**Norwegian University of Science and Technology**
**Department of Mathematical Sciences**



English

Contact during exam:
Håkon Tjelmeland 73 59 35 38

# EXAM IN TMA4300 MODERN STATISTICAL METHODS

Friday May 30th 2008
Time: 09:00–13:00

Aids: Calculator HP30S.
Statistiske tabeller og formler, Tapir forlag.
K. Rottman: Matematisk formelsamling.
One yellow paper (A4 with stamp) with own formulas and notes.

Grading: June 20th 2008.

## Problem 1

Consider a stochastic variable $X$ with density function

$$f(x) = \begin{cases} 2\alpha x e^{-\alpha x^2} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

**a)** Develop how to generate samples of $X$ by the inverse F method.

**b)** Develop how to generate samples of $X$ by rejection sampling. As proposal distribution use the exponential distribution with mean $1/\lambda$. [As usual, you can assume you have available a random number generator that generates samples from the proposal distribution.]

Explain how an optimal value for $\lambda$ can be found. (You do not need to do the calculations.)

Assume we are interested in the parameter $\theta = \mathrm{E}(\ln X)$. Clearly an unbiased estimator for this is

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \ln X_i,$$

where $X_1, \ldots, X_n$ are samples from $f(x)$ generated by the algorithm in item **a**) or **b**).

**c)** Use a variance reduction technique (control variables or antithetic variates) to define an alternative unbiased estimator $\widetilde{\theta}$ for $\theta$ with smaller variance than $\widehat{\theta}$. If important for the properties of your $\widetilde{\theta}$, specify which of the two algorithms you are using to generate samples from $f(x)$.

## Problem 2

In this problem we will consider a Bayesian model that is a (very) simplified version of a model that has been used to model seismic data.

Let $x = (x_1, \ldots, x_n)^T$ where $x_i \in \{1, \ldots, K\}$ for a given integer $K$. The $x_i$ represents the rock type in a location (node) $i$ along a vertical trace going through the underground. As a prior distribution for $x$ we use a Markov chain,

$$\pi(x) \propto p(x_1) \prod_{i=2}^{n} p(x_i|x_{i-1}),$$

where $p(x_1)$ is the marginal distribution for $x_1$ and $p(x_i|x_{i-1})$ is the transition probability for going from rock type $x_{i-1}$ to rock type $x_i$. We will assume both $p(x_1)$ and $p(x_i|x_{i-1})$ to be known functions.

Available data is $y = (y_2, \ldots, y_n)^T$ where $y_2, \ldots, y_n$ are assumed to be conditionally independent given $x$ with

$$y_i|x \sim \mathrm{N}\left(v(x_i) - v(x_{i-1}), \sigma^2(x_i)\right),$$

where $v(1), \ldots, v(K)$ and $\sigma^2(2), \ldots, \sigma^2(K)$ describe (known) physical properties of the different rock types. Thus, the (conditional) mean of $y_i$ is related to the difference in $v(\cdot)$ between the two locations $i$ and $i-1$. Moreover, the (conditional) variance of $y_i$ is different in the different rock types.

**a)** Visualise the Bayesian model specified above as a graphical model and write an expression for the posterior distribution given $y$.

Define a Markov chain Monte Carlo (MCMC) algorithm that simulates from the posterior distribution given $y$, i.e. specify what proposal distribution you will use and find formula for the corresponding acceptance probability. The acceptance probability expression should be simplified as much as possible. Summarise your chosen algorithm in pseudo code.

Assume you have run the MCMC algorithm for $M$ iterations and denote the generated states by $\{x^m\}_{m=0}^M$, where $x^m = (x_1^m, \ldots, x_n^m)$ is the generated state after $m$ iterations. In particular $x^0$ is the initial state.

It is of interest to use the MCMC output to estimate the following three quantities.

1. The (posterior) probability for rock type $k$ in node $i$, i.e.
$$\alpha_{ik} = \mathrm{P}(x_i = k|y).$$
2. The (posterior) expected fraction of nodes with rock type $k$.
3. The (posterior) probability for the fraction of nodes with rock type $k$ to be larger than a given value $r$.

**b)** Specify how one can estimate each of the three quantities above from the output of the MCMC algorithm.

## Problem 3

Assume we are in a regression problem setting with a data set $(x_i, y_i), i = 1, \ldots, n$, where $x_i$ is a vector of covariates corresponding to a (scalar) response $y_i$. Consider the regression model

$$y_i = m(\beta, x_i) + \varepsilon_i,$$

where $m(\cdot, \cdot)$ is a given (known) regression function, $\beta$ is a vector of (unknown) parameters, and $\varepsilon_1, \ldots, \varepsilon_n$ are assumed to be independent and identically distributed from an (unknown) distribution. Assume $\beta$ is estimated by the least squares estimator, so that

$$\widehat{\beta} = \underset{b}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - m(b, x_i))^2 \right\},$$

and that we are able to compute the value of $\widehat{\beta}$ by a numerical minimisation algorithm.

In the following we will consider both $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ as stochastic variables. Let $x_0$ be a new covariate vector from the same distribution as $x_1, \ldots, x_n$. The estimated $\widehat{\beta}$ can then be used to predict the response $y_0$ for $x = x_0$ as

$$\widehat{y}_0(x_0) = m(\widehat{\beta}, x_0).$$

In this problem we will consider how to estimate the resulting prediction error

$$\theta = \mathrm{E}\left( (y_0 - \widehat{y}_0(x_0))^2 \middle| (x_1, y_1), \ldots, (x_n, y_n) \right),$$

where the expectation is with respect to both $x_0$ and $y_0$.

The apparent prediction error is defined as

$$\widehat{\theta}_a = \frac{1}{n} \sum_{i=1}^{n} \Big( y_i - m(\widehat{\beta}, x_i) \Big)^2.$$

**a)** Do you think the apparent prediction error is (approximately) unbiased, or is it too optimistic or too pessimistic as an estimate of $\theta$? Give reasons for your answer.

Using the cross validation idea, propose an improved estimator for $\theta$. Introduce necessary notation so that you are able to describe the improved estimator clearly.

In the following we will consider how bootstrapping can be used to improve $\widehat{\theta}_a$.

**b)** Define the ideal bootstrap estimator for the bias of $\widehat{\theta}_a$. Introduce necessary notation to give a precise definition and specify in particular the distribution of the bootstrap samples.

Explain why it is not pratical to evaluate the ideal bootstrap estimator except for small values of $n$.

Write pseudo code for estimating the ideal bootstrap estimator for the bias of $\widehat{\theta}$ by stochastic simulation.

**c)** Using the estimated bias from the previous item, define a bias corrected version of the apparent prediction error, $\widehat{\theta}_a$.

Discuss the properties of $\widehat{\theta}_a$ relative to the properties of the bias corrected version. Which of these two estimators do you prefer? Give reasons for your answer.