Gibbs and Metropolis sampling (MCMC methods)
and relations of Gibbs to EM

Lecture Outline

1. Gibbs
    - the algorithm
    - a bivariate example
    - an elementary convergence proof for a (discrete) bivariate case
    - more than two variables
    - a counter example.

2. ***EM*** – again
    - ***EM*** as a maximization/maximization method
    - Gibbs as a variation of Generalized *EM*

3. Generating a Random Variable.
    - Continuous r.v.s and an exact method based on transforming the cdf.
    - The "accept/reject" algorithm.
    - The Metropolis Algorithm

Gibbs Sampling

We have a joint density

$$f(x, y_1, \ldots, y_k)$$

and we are interested, say, in some features of the marginal density

$$f(x) = \int\int\ldots\int f(x, y_1, \ldots, y_k)\, dy_1, dy_2, \ldots, dy_k.$$

*F*or instance, suppose that we are interested in the average

$$E[X] = \int x\, f(x) dx.$$

If we can sample from the marginal distribution, then

$$lim_{m\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = E[X]$$

without using $f(x)$ explicitly in integration. Similar reasoning applies to any other characteristic of the statistical model, i.e., of the *population*.

The Gibbs Algorithm for computing this average.

Assume we can sample the k+1-many univariate conditional densities:

$f(X \mid y_1, \ldots, y_k)$

$f(Y_1 \mid x, y_2, \ldots, y_k)$

$f(Y_2 \mid x, y_1, y_3, \ldots, y_k)$

...

$f(Y_k \mid x, y_1, y_3, \ldots, y_{k-1})$.

Choose, arbitrarily, k initial values: $Y_1 = y_1^0$, $Y_2 = y_2^0$, …., $Y_k = y_k^0$.

Create: x^1 by a draw from $f(X \mid y_1^0, \ldots, y_k^0)$

y_1^1 by a draw from $f(Y_1 \mid x^1, y_2^0, \ldots, y_k^0)$

y_2^1 by a draw from $f(Y_2 \mid x^1, y_1^1, y_3^0 \ldots, y_k^0)$

…

y_k^1 by a draw from $f(Y_k \mid x^1, y_1^1, \ldots, y_{k-1}^1)$.

This constitutes one Gibbs “pass” through the k+1 conditional distributions,

yielding values: $(x^1, y_1^1,\ y_2^1,\ \ldots., y_k^1)$.

Iterate the sampling to form the second “pass”

$$(x^2, y_1^2,\ y_2^2,\ \ldots., y_k^2).$$

Theorem: (under general conditions)

The distribution of x^n converges to $F(x)$ as $n \to \infty$.

Thus, we may take the last n X-values after many Gibbs passes:

$$\frac{1}{n}\sum_{i=m}^{m+n} X^i \approx \mathrm{E}[X]$$

or take just the last value, $x_i^{n_i}$ of n-many sequences of Gibbs passes

$(i = 1, \ldots\ n)$ $$\frac{1}{n}\sum_{i=i}^{n} X_i^{n_i} \approx \mathrm{E}[X]$$

to solve for the average, $= \int x\, f(x) dx$.

A bivariate example of the Gibbs Sampler.

Example: Let X and Y have similar truncated conditional exponential distributions:

