Gibbs sampling (an MCMC method) and relations 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 – a again (These notes will follow as a separate file.)
   - EM as a maximization/maximization method
   - Gibbs as a variation of Generalized EM
     - o an example
   - A counterexample for EM

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

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

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

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

If we can sample from the marginal distribution, then

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

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:

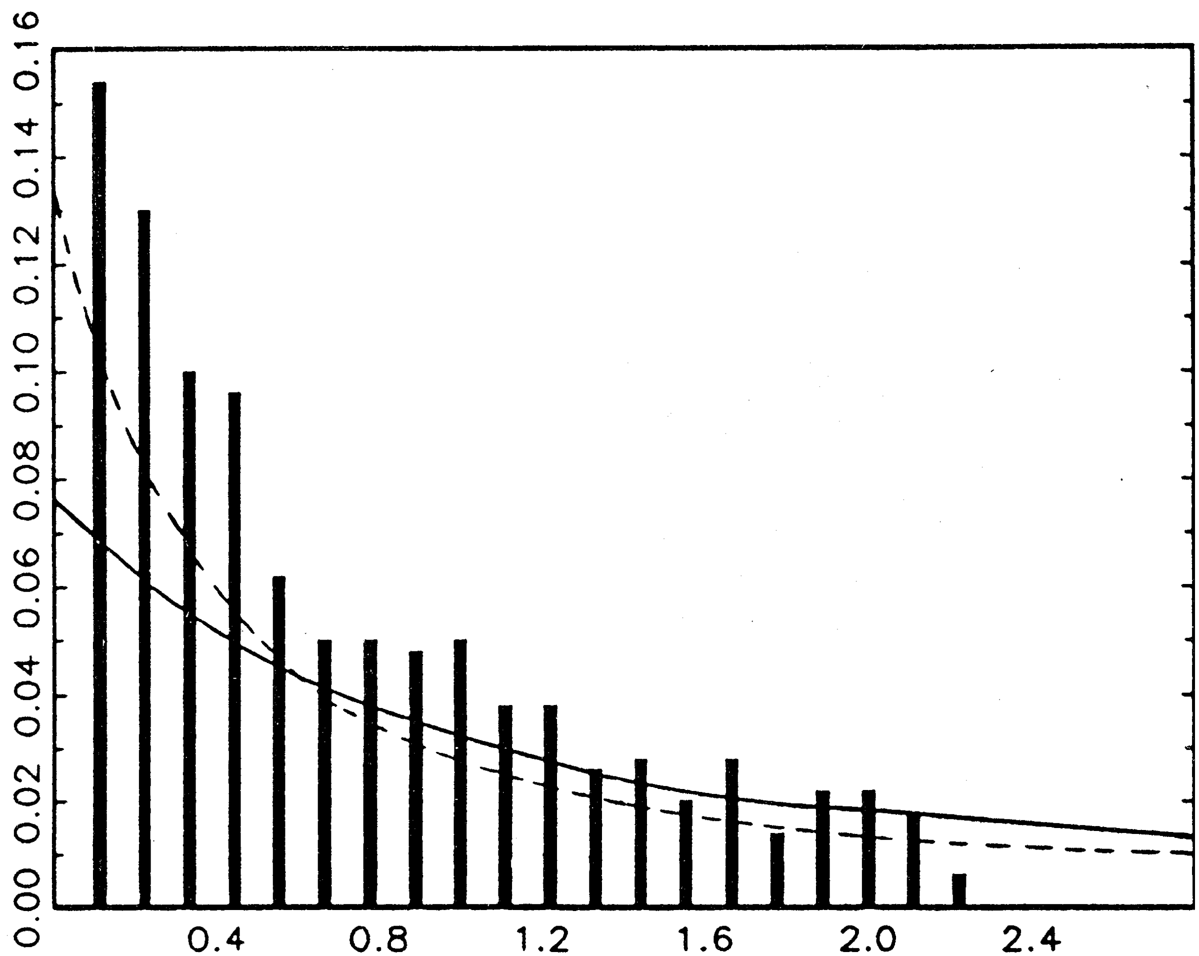
$$f(X \mid y) \propto ye^{-yx} \text{ for } 0 < \mathrm{X} < \boldsymbol{b}$$

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

where ***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, $x_i^{n_i}$ ($i = 1, \ldots, 500$, $n_i = 15$). 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):

