GENOME 541 Section 2

Lecture 3

Gibbs Sampling

Yi Yin (yy224[at]uw.edu)
Shendure Lab
Department of Genome Sciences
University of Washington
Tuesday, April 18, 2017

Review

- Posterior and prior predictive distribution (HW3)
- One parameter model and conjugacy

Prior: Ga(a, b)

$$\theta \sim Ga(a,b), \text{ where the Ga(a,b) pdf is } \frac{b^a}{\Gamma(a)}\theta^{a-1}exp(-b\theta)$$

$$E(\theta) = a/b, V(\theta) = a/b^2$$

Likelihood: Poisson(y;θ)

Posterior: $\theta|y_1,...,y_n \sim Ga(\widehat{a},\widehat{b})$ $Gamma(a+\sum y_i, b+n)$

can be read off directly from the kernel of the distribution

Now, suppose you did two experiments and collected two sets of counts, $Y^{(1)}$ and $Y^{(2)}$

$Y^{(1)}$: {1, 7, 10, 4, 3, 5, 8, 4, 6, 5}

$Y^{(2)}$: {7, 10, 6, 6, 9, 4, 7, 9, 10, 6}

How to obtain posterior summaries?

Under Ga(1,1) prior (weekly informative, with prior sample size 1), our posterior distributions are

$$\theta_1|Y^{(1)} \sim Ga(54, 11) \text{ with } \widehat{\theta_1}|Y^{(1)} = 4.9$$
$$\theta_2|Y^{(2)} \sim Ga(75, 11) \text{ with } \widehat{\theta_2}|Y^{(2)} = 6.8$$

Note that this is a bit biased, alternatively, you can center your prior mean on the MLE.

But what if we are interested more than a single point estimation like the posterior mean? For example, can we directly obtain a probability of $p(\widehat{\theta_1} < \widehat{\theta_2}|Y^{(1)}, Y^{(2)})$, or, can we alternatively obtain a predictive probability of $p(y_{n+1}^{(1)} < y_{n+1}^{(2)}|Y^{(1)}, Y^{(2)})$? (HW3 Problem 3)

Back to the posterior predictive distribution

For the Gamma-Poisson example, the posterior predictive distribution is in closed form, which is a negative binomial distribution.

Let's derive this (for short, we write $y^n=\{y_1, \ldots, y_n\}$)

General form: $$p(Y^* \mid Y) = \int p(Y^* \mid \theta) p(\theta \mid Y) d\theta$$

$$\begin{aligned}
f(y|y^n) &= \int \text{Pois}(y;\theta)\text{Ga}(\theta;\hat{a},\hat{b})d\theta. \\
&= \frac{\hat{b}^{\hat{a}}}{y!\Gamma(\hat{a})} \int \theta^{\hat{a}+y-1} \exp\{-\theta(\hat{b}+1)\} d\theta \\
&= \frac{\hat{b}^{\hat{a}}}{y!\Gamma(\hat{a})} \frac{\Gamma(\hat{a}+y)}{(1+\hat{b})^{\hat{a}+y}} \\
&= \frac{\Gamma(\hat{a}+y)}{\Gamma(y+1)\Gamma(\hat{a})} \left(\frac{\hat{b}}{\hat{b}+1}\right)^{\hat{a}} \left(\frac{1}{\hat{b}+1}\right)^{y},
\end{aligned}$$

which is neg-binomial$(\hat{a}, \hat{b}/(1+\hat{b}))$.

Negative binomial distribution

- Negative binomial distribution is an over-dispersed generalization of the Poisson

- When you marginalize θ out of the Poission (y; θ) likelihood over a gamma distribution, you can obtain a negative-binomial

	pdf	Support	Mean	Var
$NB(\alpha, p)$	$f(x) = \binom{x+\alpha-1}{x} p^{\alpha} q^{x}$	$x \in \mathbb{Z}_{+}$	$\alpha q / p$	$\alpha q / p^2$

- For $(y|y^n) \sim$ neg-binomial$(\hat{a}, \hat{b}/(1+\hat{b}))$, we have

$$\begin{aligned} \mathrm{E}(y|y^n) &= \frac{\hat{a}}{\hat{b}} = \mathrm{E}(\theta|y^n) = \text{Posterior mean} \\ \mathrm{V}(y|y^n) &= \frac{\hat{a}(\hat{b}+1)}{\hat{b}^2} = \mathrm{E}(\theta|y^n)\left(\frac{1+\hat{b}}{\hat{b}}\right), \end{aligned}$$

where variance is larger than the mean by an amount determined by $\hat{b}$

What happens when n is large?

Note that as the sample size n increases, the posterior density for θ becomes more & more concentrated

$$\mathrm{V}(\theta|y^n) = \widehat{a}/\widehat{b}^2 = (a + \sum_i y_i)/(b+n)^2 \approx \overline{y}/n \to 0$$

As we have less uncertainty about θ, inflation factor $(1+\widehat{b})/\widehat{b} \to 1$ and the predictive density $f(y|y^n) \to \text{Poisson}(\overline{y})$

In smaller samples important to inflate our predictive intervals to account for uncertainty in θ

What if you've never seen the negative-binomial distribution, or didn't recognize that this is a negative-binomial distribution?