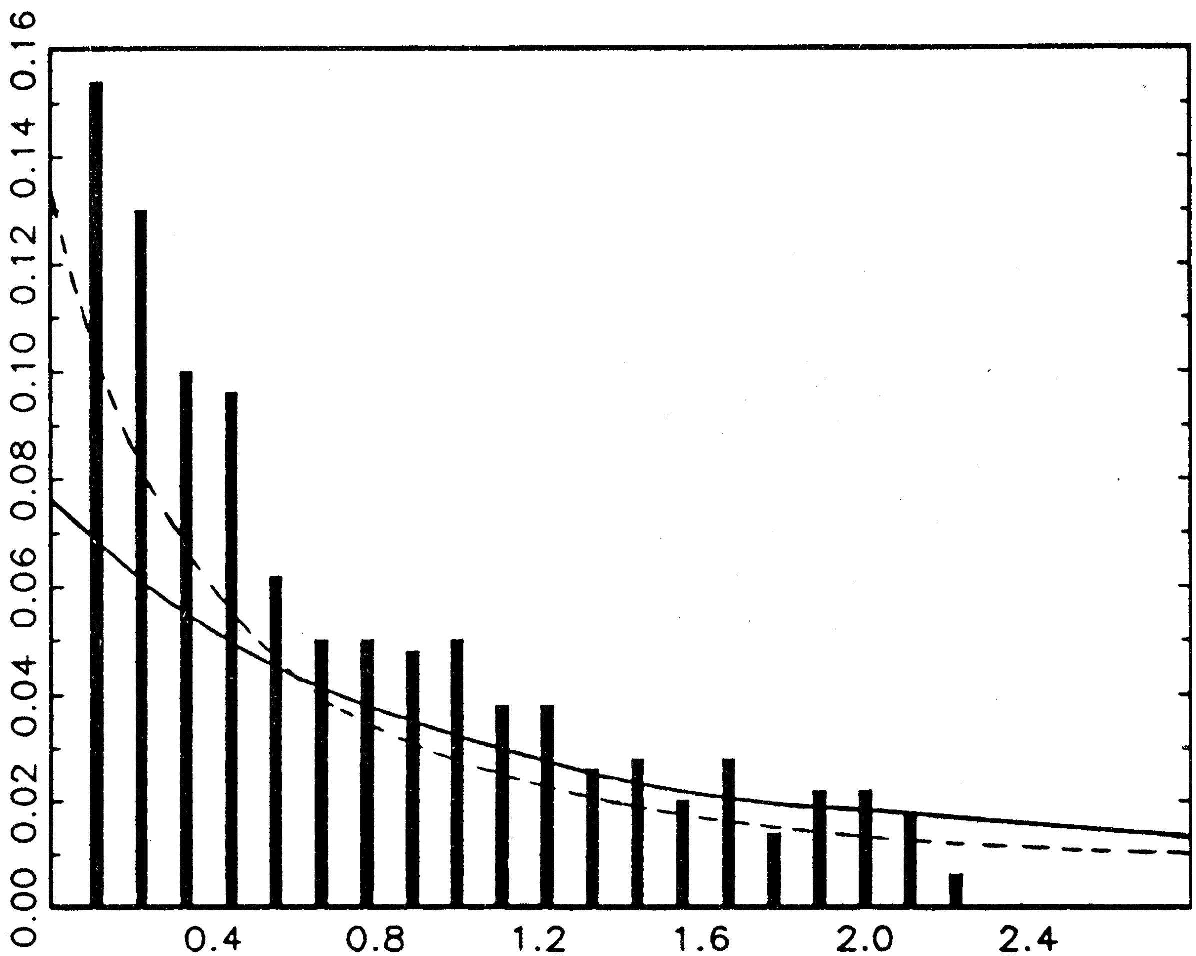
$$f(X \mid y) \propto ye^{-yx} \text{ for } 0 < \mathrm{X} < \boldsymbol{b}$$

$$f(Y \mid x) \propto xe^{-xy} \text{ for } 0 < \mathrm{Y} < \boldsymbol{b}$$

where $\boldsymbol{b}$ is a known, positive constant.

Though it is not convenient to calculate, the marginal density $f(X)$ is readily simulated by Gibbs sampling from these (truncated) exponentials.

Below is a histogram for X, $\boldsymbol{b} = 5.0$, using a sample of 500 terminal observations with 15 Gibbs' passes per trial, $x_i^{n_i}$ (i = 1,…, 500, n_i = 15) (from Casella and George, 1992).



Histogram for X, $\boldsymbol{b}$ = 5.0, using a sample of 500 terminal observations with 15 Gibbs' passes per trial, $x_i^{n_i}$ (i = 1,…, 500, n_i = 15). Taken from (Casella and George, 1992).

Here is an alternative way to compute the marginal $f(X)$ using the same Gibbs Sampler.

Recall the law of conditional expectations (assuming E[X] exists):

