Statistical Methods in Applied Computer Science, Fall 2020
Assignment 2 (actually 2, 3 and Project)

Jens Lagergren, Niharika Gauraha, Hazal Koptagel and Oskar Kviman

Deadline: see Canvas

This assignment should be done in pairs. Describe your results in a short report using at most 2 pages per problem. You will present the assignment by a written report, submitted before the deadline using Canvas. You must solve the assignment independently and it will automatically be checked for similarities to other students' solutions as well as documents on the web in general. From the report it should be clear what you have done and you need to support your claims with results. You are supposed to write down the answers to the specific questions detailed for each task. This report should clearly show how you have drawn your conclusions and explain your derivations. Your assumptions, if any, should be stated clearly.

Being able to communicate results and conclusions is a key attribute of any scientific practitioner. It is up to you as an author to make sure that the report clearly shows what you have done. Based on this, and only this, we will decide if you pass the task. No detective work should be required on our side. In particular, neat and tidy reports please! I very much recommend you to get used to LaTeX to write your report, it is a great tool.

Only *one* group member should submit the group work. Make sure to write both members' names in the report. Upload two files i) the report in PDF format ii) the codes in a compressed format e.g ZIP, TAR.

Some questions in this assignment requires you to use the datasets we generated. You can access to the datasets via this link: https://gits-15.sys.kth.se/koptagel/StatMeth20.

Grading

- **E** Correctly completed **both** Task 2.1 and 2.2
- **D** E + Correctly completed **one** of Task 2.3, 2.4.
- **C** E + Correctly completed **two** of Task 2.3, 2.4.
- **B** E + Correctly completed **two** of Task 2.3, 2.4 as well as **one** of 3.1 and 3.2
- **A** E + Correctly completed **two** of Task 2.3, 2.4 as well as **two** of 3.1 and 3.2.

These grades are valid for assignments submitted before the deadline, late assignments can at most receive the grade E.

Good Luck!

2.1 Gibbs sampler for the magic word

The following generative model generates N sequences of length M, $s^1, \ldots, s^N$ where $s^n = s_1^n, \ldots, s_M^n$. All sequences are over the alphabet $[K]$. Each of these sequences has a "*magic*" word of length W hidden in it and the rest of the sequence is called *background*.

First, for each sequence n, a start position r_n for the magic word is sampled uniformly from $[M - W + 1]$. Then the j:th positions in the magic words are sampled from $q_j(x) = \text{Cat}(x|\theta_j)$ where θ_j has a $\text{Dir}(\theta_j|\alpha)$ prior. All other positions in the sequences are sampled from the background distribution $q(x) = \text{Cat}(x|\theta)$ where θ has a $\text{Dir}(\theta|\alpha')$ prior.

We are interested in the posterior $p(r_1, \ldots, r_N|D)$ where D is a set of sequences $s^1, \ldots, s^N$ generated by the model and r_n is the start position of the magic word in the n:th sequence s^n. The following describes a Gibbs sampler that can be used for estimating the posterior over start positions after having observed $s^1, \ldots, s^N$. The sampler is collapsed and we do know the hyperparameters α and α'. The states are vectors of start positions $R = (r_1, \ldots, r_N)$.

Notice that, for the Dirchlet-Categorical distribution for the j:th position of the magic word, the marginal likelihood is

$$p(D_j|R) = \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(N + \sum_k \alpha_k)} \prod_{k=1}^{K} \frac{\Gamma(N_k^j + \alpha_k)}{\Gamma(\alpha_k)}$$

where N_k^j is the count of symbol k in the j:th column of the magic words, induced by R. For the background, the marginal likelihood is

$$p(D_B|R) = \frac{\Gamma(\sum_k \alpha'_k)}{\Gamma(B + \sum_k \alpha'_k)} \prod_{k=1}^{K} \frac{\Gamma(B_k + \alpha'_k)}{\Gamma(\alpha'_k)}$$

where B is the number of background positions (i.e., $N(M - W)$) and B_k is the count of symbol k in the background, induced by R. The full conditional $p(r_n|R_{-n}, D)$ can be expressed as follows

$$\begin{aligned}
p(r_n|R_{-n}, D) &= \frac{p(r_n \cup R_{-n}, D)}{p(R_{-n}, D)} \\
&= \frac{p(D|R)\, p(R)}{p(R_{-n}, D)} \\
&\propto p(D|R) \\
&= p(D_B|R) \prod_{j=1}^{W} p(D_j|R)
\end{aligned}$$

Implement a Gibbs sampler for the magic word model described above. Hint: Use log-scale to avoid numerical problems.

