Statistics for Data Science – Take Home Exam (due February 5, 11:59pm)

There is a total of 30 points to achieve on this take home exam. Please upload your solutions in typed or readable (!) handwritten form (scans, photographs) including your code as usual on Moodle. Good luck!

Exercise 1 (18P). *Consider the Gaussian linear model $Y \sim N(X\beta, I_n)$, where $\beta \in \mathbb{R}^d$ is the unknown parameter vector and $X \in \mathbb{R}^{n\times d}$ is a known design matrix. We put a product Laplace prior on β, that is, we consider*

$$\pi_\lambda(\beta) \propto \prod_{j=1}^{p} \exp(-\lambda|\beta_j|),$$

for a tuning parameter $\lambda > 0$.

a) ***(4P)*** *Show that the posterior mode (= maximizer of the posterior density $\beta \mapsto p(\beta|Y)$) is equal to the LASSO-estimator of β.*

b) ***(5P)*** *Consider the US crime data set but use only the variables "PctKids2Par", "NumStreet" and "racePctWhite" as well as an intercept column. Implement an appropriate MCMC algorithm that allows you to sample from the $d = 4$ dimensional posterior density $p(\beta|Y) \propto p(Y|\beta)\pi_\lambda(\beta)$.*

c) ***(5P)*** *For each parameter β_j, $j = 1, \dots, 4$, (numerically) compute a (one-dimensional) 95% Bayesian credible interval. How do these intervals change as λ changes? Provide some visualization.*

d) ***(4P)*** *Visualize the two-dimensional (marginal) posterior density of "PctKids2Par" and "NumStreet" for a few different values of λ.*

Exercise 2 (12P). *Design a pivotal bootstrap type resampling scheme for constructing confidence intervals for population graph totals in statistical network analysis.*

a) ***(8P)*** *Implement it on the Facebook data set from the lecture to obtain a bootstrap confidence interval for the population proportion of TV-politics edges and* ***plot the bootstrap distribution of your estimator*** *for just a handful of different random samples. Pick any graph sampling design you consider reasonable. Proceed as in the lecture, assuming that you don't have access to neither E nor N_E, but that you can only retrieve edge information once you have sampled one of its incident vertices.*

b) ***(4P)*** *Verify the validity of your method by running a Montecarlo simulation, recalling that you know the true population proportion since the population graph is given.*

Hint: Don't worry too much about exact statistical validity but rather try to mimic the general bootstrap idea. You need to be a bit creative and flexible when you adapt the bootstrap idea for iid data to the network setting. There isn't just one correct solution.