$$E[\, E[X \mid Y] \,] = E[X]$$

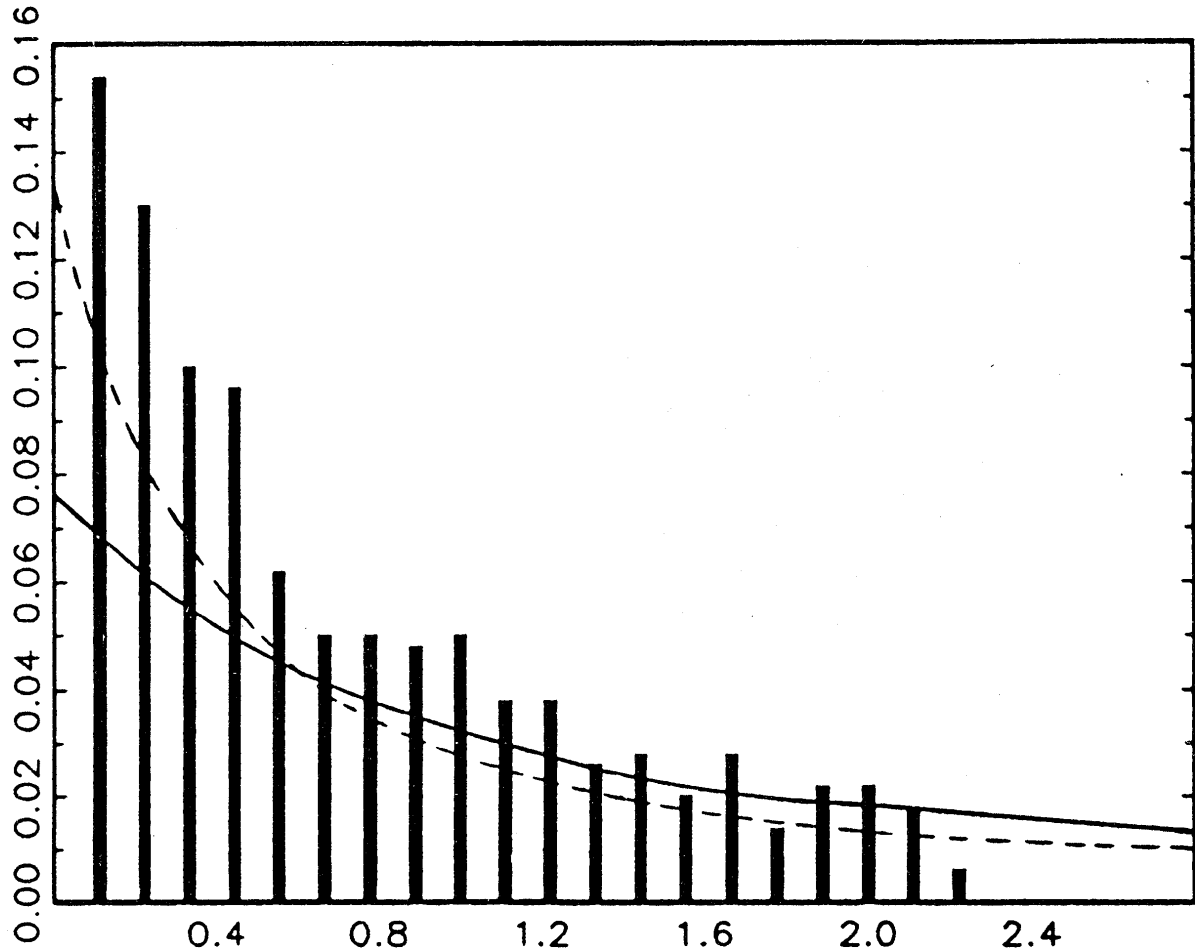
Thus $$E[f(x|Y)] = \int f(x \mid y) f(y) dy = f(x).$$

Now, use the fact that the Gibbs sampler gives us a simulation of the marginal density f(Y) using the penultimate values (for Y) in each Gibbs' pass, above: $y_i^{n_i-1}$ (i = 1, …500; n_i = 15).

Calculate $f(x \mid y_i^{n_i-1})$, which by assumption is feasible.

Then note that:

$$f(x) \approx \frac{1}{n} \sum_{i=i}^{n} \mathrm{f}(\mathrm{x} \mid y_i^{n_i-1})$$



The **solid line** graphs the alternative Gibbs Sampler estimate of the marginal $f(x)$ from eth same sequence of 500 Gibbs' passes, using $\int f(x \mid y) f(y) dy = f(x)$. The **dashed-line** is the exact solution. Taken from (Casella and George, 1992).

An elementary proof of convergence in the case of 2 x 2 Bernoulli data

Let (X,Y) be a bivariate variable, marginally, each is Bernoulli

$$\begin{array}{cc} & X \\ & \begin{array}{cc} 0 & 1 \end{array} \\ Y \begin{array}{c} 0 \\ 1 \end{array} & \begin{bmatrix} p_1 & p_2 \\ p_3 & p_4 \end{bmatrix} \end{array}$$

where $p_i \geq 0$, $\sum p_i = 1$, marginally

$$\mathbf{P}(X=0) = p_1 + p_3 \quad \text{and} \quad \mathbf{P}(X=1) = p_2 + p_4$$

$$\mathbf{P}(Y=0) = p_1 + p_2 \quad \text{and} \quad \mathbf{P}(Y=1) = p_3 + p_4.$$

The conditional probabilities $\mathbf{P}(X|y)$ and $\mathbf{P}(Y|x)$ are evident:

$\mathbf{P}(Y|x)$:

$$
\begin{array}{cc}
 & \boldsymbol{X} \\
 & \begin{array}{cc} 0 & \quad 1 \end{array} \\
\boldsymbol{Y}\begin{array}{c} 0 \\[3ex] 1 \end{array} & \begin{bmatrix} \dfrac{p_1}{p_1+p_3} & \dfrac{p_2}{p_2+p_4} \\[2ex] \dfrac{p_3}{p_1+p_3} & \dfrac{p_4}{p_2+p_4} \end{bmatrix}
\end{array}
$$

$\mathbf{P}(X|y)$:

$$
\begin{array}{cc}
 & \boldsymbol{X} \\
 & \begin{array}{cc} 0 & \quad 1 \end{array} \\
\boldsymbol{Y}\begin{array}{c} 0 \\[3ex] 1 \end{array} & \begin{bmatrix} \dfrac{p_1}{p_1+p_2} & \dfrac{p_2}{p_1+p_2} \\[2ex] \dfrac{p_3}{p_3+p_4} & \dfrac{p_4}{p_3+p_4} \end{bmatrix}
\end{array}
$$

Suppose (for illustration) that we want to generate the marginal distribution of X by the Gibbs Sampler, using the sequence of iterations of draws between the two conditional probabilites $\mathbf{P}(X|y)$ and $\mathbf{P}(Y|x)$.