> **Question 1:** *Simulate synthetic data and estimate the posterior using your Gibbs sampler. Describe your experiment setting (parameters, number of iterations etc). Provide evidence that you have succeeded and report the accuracy.*

> **Question 2:** *Use the data provided in the GitHub repository. Describe your experiment setting. Provide evidence that you have succeeded. Report the list of start positions of first 10 sequences ($R_{1:10}$) you think is the correct start positions.*

2.2 SMC for the stochastic volatility model

Consider the stochastic volatility model

$$X_1 \sim \mathcal{N}(x_1|0, \sigma^2) \tag{1a}$$

$$X_t|(X_{t-1} = x_{t-1}) \sim \mathcal{N}(x_t|\phi x_{t-1}, \sigma^2), \qquad t = 1, \dots, T, \tag{1b}$$

$$Y_t|(X_t = x_t) \sim \mathcal{N}(y_t|0, \beta^2 \exp(x_t)), \qquad t = 1, \dots, T, \tag{1c}$$

where the parameter vector is given by $\theta = \{\phi, \sigma, \beta\}$ ($\phi \in \mathbb{R}$, $\sigma > 0$, $\beta > 0$). The latent variable X_t denotes the underlying volatility, i.e. the variations in the price of some financial asset, and Y_t denotes the observed scaled log-returns from the asset.

Question 3: *Simulate $x_{1:T}$ and $y_{1:T}$ from the SV model. Use $T = 100$ and fix $\phi = 0.91, \sigma = 0.16, \beta = 0.64$.*

Question 4: *Implement sequential importance sampling (SIS) to infer x_T. i) Briefly justify the choice of proposal, derive the weight update function, and a point estimate $\hat{x}_T$ of x_T. ii) Report point estimate of x_T and compute the mean squared error of the estimate to the truth as the number of samples is increased. iii) Compute the empirical variance of the normalized weights, $\boldsymbol{w}_T$. Provide density estimation plot or histogram of the weights.*

Question 5: *Implement Bootstrap Particle Filter (BPF) with multinomial resampling. Answer parts ii and iii from SIS question above.*

Question 6: *Implement BPF with stratified resampling. Answer parts ii and iii from SIS question above.*

Question 7: *Using 2-3 sentences, compare and contrast the three schemes.*

2.3 Stochastic volatility unknown parameters part I

Assume now that the variance parameters σ and β are both unknown, but keep $\phi = 0.91$.

Question 8: *Simulate $x_{1:T}$ and $y_{1:T}$ using $\beta = 0.64, \sigma = 0.16$. Make a reasonably coarse grid for β and σ between 0 and 2 (say, 8 by 8). Use the SMC with Multinomial resampling to estimate the marginal likelihood, $p_\theta(y_{1:T})$ for each pair of (β, σ) in the grid. Run SMC 10 times (each with different random seed) for each parameter combination. Choose the best parameter combination and compare it against the ground truth.*

For numerical reasons, it is better to consider the log-likelihood:

$$\log \hat{p}(y_{1:T}) = \log \prod_{t=1}^{T} \frac{1}{N} \sum_{n=1}^{N} \underbrace{p(y_t|x_t^i)}_{\tilde{w}_t^i = \alpha_t(x_{1:t}^i)} = \sum_{t=1}^{T} \left(\log \sum_{n=1}^{N} \tilde{w}_t^i - \log N \right), \tag{2}$$

where $\tilde{w}_t^i$ denotes the incremental weight of particle i at time step t.

