Department of Mathematical Sciences

Examination paper for TMA4300 Computationally Intensive Statistical Methods

Examination date: June 8, 2023

Examination time (from-to): 15:00-19:00

Permitted examination support material: C:

- One specific basic calculator with empty memory
- Tabeller og formler i statistikk, Tapir forlag
- K. Rottman: Matematisk formelsamling
- A bilingual dictionary
- One yellow, stamped A5 sheet with personal handwritten formulas and notes (on both sides).

Academic contact during examination: John Paige
Phone: 412 25 382

Academic contact present at the exam location: No

OTHER INFORMATION

Get an overview of the question set before you start answering the questions.

All answers must be justified unless otherwise specified, and all necessary calculations must be included.

Read all questions carefully, and make your own assumptions. If a question is unclear/vague, make your own assumptions and specify them in your answer. Only contact academic contact in case of errors or insufficiencies in the question set. Address an invigilator if you wish to contact the academic contact. Write down the question in advance.

InsperaScan: All questions are meant to be answered on handwritten sheets. At the bottom of the question you will find a seven-digit code. Fill in this code in the top left corner of the sheets you wish to submit. We recommend that you do this during the exam. If you require access to the codes after the examination time ends, click “Show submission”.

Weighting: All 10 subproblems are equally weighted. It is recommended to spend an average of approximately 21 minutes per subproblems, which would leave 30 minutes to spare.

Notifications: If there is a need to send a message to the candidates during the exam (e.g. if there is an error in the question set), this will be done by sending a notification in Inspera. A dialogue box will appear. You can re-read the notification by clicking the bell icon in the top right-hand corner of the screen.

Withdrawing from the exam: If you become ill or wish to submit a blank test/withdraw from the exam for another reason, go to the menu in the top right-hand corner and click “Submit blank”. This cannot be undone, even if the test is still open.

Access to your answers: After the exam, you can find your answers in the archive in Inspera. Be aware that it may take a working day until any hand-written material is available in the archive.

Problem 1

Assume you can only generate Unif[0,1] random variables unless otherwise specified in this problem. Assume $\boldsymbol{X} \sim \text{MVN}(\boldsymbol{\mu}, \Sigma)$ is multivariate Gaussian with p dimensional mean vector $\boldsymbol{\mu} \in \mathbb{R}^p$ and positive definite covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$ with Cholesky decomposition $\Sigma = LL^T$. Further let $u \sim \chi^2_\nu$ be independent of $\boldsymbol{X}$ for integer-valued $\nu \geq 1$.

a) Write pseudocode (not necessarily R code) to generate p independent and identically distributed samples from a standard Gaussian distribution. You may assume p is even. The function you write should be called `stdGauss(p)` and use one input, p, and it should return a p-dimensional vector, $\boldsymbol{z}$.

b)

- Use `stdGauss(p)` to write pseudocode (not necessarily R code) for a function to generate N independent and identically distributed samples of $\boldsymbol{X}$. The function should be called `mvn(N, mu, L)` with inputs `N` being N, `mu` being $\boldsymbol{\mu}$, and `L` being L, and it should return a $p \times N$ matrix `xmat` where each column of `xmat` is a p-dimensional vector that is a sample from the distribution of $\boldsymbol{X}$.
- Note that $(\boldsymbol{X} - \boldsymbol{\mu})\sqrt{\nu/u} + \boldsymbol{\mu}$ is multivariate-t distributed with parameters μ, Σ, and ν. Use `mvn(N, mu, L)` and `stdGauss(p)` to write pseudocode (not necessarily R code) for a function `mvt(N, mu, L, nu)` generating N draws from a multivariate-t distribution with ν degrees of freedom, assuming $\nu \geq 1$ is an integer. The inputs `N`, `mu`, and `L` are as before, and `nu` corresponds to ν. The function should output a $p \times N$ matrix `tmat` with N columns corresponding to independent draws from the multivariate-t distribution.

Problem 2

A statistician is designing a survey with N sampled individuals divided among p distinct regions in Norway. Given that the proportion of Norway's population living in region i is π_i, and letting $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)$, the Statistician chooses to sample $\boldsymbol{X} = (X_1, \ldots, X_p)$ individuals from the p regions where $\boldsymbol{X} \sim \text{Multinomial}(N, \boldsymbol{\pi})$, and X_i is the number of individuals sampled from region i for $i = 1, \ldots, p$. Note that no pre-existing functions for drawing random values may be used in this problem except for those that draw from a standard uniform distribution, such as `runif` from `R`. We will assume throughout this problem that $N \geq p$.

a) The statistician decides to make sure that there is at least 1 sample in each region. Hence, she chooses to sample from the distribution of $\boldsymbol{X} \mid X_1 \geq 1, \ldots, X_p \geq 1$. In this subproblem we let $\boldsymbol{Y} \sim \text{Multinomial}(N - p, \boldsymbol{\pi})$ and build a rejection sampler by drawing proposals from the distribution of $\mathbf{1}+\boldsymbol{Y}$.