That is, we are interested in the sequence $< x^i$: i = 1, … $>$ created from the starting value $y^0 = 0$ or $y^0 = 1$.

Note that:

$$\mathbf{P}(X^n = 0 \,|\, x^i : \mathrm{i} = 1, \ldots, n\text{-}1) = \mathbf{P}(X^n = 0 \,|\, x^{n-1}) \quad \textbf{\textit{the Markov property}}$$

$$= \mathbf{P}(X^n{=}0 \,|\, y^{n-1}{=}0)\,\mathbf{P}(Y^{n-1}{=}0 \,|\, x^{n-1}) \;+\; \mathbf{P}(X^n{=}0 \,|\, y^{n-1}{=}1)\,\mathbf{P}(Y^{n-1}{=}1 \,|\, x^{n-1})$$

Thus, we have the four (positive) transition probabilities:

$$\mathbf{P}(\mathcal{X}^n = \mathrm{j} \mid x^{n-1} = i) = p_{\mathrm{ij}} > 0, \text{ with } \textstyle\sum_i \sum_j p_{\mathrm{ij}} = 1 \quad (i, j = 0, 1).$$

With the transition probabilities positive, it is an (old) ergodic theorem that, $\mathbf{P}(\mathcal{X}^n)$ converges to a (unique) *stationary* distribution, independent of the starting value (y^0).

Next, we confirm the easy fact that the marginal distribution $\mathbf{P}(X)$ is that same distinguished *stationary* point of this Markov process.

$$\mathbf{P}(X^n = 0)$$

$$= \quad \mathbf{P}(X^n = 0 \mid x^{n-1} = \mathit{0})\, \mathbf{P}(X^{n-1} = 0) + \mathbf{P}(X^n = 0 \mid x^{n-1} = \mathit{1})\, \mathbf{P}(X^{n-1} = 1)$$

$$= \quad \mathbf{P}(X^n=0 \mid y^{n-1}=0)\, \mathbf{P}(Y^{n-1}=0 \mid x^{n-1} = \mathit{0})\, \mathbf{P}(X^{n-1} = 0)$$

$$+ \quad \mathbf{P}(X^n=0 \mid y^{n-1}=1)\, \mathbf{P}(Y^{n-1}=1 \mid x^{n-1} = \mathit{0})\, \mathbf{P}(X^{n-1} = 0)$$

$$+ \quad \mathbf{P}(X^n=0 \mid y^{n-1}=0)\, \mathbf{P}(Y^{n-1}=0 \mid x^{n-1} = \mathit{1})\, \mathbf{P}(X^{n-1} = 1)$$

$$+ \quad \mathbf{P}(X^n=0 \mid y^{n-1}=1)\, \mathbf{P}(Y^{n-1}=1 \mid x^{n-1} = \mathit{1})\, \mathbf{P}(X^{n-1} = 1)$$

$$= \quad \mathbf{E_P}\,[\mathbf{E_P}\,[X^n=0 \mid X^{n-1}]\,]$$

$$= \quad \mathbf{E_P}\,[X^n=0\,]$$

$$= \quad \mathbf{P}(X^n = 0)\,.$$

The *Ergodic* Theorem:

<u>*Definitions*</u>:

- A *Markov chain*, $X_0, X_1, \ldots.$ satisfies

  $$\mathbf{P}(X_n | x_i: \mathrm{i} = 1, \ldots, n\text{-}1) = \mathbf{P}(X_n | x_{n-1})$$

- The distribution $\boldsymbol{F}(x)$, with density $f(x)$, for a Markov chain is *stationary* (or *invariant*) if

  $$\int_{\mathbf{A}} \boldsymbol{f}(x)\, dx = \int \mathbf{P}(X_n \in \mathbf{A} \mid x_{n-1})\, \boldsymbol{f}(x)\, dx.$$

- The Markov chain is *irreducible* if each set with positive **P**-probability is visited at some point (almost surely).

- An irreducible Markov chain is *recurrent* if, for each set $\mathbf{A}$ having positive $\mathbf{P}$-probability, with positive $\mathbf{P}$-probability the chain visits $\mathbf{A}$ infinitely often.
- A Markov chain is *periodic* if for some integer $k > 1$, there is a partition into k sets $\{\mathbf{A}_1, \ldots, \mathbf{A}_k\}$ such that

  $\mathbf{P}(X_{n+1} \in \mathbf{A}_{j+1} \mid x_n \in \mathbf{A}_j) = 1$ for all $j = 1, \ldots, k\text{-}1 \pmod{k}$. That is, the chain cycles through the partition.

