Auxiliary variable methods

The Gibbs sampler is readily implementable for the Ising model, but can be extremely slow to converge to the stationary distribution. This is due in part to high dependence between the variables.

One method for speeding up simulation of variates from the Ising model is to introduce *auxiliary* or *latent* variables that extend the state-space of the Markov chain in such a way that the full-conditionals remain tractable, but which help to break the dependence.

The most common latent variable method for the Ising model is the *Swendsen-Wang algorithm*.

Auxiliary variable methods

The general principle is as follows. Consider target distribution $\pi_X(x)$ for random variables X, and let U be a collection of random variables such that the joint distribution $\pi_{X,U}(x,u)$ has marginal $\pi_X(x)$

$$\pi_X(x) = \int \pi_{X,U}(x,u)\, du$$

In a Gibbs sampler context, we could consider constructing a Markov chain using the two conditional distributions

$$\pi_{X|U}(x|u) \qquad\qquad \pi_{U|X}(u|x)$$

and then collecting the samples for X, which are necessarily samples from $\pi_X(x)$.

Auxiliary variable methods

Note that

$$\pi_{X,U}(x,u) = \pi_X(x)\pi_{U|X}(u|x)$$

We are free to select the U variables in whichever way is appropriate. We seek to introduce U such that

- the full conditionals are straightforward to sample
- the dependence structure amongst the X variables is broken.

For example, suppose

$$\pi_X(x) \propto \pi_X^{(0)}(x) \prod_{k=1}^{K} b_k(x)$$

where $\pi_X^{(0)}(x)$ can be sampled in straightforward fashion. Let

$$U_k | X = x \sim \mathit{Uniform}(0, b_k(x)) \qquad k = 1, \ldots, K$$

be a collection of conditionally independent random variables.

$$\begin{aligned}
\pi_{X,U}(x,u) &= \pi_X(x)\pi_{U|X}(u|x) \\
&= \left\{\pi_X^{(0)}(x)\prod_{k=1}^{K} b_k(x)\right\}\left\{\prod_{k=1}^{K}\frac{1}{b_k(x)}\mathbb{1}_{[0,b_k(x)]}(u_k)\right\} \\
&= \left\{\pi_X^{(0)}(x)\right\}\left\{\prod_{k=1}^{K}\mathbb{1}_{[0,b_k(x)]}(u_k)\right\}
\end{aligned}$$

Therefore

$$\pi_{X|U}(x|u) \propto \pi_X^{(0)}(x)\prod_{k=1}^{K}\mathbb{1}_{[u_k,\infty)}(b_k(x))$$

that is, proportional to $\pi_X^{(0)}(x)$ with support restricted by

$$b_k(x) \geqslant u_k.$$

for $k = 1, \ldots, K$.

Example (Swendsen-Wang Algorithm)

In the Ising model

$$\pi_X(x) \propto \exp\left\{\beta \sum_l \sum_{m \in \partial l} \mathbb{1}_{\{x_m\}}(x_l)\right\}.$$

where ∂l constitute the neighbours of site l. In the simplest case, if site l is at Cartesian location (i, j), the neighbours are the four sites immediately

- to the left $(i-1, j)$
- to the right $(i+1, j)$
- above $(i, j+1)$
- below $(i, j-1)$

site (i, j).

Example (Swendsen-Wang Algorithm)

The full conditional distributions for the Gibbs sampler are given by

$$X_l | X_{\partial l} = x_{\partial l} \sim \textit{Bernoulli}(\theta(\beta; x_{\partial l}))$$

that is

$$\mathbb{P}[X_l = 1 | X_{\partial l} = x_{\partial l}] = \theta(\beta; x_{\partial l})$$

where

$$\theta(\beta; x_{\partial l}) = \frac{\exp\{\beta N_{l1}\}}{\exp\{\beta N_{l0}\} + \exp\{\beta N_{l1}\}}$$

This algorithm updates one x value at a time, either in a *fixed* or *random* scan fashion.

Example (Swendsen-Wang Algorithm)

The Swendsen-Wang algorithm introduces a collection of auxiliary bond variables, u. Consider two neighbouring sites, l and m, say. Let $\omega = 1 - e^{-\beta}$. Define the binary bond random variable U_{lm}, conditional on X_l and X_m, to have distribution described by

$$\mathbb{P}[U_{lm} = 1 | X_l = x_l, X_m = x_m] = \begin{cases} \omega & x_l = x_m \\ 0 & x_l \neq x_m \end{cases}$$