Consider the proposal $\boldsymbol{z} \equiv \boldsymbol{y} + \mathbf{1}$, where $\boldsymbol{y}$ is a draw from the distribution of $\boldsymbol{Y}$ and '$\equiv$' means 'is defined as'. Mathematically prove that, for the set $A = \{\boldsymbol{z} : z_i \in \{1, \ldots, N\}\ \forall i, \sum_i z_i = N, \}$, the acceptance probability of $\boldsymbol{z}$ is:

$$P(\text{accept } \boldsymbol{z}) = \frac{N - p + 1}{\prod_{i=1}^{p} z_i}, \quad \text{for } \boldsymbol{z} \in A.$$

(Hint: Let $\kappa \equiv P(X_1 \geq 1, \ldots, X_p \geq 1)$ *and show:*

$$C \equiv \sup_{\boldsymbol{z} \in A} \frac{P(\boldsymbol{X} = \boldsymbol{z} | X_1 \geq 1, \ldots, X_p \geq 1)}{P(\boldsymbol{Y} = \boldsymbol{z} - \mathbf{1})} = \frac{1}{\kappa} \sup_{\boldsymbol{z} \in A} \frac{P(\boldsymbol{X} = \boldsymbol{z})}{P(\boldsymbol{Y} = \boldsymbol{z} - \mathbf{1})} \equiv \frac{1}{\kappa} C^*.$$

You may assume $\inf_{\boldsymbol{z} \in A} \prod_{i=1}^{p} z_i = N - p + 1$. *Note that the pmf of the multinomial* X *with parameters* N *and* $\boldsymbol{\pi}$ *is*:

$$P(\boldsymbol{X} = \boldsymbol{x}) = \frac{N!}{x_1! \ldots x_p!} \pi_1^{x_1} \ldots \pi_p^{x_p}$$

for $x_1, \ldots x_p \in \mathbb{N}$ *and* $\sum_i x_i = N$*)*

b)

- If a rejection sampler were implemented using the above proposal distribution, what would the expected proportion of proposals be that were accepted if $\pi_i = \frac{1}{p}\ \forall\ i$? Derive the answer mathematically, and simplify as much as possible.
- How well computationally would this sampler perform in practice when N is much larger than p versus when N is close to p? Why?

Problem 3

In this problem, we are interested in counting cars entering m different intersections throughout n different days. Let Y_{ij} be the number of cars that visited intersection i and day j, and assume each intersection i has its own daily rate of cars traveling through it, $\lambda_i > 0$. We will use the following Bayesian hierarchical model:

$$\begin{aligned} Y_{ij} \mid \lambda_i &\sim \text{Poisson}(\lambda_i), \quad i = 1, \ldots, m, \quad j = 1, \ldots, n \\ \lambda_i \mid \alpha, \beta &\sim \text{Gamma}(\alpha, \beta), \quad i = 1, \ldots, m \\ \alpha &\sim \text{Gamma}(A, B) \\ \beta &\sim \text{Gamma}(C, D) \end{aligned}$$

for known fixed values $A, B, C, D > 0$. We assume that $Y_{ij} \perp Y_{kl} \mid \lambda_i, \lambda_k$ for all i, j, k and l, and that $\lambda_i \perp \lambda_k \mid \alpha, \beta$ for $i \neq k$. Recall the Poisson pmf is,

$$P(Y_{ij} = y \mid \lambda_i) = \frac{\lambda_i^y e^{-\lambda_i}}{y!},$$

for $y \in \{0, 1, 2, \ldots\}$, and the Gamma pdf is,

$$p(\lambda_i \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_i^{\alpha-1} e^{-\beta\lambda_i},$$

for $\lambda_i, \alpha, \beta > 0$. Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in})$ for all i and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$.

a) Derive an expression proportional to the posterior density $p(\boldsymbol{\lambda}, \alpha, \beta \mid \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m)$. Further derive the full conditional for λ_i. If it is a named distribution, then provide the name and parameters.

b)

- Assume we have a Metropolis step for $\binom{\alpha}{\beta}$ with proposal distribution

$$\binom{\alpha'}{\beta'} \sim \mathcal{N}\left(\binom{\alpha}{\beta}, \sigma^2 \boldsymbol{I}_2\right)$$

  for 2×2 identity matrix $\boldsymbol{I}_2$ and variance tuning parameter σ^2. Calculate the acceptance probability for the proposal $\binom{\alpha'}{\beta'}$.
- In practice, how should one calculate Metropolis-Hastings acceptance probabilities on a computer so as to avoid potential problems?

c)

- Assume a hybrid Metropolis-within-Gibbs sampler of the above posterior is implemented using Gibbs steps for the λ_i and the above Metropolis step proposal for (α, β). Is it possible to know if the above sampler converged? If so, how, and if not, why not?
- Describe how you could select the tuning parameter σ^2 in the above sampler. Which visual and/or mathematical criteria could you use?

Problem 4

a) Which of the following models can be fit using INLA? Choose all that apply, no explanations necessary. Here, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$ and the responses are $Y_1, \ldots, Y_n$.