  Otherwise, the chain is *aperiodic*.

Theorem: If the Markov chain $X_0, X_1, \ldots.$ is irreducible with an invariant probability distribution $\boldsymbol{F}(x)$ then:

1. the Markov chain is recurrent
2. F is the unique invariant distribution

If the chain is aperiodic, then for $\boldsymbol{F}$-almost all x_0, both

3. $lim_{n\to\infty}\, sup_{\mathbf{A}} \mid \mathbf{P}(X_{\mathrm{n}} \in \mathbf{A} \mid X_0 = x_0\,) - \int_{\mathbf{A}} \boldsymbol{f}(x)\, dx \mid = 0$

And for any function $\boldsymbol{h}$ with $\int \boldsymbol{h}(x)\, dx < \infty$,

4. $lim_{n\to\infty}\ \frac{1}{n} \sum_{i=i}^{n} h(X_i) = \int \boldsymbol{h}(x)\, \boldsymbol{f}(x)\, dx \quad (= \mathbf{E}_{\mathbf{F}}[\boldsymbol{h}(x)]\,),$

That is, the *time average* of $\boldsymbol{h}(X)$ equals its *state-average*, *a.e.* $\boldsymbol{F}$.

A (now-familiar) puzzle.

Example (continued): Let X and Y have similar conditional exponential distributions:

$$f(X \mid y) \propto ye^{-yx} \text{ for } 0 < X$$

$$f(Y \mid x) \propto xe^{-xy} \text{ for } 0 < Y$$

To solve for the marginal density $f(X)$ use Gibbs sampling from these exponential distributions. The resulting sequence does ***not*** converge!

Question: Why does this happen?

Answer: (Hint: Recall HW #1, problem 2.) Let θ be the statistical parameter for X with $f(X|\theta)$ the exponential model. What "prior" density for θ yields the *posterior* $f(\theta \mid x) \propto xe^{-x\theta}$?
Then, what is the "prior" expectation for X?

Remark: Note that $W = X\theta$ is pivotal. What is its distribution?

More on this puzzle:

The conjugate prior for the parameter θ in the exponential distribution is the Gamma $\Gamma(\alpha, \beta)$.

$$\boldsymbol{f}(\theta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} \qquad \text{for } \theta, \alpha, \beta > 0,$$

Then the posterior for θ based on $\boldsymbol{x} = (x_1, .., x_n)$, *n iid* observations from the exponential distribution is

$$\boldsymbol{f}(\theta|\boldsymbol{x}) \text{ is Gamma } \Gamma(\alpha', \beta')$$

where $\alpha' = \alpha+n$ and $\beta' = \beta + \Sigma\, x_i$.

Let n=1, and consider the limiting distribution as $\alpha, \beta \to 0$.

This produces the "posterior" density $f(\theta \mid x) \propto x e^{-x\theta}$, which is mimicked in Bayes theorem by the improper "prior" density

$f(\theta) \propto 1/\theta$. But then $E_{\boldsymbol{F}}(\theta)$ does not exist!

Part 2 EM – again

- **EM as a maximization/maximization method**
- **Gibbs as a variation of Generalized EM**

EM as a maximization/maximization method.

Recall:

$\mathbf{L}(\theta\,;\boldsymbol{x})$ is the likelihood function for θ with respect to the incomplete data $\boldsymbol{x}$.

$\mathbf{L}(\theta\,;(\boldsymbol{x},\boldsymbol{z}))$ is the likelihood for θ with respect to the complete data $(\boldsymbol{x},\boldsymbol{z})$.

And $\mathbf{L}(\theta\,;\boldsymbol{z}\mid\boldsymbol{x})$ is a *conditional likelihood* for θ with respect to $\boldsymbol{z}$, given $\boldsymbol{x}$;

which is based on $\boldsymbol{h}(\boldsymbol{z}\mid\boldsymbol{x},\theta)$: the conditional density for the data $\boldsymbol{z}$, given $(\boldsymbol{x},\theta)$.