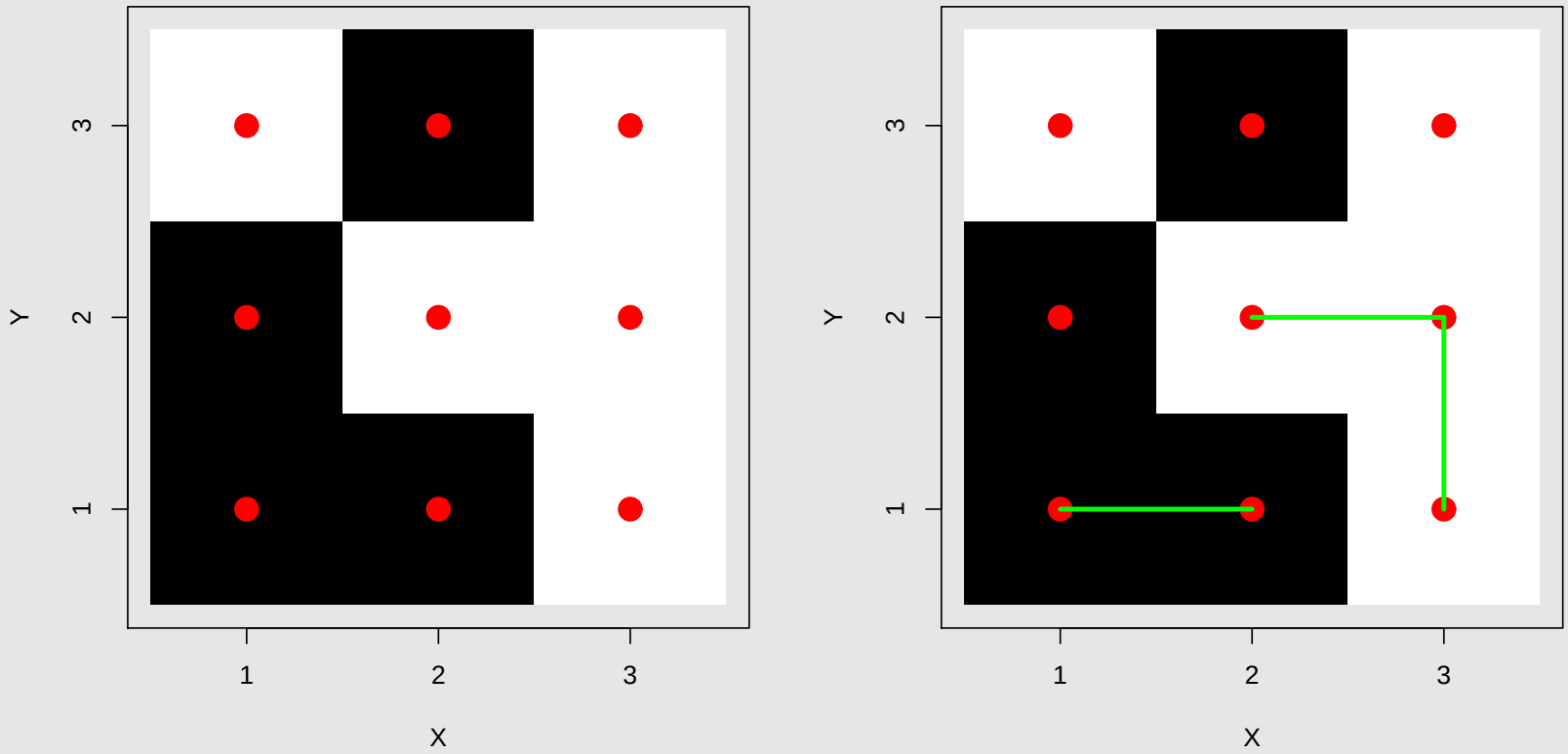
such that the collection of U variables are conditionally independent given X variables.

The bond is *present* if $U_{lm} = 1$, and *absent* otherwise.

Auxiliary variable methods

Example (Swendsen-Wang Algorithm)

Bond variables (green) between adjacent, similar x values.



Example (Swendsen-Wang Algorithm)

Consider the joint specification $\pi_{X,U}$ given by

$$\pi_{X,U}(x,u) \propto \prod_{l,m} \left\{ \omega \delta_{x_l x_m} \delta_{u_{lm} 1} + (1-\omega) \delta_{u_{lm} 0} \right\}$$

where the product is over all pairs of neighbouring sites, and $\delta_{ab} = 1$ if $a = b$, and zero otherwise. We have

$$\pi_{U_{lm}|X}(u|x) = \pi_{U_{lm}|X_l,X_m}(u|x_l,x_m) = \begin{cases} \omega^u (1-\omega)^{1-u} & x_l = x_m \\ \mathbb{1}_{\{0\}}(u) & x_l \neq x_m \end{cases}$$

for $u = 0, 1$.