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

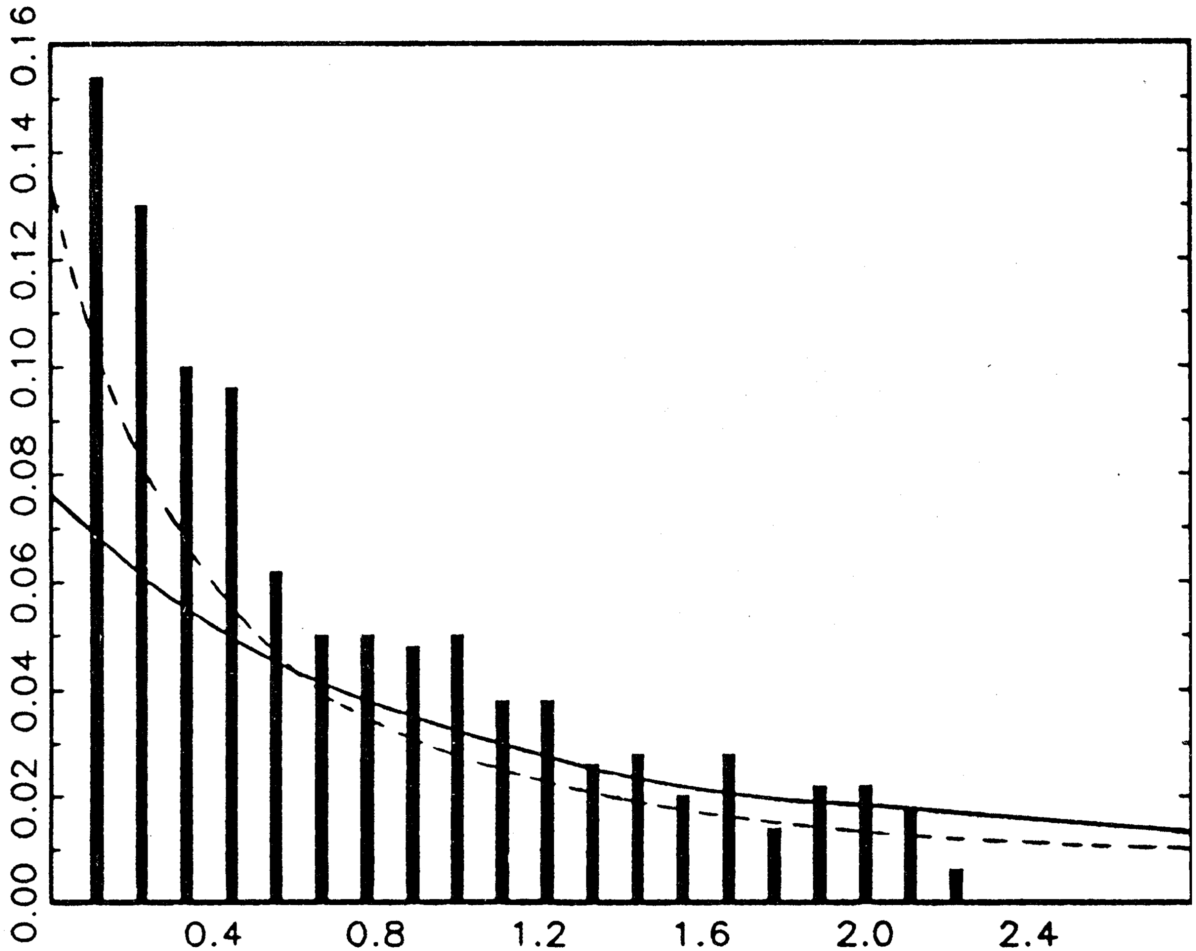
Thus $$E[f(x|Y)] = \int f(x \mid y) f(y) dy = f(x).$$

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_1 = 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

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

where $p_i > 0$, $\sum p_i = 1$, 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:

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

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

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

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

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:

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

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

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

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

With the transition probabilities positive, it is an (old) ergodic theorem that, $\mathbf{P}(X^n)$ 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.

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

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

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

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

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

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

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

$$= \quad \mathbf{E}_{\mathbf{P}}\,[X^n{=}\,0\,]$$

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

The *Ergodic* Theorem:

Definitions:

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

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

- 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 **A** having positive **P**-probability, with positive P-probability the chain visits **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 $j = 1, \ldots, k$-1 (mod 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. $lim_{n\to\infty}\, sup_{\mathbf{A}}\, |\, \mathbf{P}(X_{\mathrm{n}} \in \mathbf{A} \mid X_0 = x_0\,) - \int_{\mathbf{A}} \boldsymbol{f}(x)\, dx\ \,| = 0$

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:

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

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

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 $\Gamma(\alpha, \beta)$.

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

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

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

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!

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,