Then as $$\boldsymbol{f}(\boldsymbol{X}\mid\theta)=\boldsymbol{f}(\boldsymbol{X},\boldsymbol{Z}\mid\theta)\;/\;\boldsymbol{h}(\boldsymbol{Z}\mid\boldsymbol{x},\theta)$$

we have $$\boldsymbol{log}\,\mathrm{L}(\theta\,;\boldsymbol{x})=\boldsymbol{log}\,\mathrm{L}(\theta\,;(\boldsymbol{x},\boldsymbol{z}))-\boldsymbol{log}\,\mathrm{L}(\theta\,;\boldsymbol{z}\mid\boldsymbol{x})\qquad(*)$$

***As below, we use the* EM *algorithm to compute the* mle**

$$\hat{\theta}\;=\;\boldsymbol{argmax}_{\Theta}\,\mathrm{L}(\theta\,;\boldsymbol{x})$$

With $\hat{\theta}_0$ an arbitrary choice, define

(***E**-step)* $$\boldsymbol{Q}(\theta \mid \boldsymbol{x}, \hat{\theta}_0) = \int_Z [\boldsymbol{log}\ \mathbf{L}(\theta\,;\boldsymbol{x}, \boldsymbol{z})]\ \boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \hat{\theta}_0)\, dz$$

and

$$\boldsymbol{H}(\theta \mid \boldsymbol{x}, \hat{\theta}_0) = \int_{\boldsymbol{Z}} [\boldsymbol{log}\ \mathbf{L}(\theta\,; \boldsymbol{z} \mid \boldsymbol{x})]\ \boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \hat{\theta}_0)\, d\boldsymbol{z}.$$

then $$\boldsymbol{log}\ \mathbf{L}(\theta\,; \boldsymbol{x}) = \boldsymbol{Q}(\theta \mid \boldsymbol{x}, \theta_0) - \boldsymbol{H}(\theta \mid \boldsymbol{x}, \theta_0),$$

as we have integrated-out $\boldsymbol{z}$ from (*) using the conditional density $\boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \hat{\theta}_0)$.

The ***EM algorithm*** is an iteration of

i. the ***E***-step: determine the integral $\boldsymbol{Q}(\theta \mid \boldsymbol{x}, \hat{\theta}_j)$,

ii. the ***M***-step: define $\hat{\theta}_{j+1}$ as $\boldsymbol{argmax}_\Theta\ \boldsymbol{Q}(\theta \mid \boldsymbol{x}, \hat{\theta}_j)$.

Continue until there is convergence of the $\hat{\theta}_j$.

Now, for a *Generalized EM* algorithm.

Let be $\boldsymbol{P}(\boldsymbol{Z})$ any distribution over the augmented data $\mathbf{Z}$, with density $\boldsymbol{p}(\boldsymbol{z})$
Define the function $\boldsymbol{F}$ by:

$$\boldsymbol{F}(\theta, \boldsymbol{P}(\boldsymbol{Z})) = \int_Z [\boldsymbol{log}\ \mathrm{L}(\theta; \boldsymbol{x}, \boldsymbol{z})]\ \boldsymbol{p}(\boldsymbol{z})dz - \int_Z \boldsymbol{log}\ \boldsymbol{p}(\boldsymbol{z})\ \boldsymbol{p}(\boldsymbol{z})dz$$
$$= \mathbf{E}_{\boldsymbol{P}} [\boldsymbol{log}\ \mathrm{L}(\theta; \boldsymbol{x}, \boldsymbol{z})] - \mathbf{E}_{\boldsymbol{P}} [\ \boldsymbol{log}\ \boldsymbol{p}(\boldsymbol{z})]$$

When $\boldsymbol{p}(\boldsymbol{Z}) = \boldsymbol{h}(\boldsymbol{Z} \mid \boldsymbol{x}, \hat{\theta}_0)$ from above, then $\boldsymbol{F}(\theta, \boldsymbol{P}(\boldsymbol{Z})) = \boldsymbol{log}\ \mathrm{L}(\theta\ ; \boldsymbol{x})$.

Claim: For a fixed (arbitrary) value $\theta = \hat{\theta}_0$, $\boldsymbol{F}(\hat{\theta}_0, \boldsymbol{P}(\boldsymbol{Z}))$ is maximized over distributions $\boldsymbol{P}(\boldsymbol{Z})$ by choosing $\boldsymbol{p}(\boldsymbol{Z}) = \boldsymbol{h}(\boldsymbol{Z} \mid \boldsymbol{x}, \hat{\theta}_0)$.

Thus, the *EM* algorithm is a sequence of ***M***-***M*** steps: the old ***E***-step now is a max over the second term in $\boldsymbol{F}(\hat{\theta}_0, \boldsymbol{P}(\boldsymbol{Z}))$, given the first term. The second step remains (as in *EM*) a max over θ for a fixed second term, which does not involve θ

Suppose that the augmented data $\boldsymbol{Z}$ are multidimensional.

Consider the *GEM* approach and, instead of maximizing the choice of $\boldsymbol{P}(\boldsymbol{Z})$ over all of the augmented data – instead of the old *E*-step – instead maximize over only *one* coordinate of $\boldsymbol{Z}$ at a time, alternating with the (old) ***M***-step.

This gives us the following link with the Gibbs algorithm: Instead of maximizing at each of these two steps, use the conditional distributions, we sample from them!