Example (Swendsen-Wang Algorithm)

Also marginalizing over the u variables by summing out over $u_{lm} = 0, 1$ can be achieved over each term in the product separately. We have

$$\begin{aligned}\sum_u \pi_{X,U}(x,u) &\propto \prod_{l,m} \left\{(1-\omega) + \omega\delta_{x_l x_m}\right\} \\ &= \prod_{l,m} \left\{e^{-\beta} + (1-e^{-\beta})\delta_{x_l x_m}\right\}.\end{aligned}$$

Multiplying through each term by $e^{-\beta}$ yields

$$\sum_u \pi_{X,U}(x,u) \propto \prod_{l,m} \left\{1 + (e^{\beta}-1)\delta_{x_l x_m}\right\}.$$

Example (Swendsen-Wang Algorithm)

For each term in the product, when $x_l \neq x_m$, the term contributes 1; when $x_l = x_m$, the term contributes e^{β}. Hence

$$\sum_u \pi_{X,U}(x,u) \propto \prod_{l,m} \exp\{\beta \delta_{x_l x_m}\} = \exp\left\{\beta \sum_{l,m} \delta_{x_l x_m}\right\} = \pi_X(x)$$

and so the correct marginal model is recovered.

Example (Swendsen-Wang Algorithm)

Marginalizing over the X variables, we have

$$\begin{aligned}\sum_x \pi_{X,U}(x,u) &\propto \sum_x \prod_{l,m} \left\{\omega \delta_{x_l x_m} \delta_{u_{lm}1} + (1-\omega)\delta_{u_{lm}0}\right\} \\ &\propto \omega^{n(u)}(1-\omega)^{n_b - n(u)}\end{aligned}$$

where $n(u)$ is the total number of bonds present

$$n(u) = \sum_{l,m} u_{lm}$$

and n_b is the total number of bonds possible.

Example (Swendsen-Wang Algorithm)

We now consider running a Gibbs sampler for $\pi_{X,U}$.

- The full conditional distribution of $U|X = x$ is determined by the definition above, that is,

$$\pi_{U|X}(u|x) = \prod_{l,m} \pi_{U_{lm}|X_l,X_m}(u_{lm}|x_l, x_m)$$

where

$$\pi_{U_{lm}|X_l,X_m}(u_{lm}|x_l, x_m) = \begin{cases} \omega^u(1-\omega)^{1-u} & x_l = x_m \\ \mathbb{1}_{\{0\}}(u) & x_l \neq x_m \end{cases}$$

for $u = 0, 1$, with the product being over all neighbouring sites. Thus the U variables can be updated independently, conditional on X.

Example (Swendsen-Wang Algorithm)

- The full conditional distribution of $X|U$ is defined by the fact that

$$\pi_{X|U}(x|u) \propto \pi_{X,U}(x,u) \propto \prod_{l,m} \left\{ \omega \delta_{x_l x_m} \delta_{u_{lm} 1} + (1-\omega) \delta_{u_{lm} 0} \right\}$$

  and that bonds link "clusters" of sites that have the same value. The conditional distribution does not depend on the **value** of each x_l, merely whether or not $x_l = x_m$.

  Conditional on u, clusters can assume the label 0 or 1, **independently** with **equal probability**.

Example (Swendsen-Wang Algorithm)

Thus the Gibbs sampler can be implemented by sampling from the two full conditional distributions

$$\pi_{X|U}(x|u) \qquad\qquad \pi_{U|X}(u|x)$$

recursively, both of which involve sampling from trivial distributions.

There is some computational overhead in implementing Swendsen-Wang, in particular, tracing the clusters.

Example (Swendsen-Wang Algorithm)

For a specific computational example: with grid size $N = 128$, $\beta = 0.85$, the Gibbs sampler and Swendsen-Wang (SW) algorithms were run in R for 500 iterations.

- Gibbs sampler: 11 minutes, 30 seconds,
- Swendsen-Wang: 30 minutes, 3 seconds.

that is, roughly 2.5 slower for SW. However, the SW chain converges much more quickly; the Gibbs sampler chain exhibits slower convergence for larger values of β.

Auxiliary variable methods

Example (Swendsen-Wang Algorithm)

With $\beta = 0.85$, the performance of the two methods is not appreciably different; when $\beta = 0.95$, the behaviour is radically different.

The Ising model exhibits *critical behaviour* when $\beta = 1$: for $\beta > 1$, entire configurations of one "colour" occur with high probability.

As $\beta \uparrow 1$, the Gibbs sampler exhibits increasingly poor behaviour ("critical slowing"), whereas Swendsen-Wang does not.

Example (Swendsen-Wang Algorithm)

To demonstrate this, we track the summary statistic

$$T(x) = \sum_{l} \sum_{m \in \partial l} \mathbb{1}_{\{x_m\}}(x_l)$$

across iterations. This is the number of neighbours that are alike in the configuration.

