Homework 3

Problems 2 – 4 contain various applications of the Bootstrap filter. Read the questions carefully and formulate them in mathematical terms. Discuss with me if necessary.

Problem 1: Total variation and KL distance. Suppose P and Q are probability measures on the same probability space $(\mathcal{X}, \mathcal{B})$ with densities $dP/d\mu = p$ and $dQ/d\mu = q$ w.r.t. some dominating measure μ. Recall the total variation distance $||P - Q||_1 := \sup_{B\in\mathcal{B}} |P(B) - Q(B)|$ (the supremum is over all Borel sets; I may have defined it in the notes with a 2 before the supremum) and the Kullback-Leibler divergence $\mathrm{KL}(P||Q) = \int p \log(p/q) d\mu$.

(i) Let $\{p > q\} = \{x \in \mathcal{X} : p(x) > q(x)\}$. Show that (a picture may help)

$$||P - Q||_1 = \int_{p>q} (p - q)d\mu = \int_{q>p} (q - p)d\mu = \frac{1}{2}\int |p - q|d\mu.$$

[**Hint:** Show that the supremum is attained at the set $B^* = \{p > q\}$, or equivalently, at $(B^*)^c$]

(ii) Prove Pinsker's inequality $||P - Q||_1^2 \le \frac{1}{2}\mathrm{KL}(P||Q)$.

[**Hint.** Write $\Delta = p/q - 1$, so that $p = (1 + \Delta)q$ and $\int \Delta q d\mu = 0$. Use the inequality $(1 + x)\log(1 + x) - x \ge x^2/\{2(1 + x/3)\}$. Finally, use a Cauchy–Schwartz inequality.]

Problem 2. I have observed 20 independent samples from a $N(|\theta|, 1)$ distribution with sample mean 1.65. Assume a N(0,1) prior on θ. Run a Bootstrap filter with $T = 10^4$ samples from the prior distribution. Obtain 5000 independent samples by resampling from the obtained discrete approximation to the posterior. Use these samples to draw a histogram of the (approximated) posterior distribution of θ.
If you are using `R`, use `set.seed(729)` at the beginning of your code.

Problem 3. Generate 200 independent observations from a Beta(2, 1) density. In `R`, you will set `x = rbeta(400,2,1)`. Use `set.seed(1857)` at the beginning of your code.
Suppose we want to fit a mixture density $(1 - \nu)\mathrm{Beta}(3/2, 1) + \nu\mathrm{Beta}(5/2, 1)$ to this data. Assume a $U(0, 1)$ prior on ν. Run a Bootstrap filter with $T = 10^4$ samples from the prior distribution. Obtain 5000 independent samples by resampling from the obtained discrete approximation to the posterior. Use these samples to:
(i) Draw a histogram of the (approximated) posterior distribution on ν. Report its posterior mean and standard deviation.
(ii) Plot the *posterior mean* of the estimated density on the grid `seq(0,1,by = 0.01)`. Overlay with the true density.

Problem 4. Consider the dde and preterm birth data (posted in `datasets/dde.rtf`). The data is from the Longnecker et al. (2001, Lancet) study relating maternal serum concentration of the DDE metabolite DDE to the risk of preterm birth. [Variable key: x=dde dose, y=indicator preterm birth, z1-z5 = potential "confounders"] Consider fitting a Bayesian logistic regression

$$\text{logit } P(y_i = 1 \mid x_i) = \exp(x_i'\beta),$$

with $y_i = 1/0$ if preterm/term birth and $x_i = (1, dde_i)$. The dde values in the text-file are in the raw scale which you should standardize before analysis. Henceforth, when I mention dde, I mean the standardized scale.

Assume a $\mathrm{N}(0, 9\mathrm{I}_2)$ prior on $\beta = (\beta_1, \beta_2)'$. Run a Bootstrap filter with $T = 10^5$ draws
(i) from the prior,
(ii) from a bivariate-t importance density $t_\nu(\mu, \Sigma)$ with $\nu = 3$, $\mu = \hat{\beta}_{mle}$ the m.l.e. of β and

$\Sigma = \widehat{Cov}(\hat{\beta}_{mle})$ the approximate covariance matrix of the mle (you can obtain μ and Σ from the output of the `R` function `glm`. Your lecture notes contain the expression for the density of a multivariate t density; alternatively, you can use the `R` function `dmvt`).

Compare the two approaches (i) and (ii) and report any differences you see.

Define the dose-response curve as

$$f(z) = P(y = 1 \mid \text{dde} = z) = \frac{\exp(x'\beta)}{1 + \exp(x'\beta)}, \quad x = (1, z)'.$$

Suppose we want to estimate the function f (on a grid between -2 and 2 with equal spacing of 0.01) with uncertainty characterization. *Plot the posterior mean and point-wise 95 % credible intervals for f on the above grid using the output from (ii). Describe how you obtain these quantities from the output of (ii).*

—

Please label each page of your homework clearly with your name IN BLOCK CAPITALS. If you use more than one sheet of paper, please staple the sheets together. This homework will be due in class on **Tuesday, March 27**.