Sampling from "intractable" posterior distributions

September 8, 2016

1 Motivation

Denote the observations by $\mathbf{x}$ with sampling density (pdf/pmf) $f(\mathbf{x} \mid \theta)$ with $\theta \in \Theta \subset \mathbb{R}^d$. Let $\pi(\cdot)$ be a prior on θ. The posterior

$$\pi(\theta \mid \mathbf{x}) = \frac{f(\mathbf{x} \mid \theta)\pi(\theta)}{m(\mathbf{x})}, \quad m(\mathbf{x}) = \int f(\mathbf{x} \mid \theta)\pi(\theta)d\theta.$$

In the i.i.d. case, $f(\mathbf{x} \mid \theta) = \prod_{i=1}^n f(x_i \mid \theta)$.

In non-conjugate settings, we need a general set of tools to "compute" the posterior density. Here, by "computing" the posterior density, we mean that we should be able to calculate any posterior functional, such as the posterior mean, variance, median, quantiles of θ or of $\psi(\theta)$, where ψ is a known function. Even in conjugate settings, we have seen examples where the posterior distribution of certain $\psi(\theta)$s may be hard to obtain analytically, and we had to resort to Monte Carlo techniques. In general, our aim is to be able to (approximately) sample from the posterior distribution, so that the distribution of any posterior functional can be approximated. For example, if $\theta_1, \ldots, \theta_T$ are (approximately) independent samples from the posterior, then $\psi(\theta_1), \ldots, \psi(\theta_T)$ are samples from the posterior distribution of $\psi(\theta) \mid \mathbf{x}$, and we can use these samples to approximate the posterior mean/median/quantiles etc of $\psi(\theta)$.

The main bottleneck in sampling from the posterior is that the normalizing constant $m(\mathbf{x})$ is generally "intractable". This may be due to the fact that the integral is not analytically available, or the integral is highly expensive to compute, or a combination of both. For example, if $f(x \mid \theta) \propto [1 + (x - \theta)^2]^{-1}$, a Cauchy distribution with location θ, and $\theta \sim N(0, 1)$, then the integral is clearly not a standard one. As a second example, consider $x \mid \theta \sim 0.5N(\mu_1, 1) + 0.5N(\mu_2, 1)$, with $\mu_1, \mu_2 \sim N(0, 1)$ independently. Then,

$$f(\mathbf{x} \mid \theta) = 2^{-n} \sum_{j=0}^{n} \sum_{S:|S|=j} \left[\int \prod_{i \in S} \phi(x_i - \mu_1)\phi(\mu_1)d\mu_1 \right] \left[\int \prod_{l \in S^c} \phi(x_l - \mu_2)\phi(\mu_2)d\mu_2 \right],$$

where ϕ is the standard normal cdf and S denotes a subset of $\{1, \ldots, n\}$ with $|S|$ its size. Clearly, each of the inner integrals can be calculated analytically, but we have an outer sum over 2^n terms.

2 Some strategies for sampling from the posterior

Suppose that the observations are discrete, i.e., take values in a countable set with probability one. Let us also denote the observed data by $\mathbf{x}_{ons}$ here. Consider the following algorithm:
Algorithm (discrete ABC):

(i) Sample $\theta \sim \pi$.

(ii) Sample $\mathbf{x} \sim f(\cdot \mid \theta)$.

(iii) If $\mathbf{x} = \mathbf{x}_{obs}$, retain θ. Otherwise, discard θ.

We claim that if θ is retained at step (iii), then θ is a sample from the posterior! To see this, let B be any Borel subset of Θ. If θ is retained at step (iii), then

$$P(\theta \in B) = P(\theta \in B \mid \mathbf{x} = \mathbf{x}_{obs}) = \frac{P(\theta \in B, \mathbf{x} = \mathbf{x}_{obs})}{P(\mathbf{x} = \mathbf{x}_{obs})}$$
$$= \frac{\int_B f(\mathbf{x}_{obs} \mid \theta)\pi(\theta)d\theta}{\int f(\mathbf{x}_{obs} \mid \theta)\pi(\theta)d\theta} = \pi(\theta \in B \mid \mathbf{x}_{obs}).$$

We used the fact that since $\mathbf{x}$ is sampled from $f(\cdot \mid \theta)$, $P(\mathbf{x} = \mathbf{x}_{obs} \mid \theta) = f(\mathbf{x}_{obs} \mid \theta)$.