We do this for the three cases

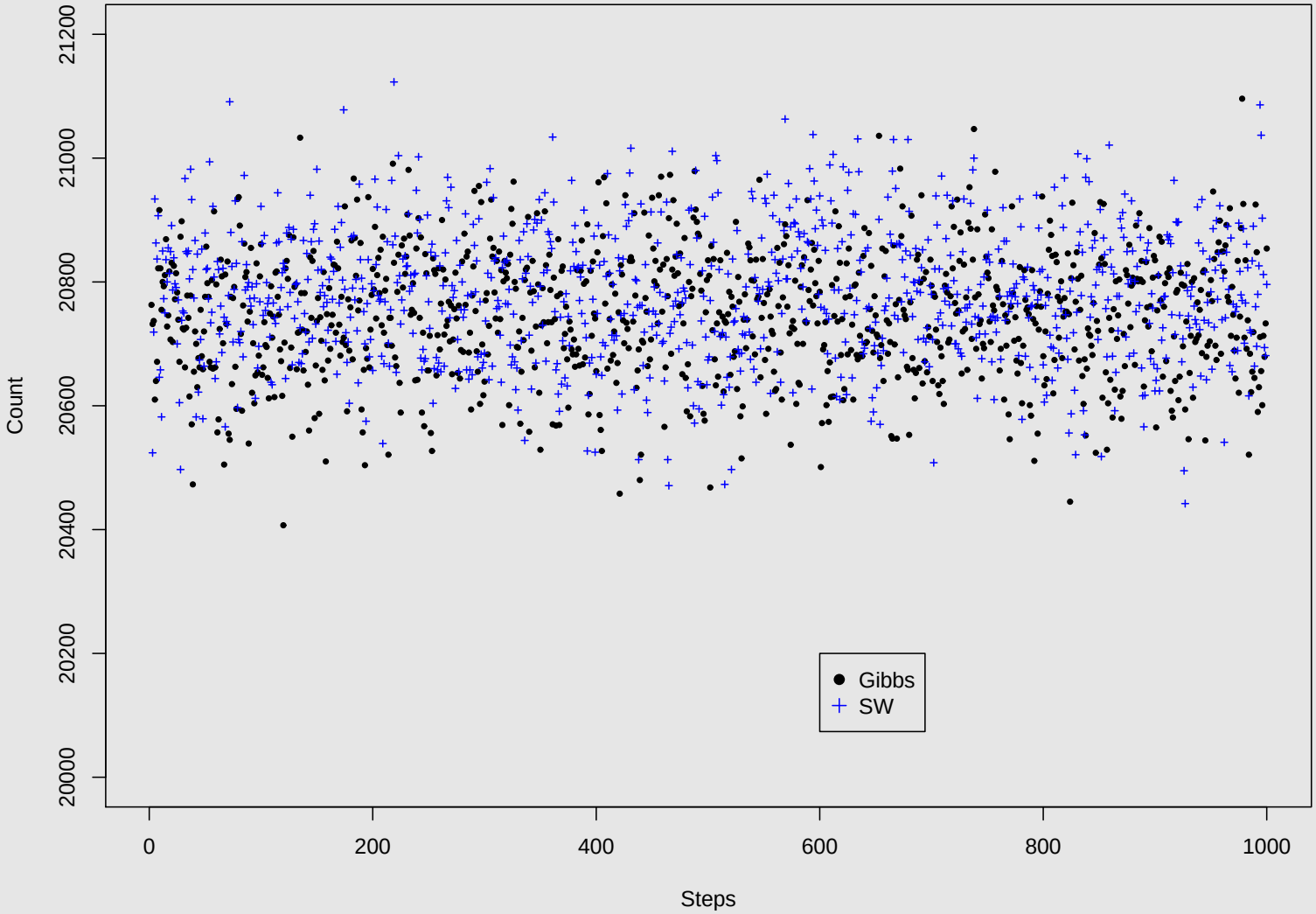
(i) $\beta = 0.5$

(ii) $\beta = 0.85$

(iii) $\beta = 0.95$.

