta_0)$ tends to a $\mathrm{N}(0, \kappa(h)^2)$, where h is the bandwidth used (held fixed, so far). For the f_1 and f_2 above, and for $\theta_0 = 0.666$, compute and display $\kappa(h)$, for a suitable window $[0, h_0]$ of h values. Discuss.

(c) Above we have worked under model conditions, with f_1 and f_2 the densities of $\mathrm{N}(0,1)$ and $\mathrm{N}(\delta,1)$, and with $\delta = 1.00$. Assume now that the real data generating distribution is of the form $f = (1-\theta_0)f_1 + \theta_0 f_2$, again with $\theta_0 = 0.666$, with f_1 as above, but that the f_2 density has not been correctly specified, and that it in fact is the density of $\mathrm{N}(\delta, \tau^2)$, with some τ not equal to 1. Explain what the three estimation methods are aiming at, exemplify, and discuss.

Exercise 4

The aim of this exercise is to look into the possibility of creating BHHJ type methods for time series models. For simplicity let us confine ourselves to models with memory length equal to one step. Let $f_\theta(y_i \,|\, y_{i-1})$ be a parametric model for Y_i given $Y_{i-1} = y_{i-1}$, perhaps an approximation to the real $g(y_i \,|\, y_{i-1})$. Consider then

$$d_\alpha(\text{truth}, \text{model}_\theta) = \frac{1}{n-1}\sum_{i=2}^{n} \int \big\{ f_\theta(y_i \,|\, y_{i-1})^{1+\alpha} - (1+1/\alpha) g(y_i \,|\, y_{i-1}) f_\theta(y_i \,|\, y_{i-1})^\alpha + (1/\alpha) g(y_i \,|\, y_{i-1})^{1+\alpha} \big\}\, \mathrm{d}y_i.$$

(a) Show that this is a divergence.

(b) Argue that choosing $\widehat{\theta}$ to minimise

$$H_n(\theta) = \frac{1}{n-1}\sum_{i=2}^{n} \int \{ f_\theta(y_i \,|\, y_{i-1})^{1+\alpha}\, \mathrm{d}y_i - (1+1/\alpha) \frac{1}{n-1}\sum_{i=2}^{n} f_\theta(y_i \,|\, y_{i-1})^\alpha$$

is aiming at the θ_0 parameter minimising the above divergence function, under some conditions.

– As a rather simple example, for exploring how this scheme might pan out, consider the first-order autocorrelation time series model with known mean and variance, taken here to be zero and one, with

$$y_i \,|\, (y_1, \ldots, y_{i-1}) \sim \mathrm{N}(\rho y_{i-1}, 1-\rho^2) \quad \text{for } i = 2, \ldots, n.$$

With initial start $y_1 \sim \mathrm{N}(0,1)$, note that this corresponds to pairwise models

$$\begin{pmatrix} y_i \\ y_{i-1} \end{pmatrix} \sim \mathrm{N}_2(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}).$$

(c) Simulate datasets of this type, of length say $n = 100$, with $\rho = 0.555$. Compute both the maximum likelihood estimate $\widehat{\rho}_{\rm ml}$ and the BHHJ estimate $\widehat{\rho}$, for simplicity with $\alpha = 1$ (corresponding to L_2), for many such datasets. Display the consequent results $(\widehat{\rho}_{\rm ml}, \widehat{\rho})$, and make so-called sensible guesses regarding the general behaviour involved.

(d) It would be worthwhile writing up a paper regarding BHHJ methods for time series. It is *not* the task of the candidate to do so (just now), but we'd like to see the candidate provide a short overview of what such a paper ought to include, if successful and publishable.

Exercise 5

Consider independent observations $Y_1, \ldots, Y_n$ from a symmetric density on the real line, where the task is to estimate the unknown symmetry point, say θ. Thus

$$Y_i = \theta + \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

where the ε_i stem from a density $f_0(x)$ symmetric around zero. A standard estimator is of course $\widetilde{\theta} = \bar{Y}$, the mean of the data, which may also be seen as the solution to $\sum_{i=1}^n (Y_i - \theta) = 0$. That method is famously rather unrobust. As a more robust alternative, study $\widehat{\theta}$, the solution to

$$\sum_{i=1}^n \arctan(Y_i - \theta) = 0.$$

(a) Show that the $\widehat{\theta}$ estimator exists and is unique. Indicate why this method may be expected to be more robust than $\widetilde{\theta}$.

(b) Show that there is limiting normality,

$$\sqrt{n}(\widehat{\theta} - \theta) \to_d \mathrm{N}(0, \kappa^2),$$

and identify the κ. In particular, in the standard case where the observations are from the $\mathrm{N}(\theta, 1)$ model, how much is lost in efficiency by using the arctan estimator? And what is the limiting correlation between $\bar{Y}$ and $\widehat{\theta}$?

(c) One may 'work backwards' from the estimating equation $\sum_{i=1}^n (Y_i - \theta) = 0$ and deduce that the estimator in question is the maximum likelihood estimator for the normal model. Attempt similarly to identify a model $f(y, \theta) = f_0(y - \theta)$ such that the arctan estimator is its maximum likelihood estimator. Explain that $\widehat{\theta}$ is a minimum divergence estimator, and exhibit the divergence in question.

(d) Generalise the setup to include also a scale parameter.