The above algorithm is the simplest version of a class of algorithms which are known as ABC (approximate Bayes computation) [When the data are discrete, there is no approximation though]. The basic idea is extremely simple: draw samples from the prior, generate "pseudo-data" conditioned on the parameter value, and check if the "pseudo-data" *matches* the observed data. If for example the data are iid samples, then $\mathbf{x} = \mathbf{x}_{obs}$ means that the same set of values are obtained, irrespective of the order.

An obvious drawback of the above algorithm is that it may be very inefficient, i.e., one may need to draw a very large number of prior samples to have a moderate number of posterior samples. A dramatic improvement can be obtained by modifying step (iii) to the condition $T(\mathbf{x}) = T(\mathbf{x}_{obs})$, where T is a sufficient statistic. [Prove this!]

Another ground for improvement is to replace the "hard" accept-reject step by a "softer" criterion, where we do not entirely discard θ, rather retain it with some probability. Since the probability of the random event $\mathbf{x} = \mathbf{x}_{obs}$ is $P(\mathbf{x} = \mathbf{x}_{obs} \mid \theta) = f(\mathbf{x}_{obs} \mid \theta)$, it makes sense to retain θ with probability proportional to $f(\mathbf{x}_{obs} \mid \theta)$. We shall see momentarily that this works even when $\mathbf{x}$ is not discrete, and the algorithm retains θ with probability proportional to the likelihood $f(\mathbf{x}_{obs} \mid \theta)$. This is also very intuitive, keep θ values which have a higher likelihood with a higher probability.

Algorithm (Bootstrap filter):

(i) Sample $\theta_1, \ldots, \theta_T \sim \pi$ independently.

(ii) Set

$$w_t = \frac{f(\mathbf{x}_{obs} \mid \theta_t)}{\sum_{j=1}^T f(\mathbf{x}_{obs} \mid \theta_j)}.$$

(iii) Keep θ_t with probability w_t. In other words, $\hat{\Pi}_T := \sum_{t=1}^T w_t \delta_{\theta_t}$ is our (random) discrete approximation to the posterior distribution. Here and elsewhere, δ_u denotes a point mass at u.

It is straightforward to show that for any Borel set B, $\hat{\Pi}_T(B) \to \Pi(B \mid \mathbf{x}_{obs})$ as $T \to \infty$. To see this,

$$\hat{\Pi}_T(B) = \frac{T^{-1}\sum_{t=1}^T f(\mathbf{x}_{obs} \mid \theta_t)\mathbb{1}(\theta_t \in B)}{T^{-1}\sum_{t=1}^T f(\mathbf{x}_{obs} \mid \theta_t)} \to \frac{\int_B f(\mathbf{x}_{obs} \mid \theta)\pi(\theta)d\theta}{\int f(\mathbf{x}_{obs} \mid \theta)\pi(\theta)d\theta},$$

almost surely by SLLN. Clearly, the last expression is the posterior probability of the set B. Along similar lines, (and maybe with a few additional assumptions), we can show that for any "nice" function $g : \Theta \to \mathbb{R}$,

$$\sum_{j=1}^T w_j g(\theta_j) \to \int g(\theta)\pi(\theta \mid \mathbf{x}_{obs})$$

almost surely as $T \to \infty$, provided the right hand side exists and is finite. This in particular

means we can approximate any posterior moments from the discrete approximation. Same is true of posterior quantiles, which allows us to construct credible intervals for the unknown parameters.

A very useful modification of the Bootstrap filter can be achieved by sampling from an *importance density* q in the first step instead of the prior π. The weights then need to be appropriately adjusted to keep the target distribution the same. This importance density may be derived from a gaussian approximation to the posterior or a kernel density estimator fitted to a previous discrete approximation to the posterior.

Algorithm (Bootstrap filter with IS): Let q be a positive density on Θ.

(i) Sample $\theta_1, \ldots, \theta_T \sim q$ independently.

(ii) Set

$$\omega_t = \frac{f(\mathbf{x}_{obs} \mid \theta_t)\pi(\theta_t)/q(\theta_t)}{\sum_{j=1}^{T} f(\mathbf{x}_{obs} \mid \theta_j)\pi(\theta_j)/q(\theta_j)}.$$

(iii) Keep θ_t with probability ω_t, i.e., $\hat{\Pi}_T^{IS} := \sum_{t=1}^{T} \omega_t \delta_{\theta_t}$ is the discrete approximation to the posterior distribution.