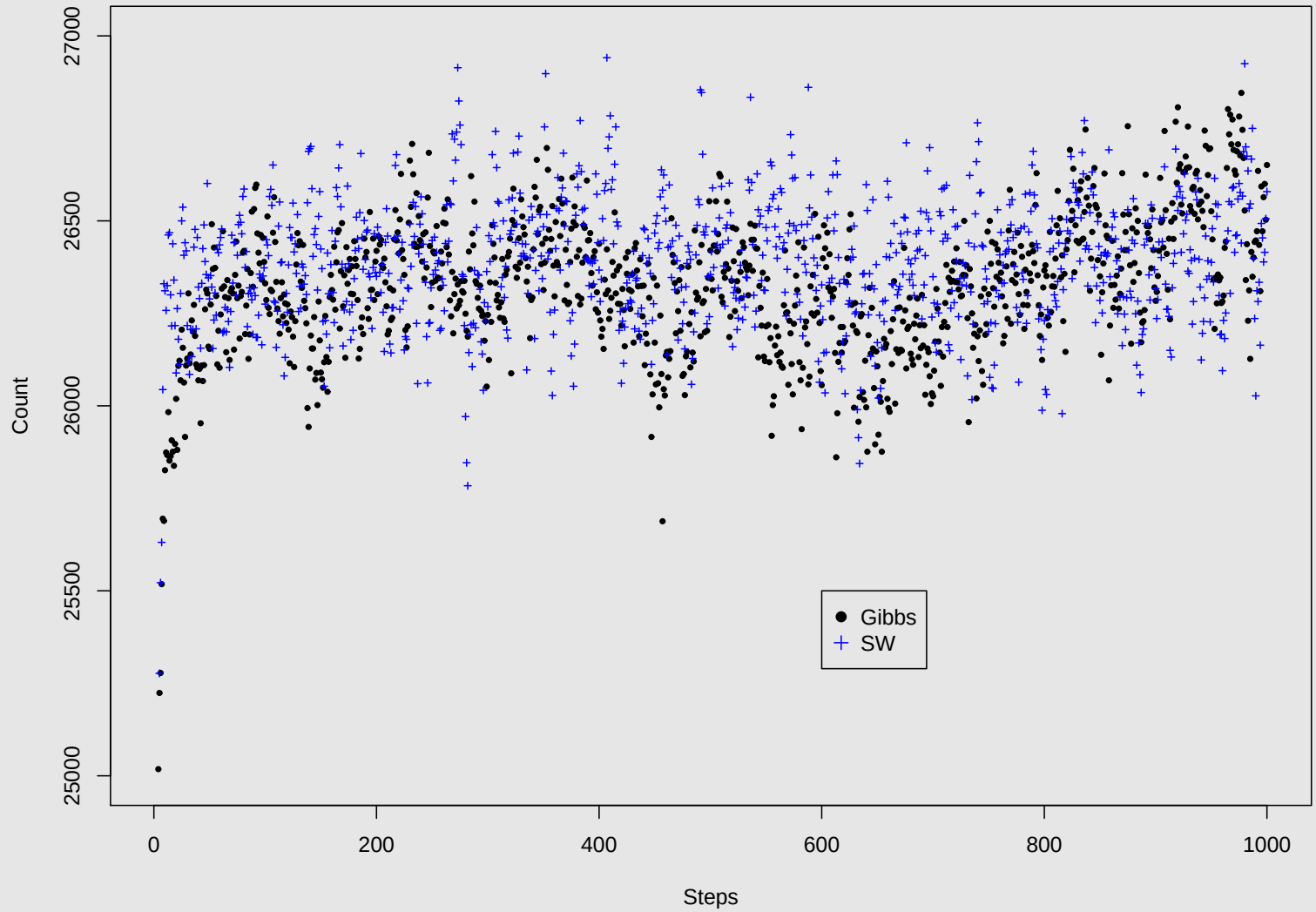
Example (Swendsen-Wang Algorithm)

