Homework 2 (due October 20, 2016)

Problem 1: [Theoretical problem] Compute the expected number of clusters induced by a Dirichlet Process on the observations $(X_1, \ldots, X_n)$ under the following hierarchical distribution: $X_i \mid P \sim P, P \sim \mathrm{DP}(\alpha G_0)$ and show that it is asymptotically of the order $\alpha \log n$ as $n \to \infty$.
Problem 2: [Numerical example]
(i) Download the Galaxy data from R using the `MASS` package. See the link below.
https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/galaxies.html
(ii) Obtain a kernel density estimate with the bandwidth chosen according to the unbiased cross validation (ucv). You may use the `kedd` package in R. See the link below.
https://cran.r-project.org/web/packages/kedd/kedd.pdf
(iii) Run the blocked Gibbs sampler to fit a Dirichlet process mixture (DPM) of normals to the data. Choose the hyperpriors and the hyperparameters carefully. Report the following a) pointwise posterior median of the density b) pointwise 95% credible intervals of the density and c) posterior histogram of the number of clusters.
(iv) Comment on the qualitative difference between the posterior median obtained in iii) and the kernel density estimate in ii).