Verify that all the properties of $\hat{\Pi}_T$ remain intact for $\hat{\Pi}_T^{IS}$. Indeed, with a "good" importance density q, $\hat{\Pi}_T^{IS}$ may be efficient by orders of magnitude. For choosing q, one thing to be careful about is that q is not too light tailed. If q has lighter tails than the posterior, then one may potentially underestimate uncertainty. A default choice is to use heavy tailed distributions like the t. The mean and covariance may be set to be the mle and a constant (> 1) multiple of the inverse Fisher information respectively in regular models.

3 Gibbs sampling

Gibbs sampling refers to a class of Markov chain Monte Carlo algorithms where one samples iteratively from the *full conditional* distributions to create a Markov chain whose stationary distribution is the posterior distribution.

3.1 Why does the Gibbs sampler work?

Suppose (u, v) have a bivariate normal distribution with $u \sim N(0,1)$, $v \sim N(0,1)$ and $\text{corr}(u, v) = \rho$. Let $r = 1 - \rho^2$. From standard bivariate normal theory, we know $u \mid v \sim N(\rho v, r)$ and $v \mid u \sim N(\rho u, r)$.

A Gibbs sampler proceeds as:

- Initialize $u = u^{(0)}$.
- For $t = 1, \ldots, T$, repeat:
  - Sample $v^{(t)} \sim N(\rho u^{(t-1)}, r)$ by letting $v^{(t)} = \rho u^{(t-1)} + \epsilon^{(t)}$, where $\epsilon^{(t)} \sim N(0, r)$ is independent of everything else.
  - Sample $u^{(t)} \sim N(\rho v^{(t)}, r)$ by letting $u^{(t)} = \rho v^{(t)} + \eta^{(t)}$, where $\eta^{(t)} \sim N(0, r)$ is independent of everything else.

If $u^{(0)} \sim N(0,1)$, then it follows from a simple calculation that $v^{(1)} \sim N(0,1)$, $u^{(1)} \sim N(0,1)$ and in fact $u^{(1)}$ and $v^{(1)}$ have a bivariate normal distribution with correlation ρ. In fact, this is true for every $u^{(t)}$ and $v^{(t)}$. This has to be the case since $N(0,1)$ is the stationary distribution of the chain $u^{(t)}$.

We argued in class that for arbitrary starting value $u^{(0)}$, the Markov chain $u^{(t)}$ "converges" to the stationary distribution. The main observation requires here is that $u^{(t)}$ is a first-order autoregressive process which is gaussian conditional on the initial value, and the impact of the initial value geometrically decreases for large t.

3.2 Some useful sampling algorithms

Sampling from a class of multivariate Gaussians: Suppose $\beta \sim N(Q^{-1}b, Q^{-1})$, where Q is a $d \times d$ positive definite matrix and b is a $d \times 1$ vector. We have seen that these type of conditional posteriors routinely arise when there is a conjugate normal prior. The following sampling algorithm (Rue, 2001) avoids calculating the inverse of Q and only requires a Cholesky factorization and a series of linear system solutions, both of which are more efficient and stable compared to computing inverse.

- Perform a Cholesky decomposition $Q = LL^{\mathrm{T}}$, where L is lower triangular.
- Draw $z \sim N(0, \mathrm{I}_d)$, solve $L^{\mathrm{T}}y = z$.
- Solve $L^{\mathrm{T}}\theta = v$, where $Lv = b$.
- Set $\beta = y + \theta$.

It is straightforward to show that β produced as above has the desired distribution.

Sampling from Dirichlet distribution: Suppose $\pi = (\pi_1, \ldots, \pi_{k-1}) \sim \mathrm{Dirichlet}(\alpha_1, \ldots, \alpha_{k-1}, \alpha_k)$. To sample π, draw Γ_i independently from $\mathrm{Gamma}(\alpha_i, 1)$ for $i = 1, \ldots, k$ and set $\pi_l = \Gamma_l / (\sum_{i=1}^{k} \Gamma_i)$.

