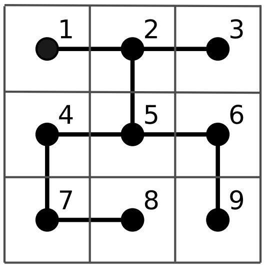
MA 589 — Computational Statistics

Final Project Suggestions

(Final Project Due: Thursday, 12/8/16)

1. In Project 5 we had an Ising model defined in the full image lattice L. It is hard, in general, to compute on the lattice, so we often resort to approximations based on a *spanning* tree of L; the figure below illustrates one such tree for the 3-by-3 regular lattice in Project 5.



The goal of this project is similar in that we seek a spanning tree that can be used to conduct inference on the land cover labels X. We use the same data (Y) and likelihood ($\mathbb{P}(Y \mid X)$) as in Project 5, and parameterize the distribution on X with the tree T:

$$\mathbb{P}_T(X) = \frac{1}{Z_T} \exp\left\{ J \sum_{(i,j)\in T} X_i X_j \right\},$$

where $Z_T := \sum_X \exp\{J \sum_{(i,j)\in T} X_i X_j\}$ is the normalizing constant. We want $T^* = \arg\max_T \mathbb{P}_T(X \mid Y)$; to obtain an estimator for T^* we employ an EM algorithm with the labels X as latent variables.

(a) Show that the E-step defines

$$\begin{aligned} Q(T; T^{(t)}) &= \mathbb{E}_{X \mid Y; T^{(t)}}\big[\log \mathbb{P}_T(X)\big] \\ &= J \sum_{(i,j)\in L} \mathbb{E}_{X \mid Y; T^{(t)}}\big[X_i X_j\big] I[(i,j) \in T] - \log Z_T. \end{aligned}$$

(b) Assume, for simplicity, that $\log Z_T$ does not change much with T, and so that Q depends only on the term with the expectation. Derive a Gibbs sampler that has $\mathbb{P}_{T^{(t)}}(X \mid Y)$ as its target, and using the samples (after convergence), estimate $\mathbb{E}_{X \mid Y; T^{(t)}}\big[X_i X_j\big]$.[1]

(c) Now use a *maximum weight spanning tree* algorithm to perform the M-step, that is, to update $T^{(t+1)} = \arg\max_T Q(T; T^{(t)})$. Report your estimate $\widehat{T}^*$ of T^* after your EM algorithm converges.

[1] Actually, these expectations can be estimated from an *exact* sampler or, even better, computed exactly using a (slightly more complicated) forward algorithm that computes pair marginals.

(d) $*$ Finally, obtain a conditional estimate of land cover classes, $\widehat{X} = \arg\max_X \mathbb{P}_{\widehat{T}^*}(X \mid Y)$.

The goal of the next two project suggestions is to conduct a bootstrap study based on a linear model using the "Old Faithful" geyser dataset[2]. The dataset has $n = 272$ observations of eruption duration—the predictor X—and waiting time to next eruption—the response Y. The linear model has two coefficients,

$$Y_i = \beta_0 + \beta_1 X_i + e_i, \quad i = 1, \ldots, n$$

and the errors e_i at each observation are *independent*, with zero mean, $\mathbb{E}[e_i] = 0$, and constant variance, $\mathbb{V}ar[e_i] = \sigma^2$, as it is usually assumed in linear regression.

2. Let us assume that the errors are normally distributed, $e_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$, and check this assumption by estimating the density of the errors.

   (a) Let us start by fitting the linear model; report $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1)$ and $\widehat{\sigma}$ (the residual standard error)[3], plot residuals $\widehat{e}_i$ and Y versus the fitted values $\widehat{Y}$, where $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$.

   (b) Conduct three bootstrap simulations using $B = 1{,}000$ replications: bootstrap (i) residuals, (ii) data pairs, and (iii) perform a parametric bootstrap assuming that $e_i \overset{\text{iid}}{\sim} N(0, \widehat{\sigma}^2)$. Plot boxplots of $\widehat{\beta}(X^*, Y^*)$, that is, of estimates using the replicated data, for each simulation. List estimates for the standard errors of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ under each simulation; how do they compare to the least-squares estimates?

   (c) Now obtain 95% *percentile* confidence intervals for each simulation and report them. In addition, report a 95% "bias corrected and accelerated" (BCA) confidence interval for the residual bootstrap only. Draw lines in the previous boxplot figure with the endpoints of the BCA interval. Comment on the coverage of the interval.

   (d) Let us estimate the distribution of the residuals using a *normal* kernel with bandwidth h,

   $$\widehat{f}_h(e) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} \phi\Big(\frac{e - \widehat{e}_i}{h}\Big),$$

   where ϕ is the standard normal density. Start by plotting the cross-validation estimate of the integrated square error J (modulo a constant) to decide the best bandwidth:

   $$\widehat{J}(h) = \int \widehat{f}_h^2(z)dz - \frac{2}{n} \sum_{i=1}^{n} \widehat{f}_{h,-i}(\widehat{e}_i),$$

   where $\widehat{f}_{h,-i}$ is the density estimate based on the data without the i-th residual. Use the interval $h \in (1, 2)$ for the plot, and report the bandwidth h^* that minimizes $\widehat{J}(h)$.

[2] `data(faithful)` in R.

[3] You can get model fit statistics using `summary(lm(y ~ x))` in R.

(e) Plot the density estimate $\widehat{f}_{h^*}$ in the range $(-20, 20)$ and 95% "studentized" confidence *bands*. Now plot the normal density that was assumed for the parametric bootstrap simulation, $N(0, \widehat{\sigma}^2)$; does it fit inside the confidence bands? In light of this result, how would you comment on the similarities or differences between the parametric confidence intervals and the other non-parametric (residual and pair) confidence intervals?

3. The usual setup of a linear model assumes that errors follow a normal distribution with zero mean and constant variance. In this case, the least-squares estimators follow a normal distribution and confidence intervals and hypothesis tests for the coefficients are simpler to build and conduct. What if the errors have a heavier tail, say, if they follow a *Cauchy* distribution? The goal of this project is to assess the sampling distribution of least-squares estimators in this case.

   (a) If C follows a Cauchy distribution with scale σ, $C \sim \text{Cauchy}(\sigma)$, the pdf of C is given by
   $$f_C(c) = \frac{1}{\pi} \frac{\sigma}{c^2 + \sigma^2}.$$
   Derive an inverse CDF method to sample from $\text{Cauchy}(\sigma)$.[4]

   (b) Same as item (a) in Suggestion 2.

   (c) Same as item (b) in Suggestion 2, but for the parametric bootstrap simulation assume that $e_i \overset{\text{iid}}{\sim} \text{Cauchy}(\widehat{\sigma})$.

   (d) Same as item (c) in Suggestion 2.

   (e) Same as item (d) in Suggestion 2.

4. Implement Marsaglia's ziggurat method using the reference paper in Blackboard, with 20 rectangles. In your project, detail how you define the endpoints of the rectangles, present a detailed algorithm, and conduct a careful simulation study (based on this paper.)

[4] The Cauchy distribution is a Student t distribution with one degree of freedom, so you can use this fact to check if your samples are correct.