Question 9: *Study how T and N affect the variance in the log-likelihood estimate. That is, fix T, vary N and plot the log-likelihood estimates and vice versa.*

Consider a Bayesian setting and place inverse Gamma priors on the variance parameters:

$$\sigma^2 \sim \mathcal{IG}(a = 0.01, b = 0.01), \tag{3a}$$
$$\beta^2 \sim \mathcal{IG}(a = 0.01, b = 0.01), \tag{3b}$$

where the the inverse Gamma PDF with parameters (a, b) is given by

$$\mathcal{IG}(x|a,b) = \frac{b^a}{\Gamma(a)} x^{-a-1} \exp\left(-\frac{b}{x}\right) \tag{4}$$

and Γ is the Gamma function.

Question 10: *Implement Particle Metropolis-Hastings (PMH) sampler for this model. i) Describe your proposal distributions for β, σ, and compute the Metropolis-Hastings acceptance probability ratio. ii) Provide evidence that you have succeeded.*

2.4 Stochastic volatility unknown parameters part II

Since the inverse Gamma priors (3) are conjugate to the model (1), the posterior distributions of the variance parameters are available in closed form and given by

$$p(\sigma^2|\phi, x_{1:T}, y_{1:T}) = \mathcal{IG}\left(\sigma^2|a + \frac{T}{2}, b + \frac{1}{2}\sum_{t=1}^{T}(x_t - \phi x_{t-1})^2\right), \tag{5a}$$
$$p(\beta^2|\phi, x_{1:T}, y_{1:T}) = \mathcal{IG}\left(\beta^2|a + \frac{T}{2}, b + \frac{1}{2}\sum_{t=1}^{T}\exp(-x_t)y_t^2\right). \tag{5b}$$

Implement a particle Gibbs sampler to compute the posterior distribution $p(\sigma^2, \beta^2|\phi, y_{1:T})$. To do so, alternately sample from

- $p(\sigma^2|\phi, x_{1:T}, y_{1:T})$ based on (5a),
- $p(\beta^2|\phi, x_{1:T}, y_{1:T})$ based on (5b),
- $p(x_{0:T}|\theta, y_{1:T}, \sigma, \beta)$ using conditional SMC.

Question 11: *Implement conditional SMC sampler for this model. Re-use the SMC implemented in 2.2 as much as possible. Report the marginal distributions of σ^2 and β^2 using histograms for the two parameters. Remember to discard the transient ('burn-in') phase of the MCMC chain. Verify that the two implemented MCMC schemes (PMH and PG) produce the same marginal distributions for the variance parameters and compare their convergence rates.*

3.1 PyClone Light

In this problem, you should apply a collapsed Gibbs sampler in order to cluster mutations based on a Dirichlet Process Mixture Model (DPMM). The probabilistic model can, following the original publication, be described by the graphical model in Figure 1, where H is a Dirichlet Process (DP).

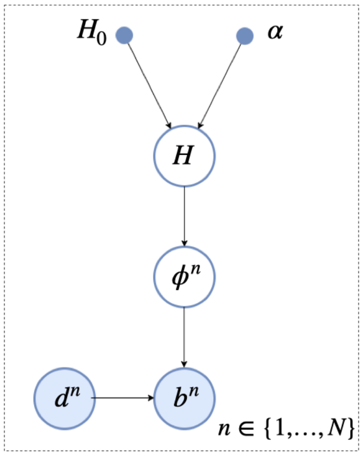

Figure 1: PyClone Light

$$
\begin{aligned}
H_0 &= \text{Beta}(a_0, a_1) \\
H | H_0, \alpha &\sim DP(\alpha, H_0) \\
\phi^n | H &\sim H \\
b^n &\sim \text{Bin}(d^n, \phi^n)
\end{aligned}
$$

Question 12: *Implement forward simulator to generate synthetic data that accepts hyper parameters and the number of data points as inputs, use $d^n \sim Poisson(1000)$ for data generation.*