3.3 Some basic Gibbs samplers

Binomial with unknown sample size: Suppose $y \mid N, p \sim \mathrm{Binomial}(N, p)$, where N and p are both unknown. Consider independent priors on N and p, with $N \sim \mathrm{Poisson}(\lambda)$, and $p \sim U(0, 1)$. The full conditionals are:

- $p \mid N, y \sim \mathrm{Beta}(y + 1, N - y + 1)$.
- To sample $N \mid p, y$, set $N = y + t$, with $t \sim \mathrm{Poisson}(\lambda(1 - p))$. [Verify.]

Linear regression with ridge prior: Suppose $y \mid \beta, \sigma^2 \sim N(X\beta, \sigma^2 \mathrm{I}_n)$, where X is a $n \times d$ matrix of covariates and $\beta \in \mathbb{R}^d$ is a vector of covariates. Consider the following prior specification on β, σ^2, with $\beta \mid \sigma^2 \sim N(0, \lambda^{-1}\sigma^2 \mathrm{I}_d)$ and $\sigma^2 \sim \mathrm{IG}(\alpha/2, \gamma/2)$. λ, α, β are hyperparameters which we fix. The full conditionals are:

- $\beta \mid \sigma^2, y \sim N_d((X^{\mathrm{T}}X + \lambda \mathrm{I}_d)^{-1}X^{\mathrm{T}}y, \sigma^2(X^{\mathrm{T}}X + \lambda \mathrm{I}_d)^{-1})$. [Note: this is of the $N(Q^{-1}b, Q^{-1})$ form.]
- $\sigma^2 \mid \beta, y \sim \mathrm{IG}((n + d + \alpha)/2, \{(y - X\beta)^{\mathrm{T}}(y - X\beta) + \lambda\beta^{\mathrm{T}}\beta + \gamma\}/2)$.

3.4 Data augmentation Gibbs samplers

Consider the following two examples. The first one is that of linear regression with with t-distributed error, i.e., $y_i = x_i^{\beta} + \epsilon_i$ with $\epsilon_i \sim t_\nu(0, \sigma^2)$ independently for $i = 1, \ldots, n$. Suppose ν is given for now. Consider the same prior on β, σ^2 as before. The joint posterior of β, σ^2 is now proportional to

$$\pi(\beta, \sigma^2 \mid y) \propto \left\{ (\sigma^2)^{-n/2} \prod_{i=1}^{n} \left[1 + \frac{(y_i - x_i^{\mathrm{T}}\beta)^2}{\nu\sigma^2} \right]^{-(\nu+1)/2} \right\} \pi(\beta, \sigma^2).$$

This clearly does not admit conjugate full conditionals.

Next, consider probit regression where the response y_i is binary and $Pr(y_i = 1 \mid \beta) = \Phi(x_i^{\mathrm{T}}\beta)$, with Φ the normal cdf. With a Gaussian prior $\beta \sim N(0, \lambda^{-1}\mathrm{I}_d)$, the posterior of β is

$$\pi(\beta \mid y) \propto \pi(\beta) \prod_{i=1}^{n} [\Phi(x_i^{\mathrm{T}}\beta)]^{y_i} [1 - \Phi(x_i^{\mathrm{T}}\beta)]^{(1-y_i)}$$

which again does not admit conjugate full conditionals.

The idea of data augmentation is to introduce latent variables so that when integrated over the distribution of the latent variables, one recovers the original likelihood. This is often an extremely useful trick to enable Gibbs sampling.

Fitting an iid t model: Suppose $y_1, \ldots, y_n \mid \mu, \sigma^2, \nu \sim t_\nu(\mu, \sigma^2)$. Consider priors $\mu \mid \sigma^2 \sim N(0, \lambda^{-1}\sigma^2)$, $\sigma^2 \sim \mathrm{IG}(\alpha_1/2, \alpha_2/2)$. Also, consider a discrete uniform prior for ν on a set of pre-specified grid points $\{\nu_1^*, \ldots, \nu_G^*\}$. Consider the following hierarchical specification of the likelihood with data augmentation:

$$\begin{aligned} y_i \mid \tau_i, \mu, \sigma^2, \nu &\sim N(\mu, \tau_i^{-1}\sigma^2), \quad i = 1, \ldots, n \\ \tau_i &\sim \mathrm{Gamma}(\nu/2, \nu/2), \quad i = 1, \ldots, n. \end{aligned}$$

Exploiting the fact that a t density can be expressed as a scale-mixture of normals, it is clear that one obtains the iid t likelihood by integrating over the τ_is. However, we retain the τ_is to facilitate Gibbs sampling, which cycles through the following steps:

- Sample $\tau_i \mid \mu, \sigma^2, \nu, y$ independently for $i = 1, \ldots, n$ from $\mathrm{Gamma}((\nu+1)/2, (y_i-\mu)^2/(2\sigma^2)+\nu/2)$ distributions.
- Sample $\mu \mid \tau, \sigma^2, \nu, y$ from a $N((\sum_{i=1}^n \tau_i y_i)/(\sum_{i=1}^n \tau_i + \lambda), \sigma^2/(\sum_{i=1}^n \tau_i + \lambda)$ distribution.
- Sample $\sigma^2 \mid \mu, \tau, \nu, y$ from an $\mathrm{IG}((n + 1 + \alpha_1)/2, \{\sum_{i=1}^n \tau_i(y_i - \mu)^2 + \lambda\mu^2 + \alpha_2\}/2)$ distribution.
- To sample ν from its discrete conditional posterior, we have two options:
  - Draw $\nu \mid \tau, \mu, \sigma^2, y$ from the discrete distribution $Pr(\nu = \nu_g^* \mid \tau, \mu, \sigma^2, y) = w_g$, where $w_g \propto \prod_{i=1}^n \tau_i^{\nu/2-1}\{(\nu/2)^{\nu/2}/\Gamma(\nu/2)\}\tau_i^{\nu/2-1}e^{-\nu\tau_i/2}$.
  - Marginalize τ to draw $\nu \mid \mu, \sigma^2, y$ from the discrete distribution $Pr(\nu = \nu_g^* \mid \tau, \mu, \sigma^2, y) = w_g$, where $w_g \propto \prod_{i=1}^n t_{\nu_g^*}((y_i - \mu)/\sigma)$, with $t_\nu(x)$ denoting the standard t density with ν degrees of freedom evaluated at x.

Linear regression with t error: The above data augmentation scheme can be trivially extended to linear regression with t error by considering

$$\begin{aligned} y_i \mid \tau_i, \beta, \sigma^2, \nu &\sim N(x_i^{\mathrm{T}}\beta, \tau_i^{-1}\sigma^2), \quad i = 1, \ldots, n \\ \tau_i &\sim \mathrm{Gamma}(\nu/2, \nu/2), \quad i = 1, \ldots, n. \end{aligned}$$

The full conditionals are straightforward to work out (next exercise!).

Probit regression: The data augmentation scheme here is from a famous paper by Albert & Chib (1993). Let $y_i = \mathbb{1}(z_i > 0)$, where $z_i \sim N(x_i^{\mathrm{T}}\beta, 1)$. Clearly, this implies $Pr(y_i = 1) = \Phi(x_i^{\mathrm{T}}\beta)$. Letting $z = (z_1, \ldots, z_n)^{\mathrm{T}}$, the joint posterior of β, z is

$$\pi(\beta, z \mid y) \propto \pi(\beta) \prod_{i=1}^{n} \big[\mathbb{1}(z_i > 0)\mathbb{1}(y_i = 1) + \mathbb{1}(z_i \leq 0)\mathbb{1}(y_i = 0)\big]\phi(z_i - x_i^{\mathrm{T}}\beta).$$

The Gibbs sampler cycles through:

- Sample $\beta \mid y, z$ from $N((X^{\mathrm{T}}X + \lambda \mathrm{I}_d)^{-1}X^{\mathrm{T}}z, (X^{\mathrm{T}}X + \lambda \mathrm{I}_d)^{-1})$.
- For $i = 1, \ldots, n$, independently sample $z_i \mid y, \beta$ from
  - A $N(x_i^{\mathrm{T}}\beta, 1)$ truncated to $(0, \infty)$ if $y_i = 1$.
  - A $N(x_i^{\mathrm{T}}\beta, 1)$ truncated to $(-\infty, 0)$ if $y_i = 0$.

Mixture models: Mixture models are an extremely useful class of models that are used for semi-parametric density estimation, classification, regression and various other tasks. We shall focus on density estimation for illustration. Say $\{f_h : h = 1, \ldots, k\}$ are a set of density functions that are specified up to a few parameters. For example, f_h can be the density function of $N(\mu_h, \tau_h^{-1})$. A mixture of the f_hs take the form $f = \sum_{h=1}^k \pi_h f_h$, where $\pi_h \geq 0$ and $\sum_{h=1}^k \pi_h = 1$. Clearly, f is a density.

