IEOR E4525
Martin Haugh

Due: Wednesday 20^{th} April 2016

Assignment 7

1. *(Conjugate Priors)*

   (a) Consider the following form of the Normal distribution

   $$p(x \mid \mu, \kappa) = \frac{\kappa^{\frac{1}{2}}}{\sqrt{2\pi}} e^{-\frac{\kappa(x-\mu)^2}{2}}$$

   where κ (the variance inverse) is called the precision parameter. Show that this distribution can be written as an Exponential Family distribution of the form

   $$p(x \mid \theta_1, \theta_2) = h(x) e^{-\frac{\theta_1 x^2}{2} + \theta_2 x - \psi(\theta_1, \theta_2)}$$

   Characterize $h(x)$, (θ_1, θ_2) and the function $\psi(\theta_1, \theta_2)$.

   (b) Recall that the generic conjugate prior for an exponential family distribution is given by

   $$\pi(\theta_1, \theta_2) \propto e^{a_1\theta_1 + a_2\theta_2 - \gamma\psi(\theta_1, \theta_2)}.$$

   Substitute your expression for (θ_1, θ_2) from part (a) to show that the conjugate prior for the Normal model is of the form

   $$\pi(\kappa \mid a_0, b_0) \cdot \pi(\mu \mid \mu_0, \gamma\kappa) \propto \underbrace{\kappa^{a_0-1} e^{-\frac{\kappa}{b_0}}}_{\text{Gamma}(\kappa|a_0,b_0)} \cdot \underbrace{\kappa^{\frac{1}{2}}\, e^{-\frac{\gamma\kappa}{2}(\mu-\mu_0)^2}}_{\text{Normal}(\mu|\mu_0,\gamma\kappa)}.$$

   Your expressions for a_0, b_0 and μ_0 should be in terms of γ, a_1 and a_2. (This prior is known as the Normal-Gamma prior.)

   (c) Suppose $(\mu, \kappa) \sim \text{Normal-Gamma}(a_0, b_0, \mu_0, \gamma)$, and the likelihood of the data, x, is $p(x \mid \mu, \kappa) = \frac{\kappa^{\frac{1}{2}}}{\sqrt{2\pi}} e^{-\frac{\kappa(x-\mu)^2}{2}}$. Compute the posterior distribution after you see N IID samples $\{x_1, \ldots, x_N\}$.

2. *(Conjugate Priors ... Again)*

   Suppose the data x satisfies $x \,|\, p \sim \text{Bernoulli}(p)$. Repeat steps (a)-(c) from the previous question for this model. In step (b), the natural prior you obtain should be the beta distribution. (This is known as the Beta-Bernoulli model.)

3. *(Convergence Diagnostics)*

In the lecture slides we defined

$$\widehat{\mathrm{Var}}^{+}(\psi \mid \mathbf{X}) := \frac{n-1}{n} W \; + \; \frac{1}{n} B \tag{1}$$

where

$$\begin{aligned} B \;&:=\; \frac{n}{m-1} \sum_{j=1}^{m} \left(\bar{\psi}_{.j} - \bar{\psi}_{..}\right)^2 \\ W \;&:=\; \frac{1}{m} \sum_{j=1}^{m} s_j^2 \quad \text{where} \quad s_j^2 := \frac{1}{n-1} \sum_{i=1}^{n} \left(\psi_{ij} - \bar{\psi}_{.j}\right)^2 . \end{aligned}$$

These definitions were based on having m chains each with n samples after discarding the burn-in samples and ψ is some scalar function of the parameters / hidden variables over which the posterior is defined. We claimed that $\widehat{\mathrm{Var}}^{+}(\psi \mid \mathbf{X})$ was an unbiased estimator for $\mathrm{Var}^{+}(\psi \mid \mathbf{X})$ under stationarity. In this question, we will justify this claim.

(a) Suppose $Y_1, \ldots, Y_n$ is a sample from a stationary process with mean μ and autocovariance function $\gamma(h)$. Show that

$$\mathrm{Var}(\bar{Y}) = \frac{\gamma(0)}{n} R_n \tag{2}$$

where $R_n := 1 + 2\sum_{h=1}^{n-1} \rho(h)\left(1 - \frac{h}{n}\right)$ and $\rho(h) := \gamma(h)/\gamma(0)$ is the autocorrelation function. Note that $\gamma(0) = \mathrm{Var}(Y)$. (If you don't know what the autocovariance function is try `Google`, `Wikipedia` or any time-series book.) Most stationary processes generated by MCMC have $\rho(h) \geq 0$ so that if we use (2) to estimate $\mathrm{Var}(Y)$ then we need to take this autocorrelation into account.