Use simulation as approximation, which means:
You can still draw samples…

step 1: draw from $\text{Ga}(\theta; \hat{a}, \hat{b})$ with $\theta \leftarrow$ rgamma(1, $\hat{a}, \hat{b}$)
step 2: plug the θ into Poisson with y $\leftarrow$ rpois(1, θ)

This is Monte Carlo sampling directly from the posterior predictive distribution

Suppose we can sample S values from the posterior distribution of θ, so that

$$\theta^{(1)}, \ldots, \theta^{(S)} \overset{\text{iid}}{\sim} p(\theta \mid Y)$$

for large S.

Law of Large Numbers

$$\frac{1}{S}\sum \theta^{(i)} \rightarrow E[\theta \mid Y]$$
$$\frac{1}{S}\sum g(\theta)^{(i)} \rightarrow E[g(\theta) \mid Y]$$

Sample means converge to their expectations.

Prior predictive distribution

In addition to posterior predictive distribution, it's often very helpful to plot your prior predictive distribution and see if it looks like your data before doing any posterior computation/simulation etc.

$$f(y) = \int L(y;\theta)p(\theta)d\theta = \int_0^\infty Pois(y;\theta)Ga(\theta;a,b)d\theta$$

Used the likelihood function, but didn't use the data

In practice, should always try to draw samples from prior predictive distribution and examine whether they look like your data!

Preview

- Going beyond one parameter:
    - Joint posterior is hard to sample from
    - But, full conditional distribution of each parameter is in closed-form
    - e.g., Normal distribution

- MCMC (Markov Chain Monte Carlo)
    - Gibbs sampling
    - Metropolis

Perhaps something more useful: Normal (Gaussian) models

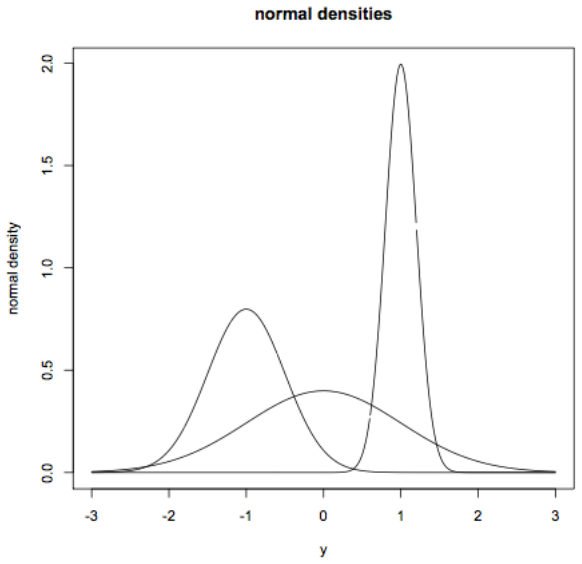
For a random variable $Y \sim N(\mu, \sigma^2)$, the pdf is

$$f(y|\theta,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\theta)^2\right\}, \quad y \in \Re = (-\infty, \infty).$$

The normal density is symmetric about the mean, median & mode θ

95% probability within $\pm 1.96\sigma$ (approximately two standard deviations) of the mean

rnorm, dnorm, pnorm, qnorm in R take mean and standard deviation σ as arguments



It is amazing how often real data are close to normally distributed

Likely a consequence of central limit theorem - sum or mean of a set of random variables is normally distributed

Occurs under very general conditions

A note on parameterization:

Independent observations $Y = (y_1, y_2, \ldots y_n)$

$$y_i \mid \mu, \sigma^2 \sim N(\mu, \sigma^2)$$

unknown parameters μ and σ^2.

Some prefer to work with the *precision*, ϕ, where $\phi = 1/\sigma^2$.

Likelihood:

$$\begin{aligned} L(Y; \mu, \phi) &\propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \phi^{1/2} \exp\left\{ -\frac{1}{2}\phi(y_i - \mu)^2 \right\} \\ &\propto \phi^{n/2} \exp\left\{ -\frac{1}{2}\phi \sum_i (y_i - \mu)^2 \right\} \end{aligned}$$

Now we have two parameters μ and ϕ, how should we pick a prior distribution?