Gibbs sampling for location-scale mixtures of normals again uses a data augmentation trick. Suppose $f_h \equiv N(\mu_h, \tau_h^{-1})$. The unknown parameters here are $\{\pi_h, \mu_h, \tau_h\}_{h=1}^k$, which are endowed with the following priors:

$$\begin{aligned}
&\pi \sim \text{Dirichlet}(\alpha, \ldots, \alpha), \\
&\mu_h \mid \tau_h \sim N(\mu_0, \tau_0^{-1}\tau_h^{-1}), \\
&\tau_h \sim \text{Gamma}(a_\tau, b_\tau).
\end{aligned}$$

Note: to specify the above prior, we need to specify 5 hyperparameters: $\alpha, \mu_0, \tau_0, a_\tau, b_\tau$. A default choice of $\alpha = 1/k$. [Has connections with Dirichlet process mixtures, which are a class of infinite mixture models]

The main difficulty with mixture models is that the likelihood is intractable due to its combinatorial nature: the joint likelihood of $y_1, \ldots, y_n$ iid from $f = \sum_{h=1}^k \pi_h f_h$ can be written as

$$\prod_{i=1}^n \sum_{h=1}^k \pi_h f_h(y_i) = \sum_{\gamma_1=1}^k \cdots \sum_{\gamma_n=1}^k \pi_1^{n_1} \ldots \pi_k^{n_k} f_{\gamma_1}(y_1) \ldots f_{\gamma_n}(y_n),$$

where $n_j = \#(\gamma_i = j)$. To see this, start by noting that both sides have k^n terms in the summand.

The above likelihood is clearly intractable to deal with due to the sum over the exponential number of terms. However, the way the sum is arranged gives us the idea of data augmentation. For each individual i, introduce a latent index $\gamma_i \in \{1, \ldots, k\}$ with $\text{pr}(\gamma_i = h) = \pi_h$. Then, we can write the model hierarchically as:

$$\begin{aligned}
&y_i \mid z_i = h, \mu, \tau, \pi \sim N(\mu_h, \tau_h^{-1}), \quad i = 1, \ldots, n \\
&\pi \sim \text{Dirichlet}(\alpha, \ldots, \alpha), \\
&\mu_h \mid \tau_h \sim N(\mu_0, \tau_0^{-1}\tau_h^{-1}), \\
&\tau_h \sim \text{Gamma}(a_\tau, b_\tau).
\end{aligned}$$

Verify that the Gibbs sampler cycles through the steps:

1. Update γ_i from its discrete conditional posterior with

$$\Pr(\gamma_i = h \mid -) = \frac{\pi_h \phi(y_i; \mu_h, \tau_h^{-1})}{\sum_{l=1}^k \pi_l \phi(y_i; \mu_l, \tau_l^{-1})}.$$

2. Update μ_h, τ_h from their conditional posterior:

$$\tau_h \mid - \sim \text{Gamma}(a, b), \quad a = a_\tau + n_h/2,$$
$$b = b_\tau + \frac{1}{2}\Big[\sum_{i:\gamma_i=h}(y_i - \bar{y}_h)^2 + \frac{\tau_0 n_h}{\tau_0 + n_h}(\bar{y}_h - \mu_0)^2\Big],$$
$$\mu_h \mid - \sim N(\hat{\mu}_h, \eta_h^{-1}\tau_h^{-1}),\ \eta_h = (n_h + \tau_0), \hat{\mu}_h = (\tau_0\mu_0 + n_h\bar{y}_h)/(n_h + \tau_0).$$

   In the above display, $n_h = \#(i : \gamma_i = h)$ and $\bar{y}_h = (\sum_{i:\gamma_i=h} y_i)/n_h$.

3. Update $\pi \mid - \sim \text{Dirichlet}(\alpha + n_1, \ldots, \alpha + n_k)$.

Check the correctness of the above steps. Straightforward to implement the above Gibbs sampler. Discarding a burn-in period, we can plot $f^t(y_g) = \sum_{h=1}^k \pi_h^t \phi(y_g; \mu_h^t, (\tau_h^{-1})^t)$ at MCMC iteration t on a fine set of grid points y_g. Can compute posterior mean and pointwise intervals. Can also do inference on any functional of the density. For example, suppose we are interested in obtaining a confidence interval for the hazard function $h(t) = 1 - F(t)$, with F the cdf of f, at a given point t.