(b) Suppose now that Y follows an $AR(1)$ process (a reasonable approximation to an MCMC process) so that $Y_n = \phi Y_{n-1} + \epsilon$. In that case it is straightforward to check that $\rho(h) = \phi^h$. Now justify the approximation

$$R_n \approx \frac{1+\phi}{1-\phi}.$$

(c) Use the identity

$$\sum_{i=1}^{n} (Y_i - \mu)^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 + n(\bar{Y} - \mu)^2$$

and (2) to show that $\mathrm{E}\left[\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right] = \gamma(0)(n - R_n)$. Argue then that

$$\widehat{\mathrm{Var}}(Y) := \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 + \widehat{\gamma(0)R_n}}{n}$$

is an unbiased estimator of $\mathrm{Var}(Y)$ when $\widehat{\gamma(0)R_n}$ is an unbiased estimator of $\gamma(0)R_n$.

(d) Explain how you could construct such an unbiased estimator of $\gamma(0)R_n$ using m realizations (each of length n) of the process. Now justify (1).

4. *(Gibbs and the Hierarchical Normal Model)*

   Consider the hierarchical Normal model from lecture slides #29 to #33. (This model is taken from Gelman et al's *Bayesian Data Analysis*.)

   (a) Write your own Gibbs sampler code in the language of your choice to sample from the posterior distribution.

   Hint: To simulate $X \sim \text{Inv-}\chi^2(\nu, s^2)$ first simulate Y from the χ^2_ν distribution and then set $X = \nu s^2/Y$.

   (b) Implement the Gelman-Rubin diagnostic by running 4 chains from over-dispersed starting points, discarding the first 50% of samples etc.

   (c) After running your code from (a) and (b) (and checking that the convergence diagnostics are satisfied!) report posterior quantiles (at the 2.5%, 25%, 50%, 75% and 97.5% levels) for θ_1, θ_2, θ_3, θ_4, μ, σ and τ. (Figure 1 displays results from Gelman et al's *Bayesian Data Analsyis*. You should obtain similar results.)

5. *(Decoding English Text)*

   In this problem you have to construct an MCMC algorithm to decode an English sentence that has been encoded using a substitution cipher. The file `Assign7_DecodingEnglishText.m` contains some code snippets as well as the coded text. (The coded text is available as text and as an array of numbers where the letters a-z are encoded using the numbers $1-26$ and space is encoded as the number 27.)

   The single-letter transition matrix $A(i,j) = \mathbb{P}(x_t = i \mid x_{t-1} = j)$ and the two-letter transition matrices $S(i,j,k) = \mathbb{P}(x_t = i \mid x_{t-1} = j, x_{t-2} = k)$ are in the file `English_trans.mat`.

Estimand	Posterior quantiles 2.5%	25%	median	75%	97.5%	$\widehat{R}$
θ_1	58.9	60.6	61.3	62.1	63.5	1.01
θ_2	63.9	65.3	65.9	66.6	67.7	1.01
θ_3	66.0	67.1	67.8	68.5	69.5	1.01
θ_4	59.5	60.6	61.1	61.7	62.8	1.01
μ	56.9	62.2	63.9	65.5	73.4	1.04
σ	1.8	2.2	2.4	2.6	3.3	1.00
τ	2.1	3.6	4.9	7.6	26.6	1.05
$\log p(\mu, \log\sigma, \log\tau \vert y)$	−67.6	−64.3	−63.4	−62.6	−62.0	1.02
$\log p(\theta, \mu, \log\sigma, \log\tau \vert y)$	−70.6	−66.5	−65.1	−64.0	−62.4	1.01

Figure 1: Results for Exercise 4 from Gelman et al.'s *Bayesian Data Analysis.*

6. *(Bayesian Estimation of Covariance Matrices)*

   Read Section 20.9 of Ruppert and Matteson's *Statistics and Data Analysis for Financial Engineering.* The Wishart and inverse-Wishart distributions are distributions over symmetric positive definite matrices, i.e. covariance matrices. The Wishart distribution is a conjugate prior for the precision matrix, Σ^{-1}, of a multivariate normal distribution. It can also be used (with MCMC) as a prior distribution for the scale matrix of a multivariate t-distribution.

   You don't have to submit anything for this question!