$\beta = 0.5$



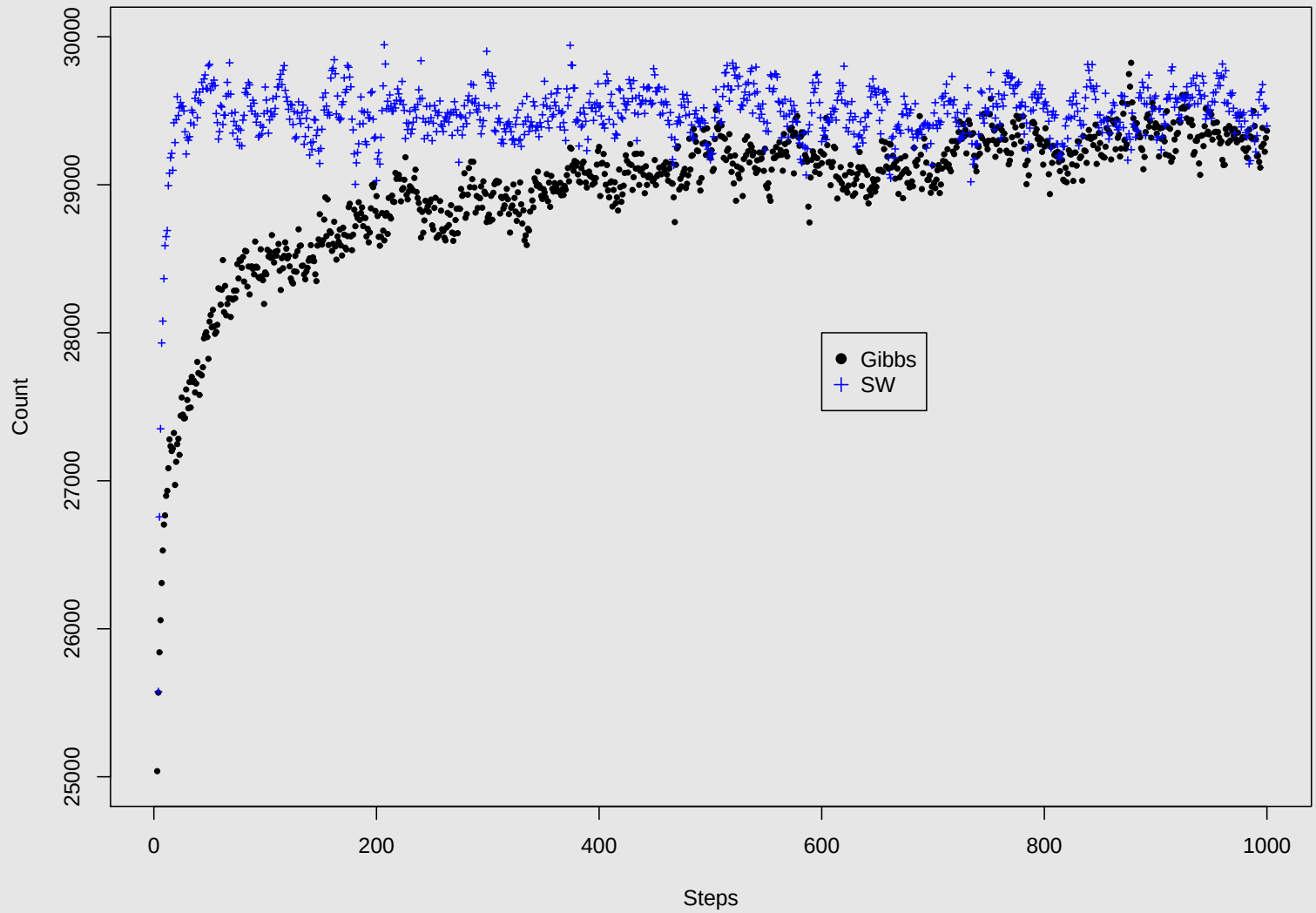
Example (Swendsen-Wang Algorithm)

$\beta = 0.85$



Example (Swendsen-Wang Algorithm)

$\beta = 0.95$



Auxiliary variable methods: Mixture Models

Example (Bayesian analysis of Finite Mixture Models)

Suppose that $Y_1 \ldots, Y_n$ are a random sample from the K component finite mixture model for $y \in \mathbb{R}$

$$f_{Y|\theta}(y|\theta) = \sum_{k=1}^{K} \omega_k f_k(y|\theta_k) \qquad \sum_{k=1}^{K} \omega_k = 1$$

for $0 < \omega_k < 1$ where $f_1, \ldots, f_K$ are component densities. The likelihood is

$$L_Y(\theta; \mathbf{y}) = \prod_{i=1}^{n} \left\{ \sum_{k=1}^{K} \omega_k f_k(y_i|\theta_k) \right\}$$

For a Bayesian analysis, we specify a prior distribution

$$p(\theta) = p(\omega_1, \ldots, \omega_K, \theta_1, \ldots, \theta_K)$$

Auxiliary variable methods: Mixture Models

Example (Bayesian analysis of Finite Mixture Models)

Here

$$\pi(\theta|\boldsymbol{y}) \propto L_Y(\theta;\boldsymbol{y})p(\theta).$$

Consider the discrete auxiliary variables $U_1, \ldots, U_n$ introduced into the likelihood density for each i via the identity

$$f_{Y|\theta}(y|\theta) = \sum_{k=1}^{K} f_{Y|U,\theta}(y|u,\theta) f_U(u)$$

where U has the $Multinomial(1, \omega_1, \ldots, \omega_K)$ distribution

$$f_U(u) = \prod_{k=1}^{K} \omega_k^{\mathbb{1}_{\{k\}}(u)}$$

and $f_{Y|U,\theta}(y|u,\theta) \equiv f_u(y|\theta)$.

Auxiliary variable methods: Mixture Models

Example (Bayesian analysis of Finite Mixture Models)

The U variables are discrete labels indicating which of the component densities, $f_1, \ldots, f_K$, each data point is sampled from, with

$$\mathbb{P}[U = k] = \omega_k \qquad k = 1, \ldots, K$$

Consider the joint posterior distribution on the unknown (unobservable) quantities

$$\pi_{\theta,\boldsymbol{u}}(\theta, \boldsymbol{u}|\boldsymbol{y}) \propto L(\boldsymbol{y}; \theta, \boldsymbol{u}) p(\theta, \boldsymbol{u})$$

Here

$$p(\theta, \boldsymbol{u}) = p(\theta) p(\boldsymbol{u}|\theta) = p(\theta) \prod_{i=1}^{n} \left\{ \prod_{k=1}^{K} \omega_k^{\mathbb{1}_{\{k\}}(u)} \right\}$$

where $p(\theta)$ is a joint prior on the ω and θ variables.