Part 3) Generating a Random Variable

3.1) Continuous r.v.'s – an Exact Method using transformation of the CDF

- Let $\boldsymbol{Y}$ be a continuous r.v. with **cdf** $\boldsymbol{F_Y}(\bullet)$ Then the range of $\boldsymbol{F_Y}(\bullet)$ is (0, 1), and as a r.v. $\boldsymbol{F_Y}$ it is distributed $\boldsymbol{U} \sim$ Uniform (0,1). Thus the ***inverse*** tranformation $\boldsymbol{F_Y^{-1}}(\boldsymbol{U})$ gives us the desired distribution for $\boldsymbol{Y}$.

Examples:

- If Y ~ Exponential(λ) then $\boldsymbol{F_Y^{-1}}(\mathrm{U}) = -\lambda \, ln(1\text{-U})$ is the desired Exponential.

And from known relationships between the Exponential distribution and other members of the Exponential Family, we may proceed as follows.

Let $\boldsymbol{U}_j$ be *iid* Uniform(0,1), so that $Y_j = -\lambda \, ln(\boldsymbol{U}_j)$ are *iid* Exponential(λ)

- $\boldsymbol{Z} = -2 \sum_{j=1}^{n} ln(U_j) \sim \chi^2_{2n}$ a Chi-squared distribution on *2n degrees of freedom*
  Note only even integer values possible here, alas!

- $\boldsymbol{Z} = -\beta \sum_{j=1}^{a} ln(U_j) \sim$ Gamma $\Gamma(a, \beta)$ – with integer values only for a.

- $\boldsymbol{Z} = \dfrac{\sum_{j=1}^{a} \ln(U_j)}{\sum_{j=1}^{a+b} \ln(U_j)} \sim$ Beta(a,b) – with integer values only for a.

3.2) The "Accept/Reject" algorithm for approximations using pdf's.

Suppose we want to generate $\boldsymbol{Y} \sim \text{Beta}(a,b)$, for non-integer values of a and b, say $a = 2.7$ and $b = 6.3$.

Let $(\boldsymbol{U},\boldsymbol{V})$ be independent Uniform(0, 1) random variables. Let $\boldsymbol{c} \geq \max_y \boldsymbol{f_Y}(y)$

Now calculate $\boldsymbol{P}(\boldsymbol{Y} \leq y)$ as follows:

$$\boldsymbol{P}(\boldsymbol{V} \leq y, \boldsymbol{U} \leq (1/\mathbf{c})\, \boldsymbol{f_Y}(\boldsymbol{V})\,) = \int_0^y \int_0^{f_Y(v)/c} \boldsymbol{du\,dv}$$

$$= (1/\mathbf{c}) \int_0^y \boldsymbol{f_Y}(\boldsymbol{v})\boldsymbol{dv}$$

$$= (1/\mathbf{c})\, \boldsymbol{P}(\boldsymbol{Y} \leq y).$$

So: (i) generate independent $(\boldsymbol{U},\boldsymbol{V})$ Uniform(0,1)

(ii) If $\boldsymbol{U} < (1/\mathbf{c})\boldsymbol{f_Y}(\boldsymbol{V})$, set $\boldsymbol{Y} = \boldsymbol{V}$, otherwise, return to step (i).

Note: The waiting time for one value of $\boldsymbol{Y}$ with this algorthim is $\boldsymbol{c}$, so we want $\boldsymbol{c}$ small. Thus, choose $\boldsymbol{c} = \max_y \boldsymbol{f_Y}(y)$. But we waste generated values of $\boldsymbol{U},\boldsymbol{V}$ whenever $\boldsymbol{U} \geq (1/\mathbf{c})\boldsymbol{f_Y}(\boldsymbol{V})$, so we want to choose a better approximation distribution for $\boldsymbol{V}$ than the uniform.

Let $\boldsymbol{Y} \sim \boldsymbol{f}_{\mathrm{Y}}(y)$ and $\boldsymbol{V} \sim \boldsymbol{f}_{\mathrm{V}}(v)$.

- Assume that these two have common support, i.e., the smallest closed sets of measure one are the same.

- Also, assume that $\mathbf{M} = sup_{\mathrm{y}}\, [\boldsymbol{f}_{\mathrm{Y}}(y) \,/\, \boldsymbol{f}_{\mathrm{V}}(y)]$ exists, i.e., $\mathbf{M} < \infty$.

Then generate the *r.v.* $\boldsymbol{Y} \sim \boldsymbol{f}_{\mathrm{Y}}(y)$ using

$\boldsymbol{U} \sim$ Uniform(0,1) and $\boldsymbol{V} \sim \boldsymbol{f}_{\mathrm{V}}(v)$, with ($\boldsymbol{U}$, $\boldsymbol{V}$) independent, as follows:

(i) Generate values (u, v).