1. For fixed known $N_1, \ldots, N_n$,

$$\begin{aligned} Y_i \mid \boldsymbol{\eta} &\overset{iid}{\sim} \text{Binomial}(N_i, \text{expit}(\eta_i)), \quad \forall\, i \\ \eta_i &= \alpha + \boldsymbol{x_i}^T \boldsymbol{\beta} + \epsilon_i, \quad \forall\, i \\ \alpha &\sim N(0, \infty) \\ \beta_j &\sim N(0, 100^2), \quad \forall\, j \\ \epsilon_i \mid \sigma_\epsilon^2 &\sim N(0, \sigma_\epsilon^2) \\ \sigma_\epsilon^2 &\sim \text{Gamma}(5, 5) \end{aligned}$$

2.

$$\begin{aligned} Y_i \mid \boldsymbol{\eta}, \sigma^2 &\overset{iid}{\sim} \mathcal{N}(\eta_i, \sigma^2), \quad \forall\, i \\ \eta_i &= \boldsymbol{x_i}^T \boldsymbol{\beta} + \epsilon_i, \quad \forall\, i \\ \epsilon_i &\sim \text{Exp}(5) \\ \beta_j &\sim N(0, 100^2), \quad \forall\, j \end{aligned}$$

3. For known, sparse positive definite matrix $\boldsymbol{Q}$,

$$\begin{aligned} Y_i \mid \boldsymbol{\eta} &\overset{iid}{\sim} \text{Poisson}(e^{\eta_i}), \quad \forall\, i \\ \eta_i &= \alpha + \boldsymbol{x_i}^T \boldsymbol{\beta} + \boldsymbol{z_i}^T \boldsymbol{\gamma} + \epsilon_i, \quad \forall\, i \\ \alpha &\sim N(0, \infty) \\ \beta_j &\sim N(0, 100^2), \quad \forall\, j \\ \boldsymbol{\gamma} &\sim MVN(\mathbf{0}, \boldsymbol{Q}^{-1}) \\ \epsilon_i \mid \sigma_\epsilon^2 &\sim N(0, \sigma_\epsilon^2) \\ \sigma_\epsilon^2 &\sim \text{Gamma}(5, 5) \end{aligned}$$

4. For unknown random vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$:

$$\begin{aligned} Y_i \mid \boldsymbol{\eta}, \sigma^2 &\overset{iid}{\sim} \mathcal{N}(\eta_i, \sigma^2), \quad \forall\, i \\ \eta_i &= \boldsymbol{x_i}^T \boldsymbol{\beta} + \gamma_i, \quad \forall\, i \\ (\boldsymbol{\gamma} \mid \tau_\gamma) &\text{ is RW(1) with time index } i \text{ and precision parameter } \tau_\gamma \\ \beta_j &\sim N(0, 100^2), \quad \forall\, j \\ \tau_\gamma &\sim \text{InvGamma}(2, 10) \end{aligned}$$

Problem 5

Assume we have the observed response random vector $\boldsymbol{z} = (z_1, \ldots, z_n)$. In addition, we can define another sequence of unobserved random variables, $\boldsymbol{u} = (u_1, \ldots, u_n)$, where the z_i's are independent conditional on the u_i's (i.e. $z_i \perp z_j \mid u_i$ for all i and j), z_i has conditional density

$$f(z_i \mid u_i) = u_i \cdot e^{-z_i} + (1 - u_i) \cdot 3e^{-3z_i},$$

and $u_i \overset{iid}{\sim}$ Bernoulli(p) so that

$$f(u_i) = \begin{cases} p, & \text{if } u_i = 1 \\ 1 - p, & \text{if } u_i = 0 \\ 0, & \text{otherwise} \end{cases}$$

for unknown but fixed $0 \leq p \leq 1$. In this problem we will use expectation maximization to estimate p.

a)

- Assuming we have observed both $\boldsymbol{z}$ and $\boldsymbol{u}$, derive an expression for the full log likelihood, $\ell(p; \boldsymbol{z}, \boldsymbol{u}) = \log(f(\boldsymbol{z}, \boldsymbol{u} \mid p))$.
- Let $\delta_{p^{(t)}}(z_i) = E[u_i \mid z_i, p^{(t)}]$ for each i. For

  $$Q(p) = Q(p \mid p^{(t)}) = E[\ell(p) \mid \boldsymbol{z}, p^{(t)}],$$

  show that:

  $$\begin{aligned} Q(p \mid p^{(t)}) = &\sum_{i=1}^{n} \Big(\delta_{p^{(t)}}(z_i) \cdot (-z_i) + (1 - \delta_{p^{(t)}}(z_i)) \cdot (\log(3) - 3z_i) \Big) \\ &+ \log(p) \sum_{i=1}^{n} \delta_{p^{(t)}}(z_i) + \log(1 - p) \sum_{i=1}^{n} (1 - \delta_{p^{(t)}}(z_i)). \end{aligned}$$

b)

- Maximize the above expression for $Q(p \mid p^{(t)})$ with respect to p to obtain $p^{(t+1)}$ in terms of $\delta_{p^{(t)}}(z_i)$.
- Assuming you use the E and M steps derived above to estimate p as $\hat{p}$ via the EM algorithm, describe in detail how you could estimate the standard error of $\hat{p}$.