Auxiliary variable methods: Mixture Models

Example (Bayesian analysis of Finite Mixture Models)

Running the Gibbs sampler on the full conditional distributions

$$\pi_{\theta|\boldsymbol{U}}(\theta|\boldsymbol{u},\boldsymbol{y}) \qquad\qquad \pi_{\boldsymbol{U}|\theta}(\boldsymbol{u}|\theta,\boldsymbol{y})$$

yields samples from the joint posterior. Marginally, the Gibbs sampler produces a sample from

$$\pi_{\theta}(\theta|\boldsymbol{y})$$

which is the target posterior of interest.

For the first conditional, given $\boldsymbol{U} = \boldsymbol{u}$, we have a factorization

$$\pi_{\theta|\boldsymbol{U}}(\theta|\boldsymbol{u},\boldsymbol{y}) = \pi_{\theta_1,\ldots,\theta_K|\boldsymbol{U}}(\theta_1,\ldots,\theta_K|\boldsymbol{u},\boldsymbol{y})\pi_{\omega|\boldsymbol{U}}(\omega_1,\ldots,\omega_K|\boldsymbol{u},\boldsymbol{y})$$

and can update these parameters independently in two parts.

Auxiliary variable methods: Mixture Models

Example (Bayesian analysis of Finite Mixture Models)

- $\pi_{\theta|\boldsymbol{u}}(\theta|\boldsymbol{u},\boldsymbol{y})$ (Part I): if the prior on the θ parameters factorizes into independent components, then

$$\pi_{\theta_1,\ldots,\theta_K|\boldsymbol{u}}(\theta_1,\ldots,\theta_K|\boldsymbol{u},\boldsymbol{y}) = \prod_{k=1}^{K} \pi_{\theta_k|\boldsymbol{u}}(\theta_k|\boldsymbol{u},\boldsymbol{y})$$

where

$$\pi_{\theta_k|\boldsymbol{u}}(\theta_k|\boldsymbol{u},\boldsymbol{y}) \propto \left\{ \prod_{i:u_i=k} f_k(y_i|\theta_k) \right\} p(\theta_k)$$

and we can update the θ parameters independently conditional on $\boldsymbol{u}$. This may require a Metropolis-Hastings update.

Auxiliary variable methods: Mixture Models

Example (Bayesian analysis of Finite Mixture Models)

- $\pi_{\theta|\boldsymbol{U}}(\theta|\boldsymbol{u},\boldsymbol{y})$ (Part II): a standard choice for the prior for the ω parameters is the Dirichlet prior

$$p(\omega_1,\ldots,\omega_K) \propto \prod_{k=1}^{K} \omega_k^{\alpha_k-1}$$

$$\Longrightarrow p(\omega_1,\ldots,\omega_K|\boldsymbol{u},\boldsymbol{y}) \propto \prod_{k=1}^{K} \omega_k^{n_k+\alpha_k-1}$$

where, for $k = 1,\ldots,K$, $n_k = \sum_{i=1}^{n} \mathbb{1}_{\{k\}}(u_i)$ is the count of the U variables that are set equal to k. Thus the full conditional distribution is also Dirichlet, and can be sampled directly.

Auxiliary variable methods: Mixture Models

Example (Bayesian analysis of Finite Mixture Models)

- $\pi_{\boldsymbol{U}|\theta}(\boldsymbol{u}|\theta, \boldsymbol{y})$: by conditional independence of the Y given θ, we must have

$$\pi_{\boldsymbol{U}|\theta}(\boldsymbol{u}|\theta, \boldsymbol{y}) = \prod_{i=1}^{n} \pi_{U_i|\theta}(u_i|\theta, y_i)$$

where, by Bayes theorem

$$\pi_{U_i|\theta}(u_i|\theta, y_i) = \frac{f_{Y_i}(y_i|u_i, \theta) f_{U_i}(u_i|\theta)}{\sum_{k=1}^{K} f_{Y_i}(y_i|k, \theta) f_{U_i}(k|\theta)} = \frac{f_{u_i}(y_i|\theta_{u_i})\omega_{u_i}}{\sum_{k=1}^{K} f_k(y_i|\theta_k)\omega_k}$$

which is a discrete distribution on $\{1, \ldots, K\}$. Thus updating the U variables is straightforward.

Auxiliary variable methods: Mixture Models

Example (Bayesian analysis of Finite Mixture Models)

Note: this formulation reveals a connection with the EM algorithm; the full conditional distribution of U_i is identical to the conditional distribution of the missing data, given the observed data (see p. 128).