(ii) If $u < (1/\mathbf{M})\, \boldsymbol{f}_{\mathrm{Y}}(v) \,/\, \boldsymbol{f}_{\mathrm{V}}(y)$ then set $y = v$.

If not, return to step (i) and redraw (u,v).

Proof of correctness for the *accept/reject* algorithm:

The generated r.v. $\boldsymbol{Y}$ has a *cdf*

$$\boldsymbol{P}(\boldsymbol{Y} \le y) = \boldsymbol{P}(\boldsymbol{V} \le y \mid \text{stop})$$

$$= \boldsymbol{P}\,(\boldsymbol{V} \le y \mid \ \boldsymbol{U} < (1/\mathbf{M})\, \boldsymbol{f}_{\mathrm{Y}}(v) \,/\, \boldsymbol{f}_{\mathrm{V}}(y)\,)$$

$$= \frac{\boldsymbol{P}(\boldsymbol{V} \le \boldsymbol{y}, \boldsymbol{U} < (1/\boldsymbol{M})\, \boldsymbol{f}_{\boldsymbol{Y}}(\boldsymbol{V}) / \boldsymbol{f}_{\boldsymbol{V}}(\boldsymbol{V}))}{\boldsymbol{P}(\boldsymbol{U} < (1/\boldsymbol{M})\, \boldsymbol{f}_{\boldsymbol{Y}}(\boldsymbol{V}) / \boldsymbol{f}_{\boldsymbol{V}}(\boldsymbol{V}))}$$

$$= \frac{\int_{-\infty}^{\boldsymbol{y}} \int_0^{(1/\boldsymbol{M})\, \boldsymbol{f}_{\boldsymbol{Y}}(\boldsymbol{v}) / \boldsymbol{f}_{\boldsymbol{V}}(\boldsymbol{v})} \boldsymbol{du} \boldsymbol{f}_{\boldsymbol{V}}(\boldsymbol{v}) \boldsymbol{dv}}{\int_{-\infty}^{\infty} \int_0^{(1/\boldsymbol{M})\, \boldsymbol{f}_{\boldsymbol{Y}}(\boldsymbol{v}) / \boldsymbol{f}_{\boldsymbol{V}}(\boldsymbol{v})} \boldsymbol{du} \boldsymbol{f}_{\boldsymbol{V}}(\boldsymbol{v}) \boldsymbol{dv}}$$

$$= \int_{-\infty}^{\boldsymbol{y}} \boldsymbol{f}_{\boldsymbol{Y}}(\boldsymbol{v}) \boldsymbol{dv}.$$

Example: Generate $\boldsymbol{Y} \sim$ Beta(2.7,6.3).

Let $\boldsymbol{V} \sim$ Beta(2,6). Then $\mathbf{M} = 1.67$ and we may proceed with the algorithm.

3.3) Metropolis algorithm for "heavy-tailed" target densities.

As before, let $\boldsymbol{Y} \sim \boldsymbol{f}_{\mathrm{Y}}(y)$, $\boldsymbol{V} \sim \boldsymbol{f}_{\mathrm{V}}(v)$, $\boldsymbol{U} \sim$ Uniform(0,1), with $(\boldsymbol{U},\boldsymbol{V})$ independent.

Assume only that $\boldsymbol{Y}$ and $\boldsymbol{V}$ have a common support.

Metropolis Algorithm:

Step$_0$: Generate $\boldsymbol{v}_0$ and set $\boldsymbol{z}_0 = \boldsymbol{v}_0$. For $\boldsymbol{i} = 1, \ldots,$

Step$_{\mathbf{i}}$: Generate $(\boldsymbol{u}_i, \boldsymbol{v}_i)$

Define $$\rho_i = \min \left\{ \frac{f_Y(v_i)}{f_V(v_i)} \times \frac{f_V(z_{i-1})}{f_Y(z_{i-1})}, 1 \right\}$$

Let $$\mathbf{z}_i = \begin{cases} \mathbf{v}_i & \text{if } \mathbf{u}_i \le \rho_i \\ \mathbf{z}_{i-1} & \text{if } \boldsymbol{u}_i > \rho_i. \end{cases}$$

Then, as i $\to\infty$, the *r.v.* $\boldsymbol{Z}_i$ converges in distribution to the random variable $\boldsymbol{Y}$.

Additional References

Casella, G. and George, E. (1992) "Explaining the Gibbs Sampler," *Amer. Statistician* **46**, 167-174.

Flury, B. and Zoppe, A. (2000) "Exercises in EM," *Amer. Staistican* **54**, 207-209.

Hastie, T., Tibshirani, R, and Friedman, J. *The Elements of Statistical Learning*. New York: Spring-Verlag, 2001, sections 8.5-8.6.

Tierney, L. (1994) "Markov chains for exploring posterior distributions" (with discussion) *Annals of Statistics* **22**, 1701-1762.