Question 13: *Implement collapsed Gibbs sampler for this DPMM. Test it on synthetic data with the following values of the hyper parameters: $\alpha = 1$, $a_0 = a_1 = 1$. Provide a figure depicting accuracy of clustering.*

Question 14: *The number of clusters can be controlled via the parameter α. Study how the performance deteriorates with increasing number of clusters. How is the number of data points affecting the performance?*

3.2 Hierarchical Finite Mixture Model

In this problem, you should apply two approaches to MCMC estimation random walk Metropolis algorithm (RWM) and Hamiltonian Monte Carlo (HMC) algorithm (used by STAN) to estimate the parameters of a hierarchical mixture model. Figure 2 shows the model graphically. The plates indicate replication over i, j and k. Consider $Y_i = \{Y_{i,1}, \ldots Y_{i,J_i}\}$ for $i = 1, \ldots, I$ are observed. Let $Y^*_{i,j,k}$ be the value of $Y_{i,j}$, which would be realized when $Z_{i,j} = k$, and let $\sigma_{i,k} = \frac{1}{\tau_{i,k}}$. The model parameters π_i, σ_i, μ_i are assumed to be unknown and the hyper parameter vectors $\alpha, \mu_0, \beta_0, \tau_0, \gamma_0$ can be assumed to be known or you could also put prior distribution on these parameters.

For details, see http://ceur-ws.org/Vol-1218/bmaw2014_paper_1.pdf.

$$
\begin{aligned}
\pi_i = (\pi_{i,1}, \dots, \pi_{i,K}) &\sim \text{Dirichlet}(\alpha_1, \dots, \alpha_K) \\
Z_{i,j} &\sim \text{Cat}(\pi_i) \\
Y^*_{i,j,k} &\sim \mathcal{N}(\mu_{i,k}, \sigma_{i,k}) \\
\mu_{i,k} &\sim \mathcal{N}(\mu_{0,k}, \beta_{0,k}) \\
\log \tau_{i,k} &\sim \mathcal{N}(\log(\tau_{0,k}), \gamma_{0,k})
\end{aligned}
$$

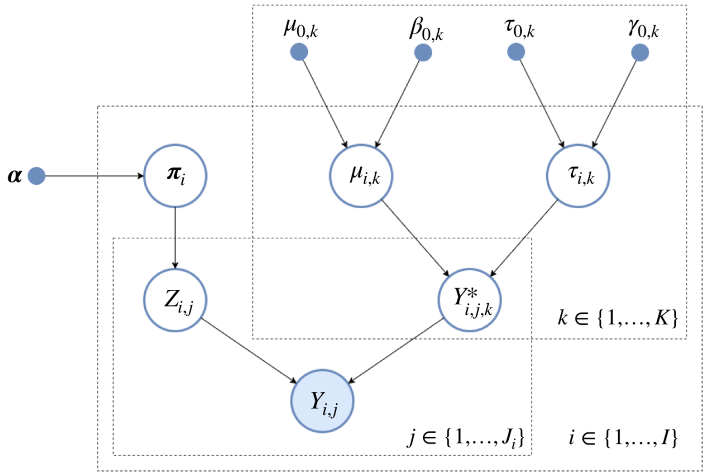

Figure 2: Hierarchical Mixture Model

Question 15: *Implement a generator to generate synthetic data for* $I = 100, K = 5$, $J_i \sim Poisson(10)$, $\alpha = (1,1,1,1,1), \mu_0 = (0,0,0,0,0), \beta_0 = (1,1,1,1,1), \tau_0 = (1,1,1,1,1)$ *and* $\gamma_0 = (1,1,1,1,1)$.

Question 16: *Implement a Random Walk Metropolis algorithm (RWM) for this model.*

Question 17: *Implement a Hamiltonian Monte Carlo (HMC) algorithm in STAN programming language for this model.*

Question 18: *Plot the auto-correlation function of your chains for each MCMC approach (RWM and HMC).*

Question 19: *How does STAN compare to RWM?*

Acknowledgement: Course in Sequential Monte Carlo Methods http://www.it.uu.se/research/systems_and_control/education/2017/smc.