Indeed, the entire formulation is similar, the only difference is that we replace the Expectation and Maximization steps by sampling steps. The Gibbs sampler does not perform maximization to get the ML estimate, it produces samples from the posterior distribution of the parameters.

Auxiliary variable methods: Mixture Models

Example (Bayesian analysis of Finite Mixture Models)

Some care is needed in interpreting the Gibbs sampler output, as the component labels are, in fact, arbitrarily assigned in the mixture formulation. Therefore, the U variables can exhibit *label-switching*.

A simple solution is to add an additional Metropolis-Hastings step at the end of each Gibbs sampler iteration that permutes the labels on the U variables, that is, for $K = 3$, if the random permutation is $(3, 1, 2)$, then

- All Us labelled 1 are relabelled 3
- All Us labelled 2 are relabelled 1
- All Us labelled 3 are relabelled 2

This Metropolis-Hastings proposal is **always accepted**, as it does not change the posterior density value.

Auxiliary variable methods: Missing data

Auxiliary variable methods can also be used to handle *missing data* problems. Missing data can arise in problems due to

- drop-out
- data handling issues
- censoring/truncation

The missingness can occur

- *completely at random* (independently of the observed and unobserved data)
- *at random* (dependent on the observed data, but independent of the unobserved data)
- *not at random* (dependent on the observed data and unobserved data)

Auxiliary variable methods: Missing data

Missingness can occur in the

- *response* variable Y
- *covariate* X

We will denote the missing values by $X^\star$ and $Y^\star$.

We consider a Bayesian formulation of the missing data inference problem; maximum likelihood estimation can be handled using the EM algorithm. We consider the joint likelihood of the observed and unobserved data, and treat the unobserved values as quantities to be updated during the MCMC.

The parameter vector θ can now also include parameters in the stochastic missingness mechanism.

Auxiliary variable methods: Missing data

In the case of i.i.d. data, where the missing data are considered independent of the observed data, the **complete** data $(\boldsymbol{Y}, \boldsymbol{Y}^\star)$ have likelihood

$$L(\boldsymbol{y}, \boldsymbol{y}^\star; \theta) = \left\{ \prod_{i=1}^{n} f_Y(y_i; \theta) \right\} \left\{ \prod_{i=1}^{n^\star} f_Y(y_i^\star; \theta) \right\}$$

In this case, a Gibbs sampler can be used to update the quantities $(\theta, \boldsymbol{Y}^\star)$ from their full conditionals. For θ,

$$\pi_{\theta|\boldsymbol{Y},\boldsymbol{Y}^\star}(\theta|\boldsymbol{y}, \boldsymbol{y}^\star) \propto L(\boldsymbol{y}, \boldsymbol{y}^\star; \theta) p(\theta)$$

that is, we perform Bayesian inference using the complete data likelihood.

Auxiliary variable methods: Missing data

For elements of $\boldsymbol{Y}^\star$, the full conditional distribution is merely the density $f_Y(y_i^\star;\theta)$. Most typically, however, this kind of missing data occurs due to censoring, that is, we have information that

$$Y_i^\star > c_i$$

say. In this case, the full conditional distribution used in the Gibbs sampler must respect the censoring constraint, that is, the full conditional is

$$f_Y(y_i^\star;\theta,c_i) \propto \mathbb{1}_{(c_i,\infty)}(y_i^\star) f_Y(y_i^\star;\theta)$$

Conditional on $(\boldsymbol{y},\boldsymbol{y}^\star)$, the full conditional distribution of θ is not affected by the censoring, so is identical to above.

Auxiliary variable methods: Missing data

For missing covariate data in a regression setting, sampling the missing covariates involves an extra component. Consider the full conditional distribution for missing covariates $\boldsymbol{X}^\star$

$$\pi_{\boldsymbol{X}^\star|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Y}^\star,\theta}(\boldsymbol{x}^\star|\boldsymbol{x},\boldsymbol{y},\boldsymbol{y}^\star,\theta)$$

By Bayes theorem, this is proportional to

$$f_{\boldsymbol{Y},\boldsymbol{Y}^\star|\boldsymbol{X},\boldsymbol{X}^\star,\theta}(\boldsymbol{y},\boldsymbol{y}^\star|\boldsymbol{x},\boldsymbol{x}^\star,\theta)\, f_{\boldsymbol{X}^\star|\boldsymbol{X},\theta}(\boldsymbol{x}^\star|\boldsymbol{x},\theta)$$

The first term is the complete data likelihood for a regression problem. The second term most typically simplifies to

$$f_{\boldsymbol{X}^\star|\theta}(\boldsymbol{x}^\star|\theta)$$

under the assumption of random sampling of the response/covariate pairs.