You need a joint prior $p(\mu, \phi)$

Conjugate prior (Normal-Gamma)

$$p(\mu, \phi) = p(\mu|\phi)p(\phi)$$

Semi-conjugate prior (Gibbs sampling)

$$p(\mu, \phi) = p(\mu)p(\phi)$$

Prior:

The conjugate prior for (μ, ϕ) is Normal-Gamma.

$$\begin{aligned} \mu|\phi &\sim \mathrm{N}(\mu_0, 1/(\kappa_0\phi)) \\ \phi &\sim \mathrm{G}(v_0/2, \mathrm{SS}_0/2) \end{aligned}$$

Note that μ depends on ϕ

where $-\infty < \mu_0 < \infty, \kappa > 0, \mathrm{SS}_0 > 0, v_0 > 0$

This can be expressed as

$$p(\mu, \phi) \propto \phi^{v_0/2-1} \exp\left\{-\phi\frac{\mathrm{SS}_0}{2}\right\}\phi^{1/2}\exp\left\{-\phi\frac{\kappa_0}{2}(\mu-\mu_0)^2\right\}$$

Likelihood:

$$\begin{aligned} L(\mu, \phi|Y) &\propto \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}\phi^{1/2}\exp\left\{-\frac{1}{2}\phi(y_i-\mu)^2\right\} \\ &\propto \phi^{n/2}\exp\left\{-\frac{1}{2}\phi\sum_i(y_i-\mu)^2\right\} \end{aligned}$$

Posterior:

$$\mu \mid \phi, Y \sim \mathrm{N}\left(\mu_n, \frac{1}{\kappa_n\phi}\right)$$

$$\phi \mid Y \sim \mathrm{G}\left(\frac{v_n}{2}, \frac{\mathrm{SS}_n}{2}\right)$$

where

$$\kappa_n = \kappa_0 + n$$

$$\mu_n = \frac{\phi n\bar{y} + \phi\kappa_0\mu_0}{\phi\kappa_n}$$

$$v_n = v_0 + n$$

$$\mathrm{SS}_n = \mathrm{SS}_0 + \mathrm{SS} + \frac{n\kappa_0}{\kappa_n}(\bar{y}-\mu_0)^2$$

Interpretation:

κ_n: like sample size for estimating μ (precision $= \phi\kappa_n$)

μ_n: expected value for μ is weighted average

$$\mu_n = \frac{n}{\kappa_n}\bar{y} + \frac{\kappa_0}{\kappa_n}\mu_0$$

v_n: degrees of freedom for estimating ϕ

$\phi \sim G(a/2, b/2) \Leftrightarrow \phi b \sim \chi^2_a$ with degrees of freedom a

$\mathrm{SS}_n = \mathrm{SS}_0 + \mathrm{SS} + \frac{n\kappa_0}{\kappa_n}(\bar{y}-\mu_0)^2$: total posterior variation

- prior variation,
- observed variation (sum of squares),
- variation between prior mean and sample mean

What if we want to model μ and ϕ independently?

Let's explore Gibbs sampler with this example

Prior: $p(\mu, \phi) = p(\mu)p(\phi)$ $\quad \mu \sim N(\mu_0, 1/\tau_0)$

$\phi \sim G(\nu_0/2, SS_0/2)$ for simplicity, $\phi \sim G(a, b)$

Likelihood:

$$L(\mu, \phi|Y) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \phi^{1/2} \exp\left\{ -\frac{1}{2}\phi(y_i - \mu)^2 \right\}$$

$$\propto \phi^{n/2} \exp\left\{ -\frac{1}{2}\phi \sum_i (y_i - \mu)^2 \right\}$$

Posterior:

$$\propto \tau_0^{1/2} exp(-\tfrac{\tau_0}{2}(\mu - \mu_0)^2) \times \phi^{a-1} exp(-b\phi) \times \phi^{n/2} exp(-\tfrac{\phi}{2} \textstyle\sum_{i=1}^{n} (y_i - \mu)^2)$$

For Gibbs sampling, we can derive full conditional distributions

$$p(\mu|\phi, Y), \text{ pretend that } \phi \text{ is constant}$$

$$p(\phi|\mu, Y), \text{ pretend that } \mu \text{ is constant}$$

Markov chain Monte Carlo (MCMC)

Basic idea

Markov chain Monte Carlo (MCMC) provides an approach for generating samples from the posterior distribution

Note that this does not give us an approximation to $\pi(\theta \mid \mathbf{y})$ directly (this is different from the MC example in the posterior predictive case)

However, from these samples we can obtain summaries of the posterior distribution for θ

Summaries of exact posterior distributions of $g(\theta)$, for any functional $g(\cdot)$, can also be obtained.

How does MCMC work?

- Let $\theta^t = (\theta_1^t, \ldots, \theta_p^t)$ denote the value of the $p \times 1$ vector of parameters at iteration t.
- $\theta^0 =$ initial value used to start the chain (*shouldn't be sensitive*)
- MCMC generates θ^t from a distribution that depends on the data & potentially on θ^{t-1}, but not on $\theta^1, \ldots, \theta^{t-2}$.
- This results in a Markov chain with stationary distribution $\pi(\theta \mid \mathbf{y})$ under some conditions on the sampling distribution

Gibbs Sampling

Start with initial value $\theta^0 = (\theta_1^0, \ldots, \theta_p^0)$

For iterations $t = 1, \ldots, T$,

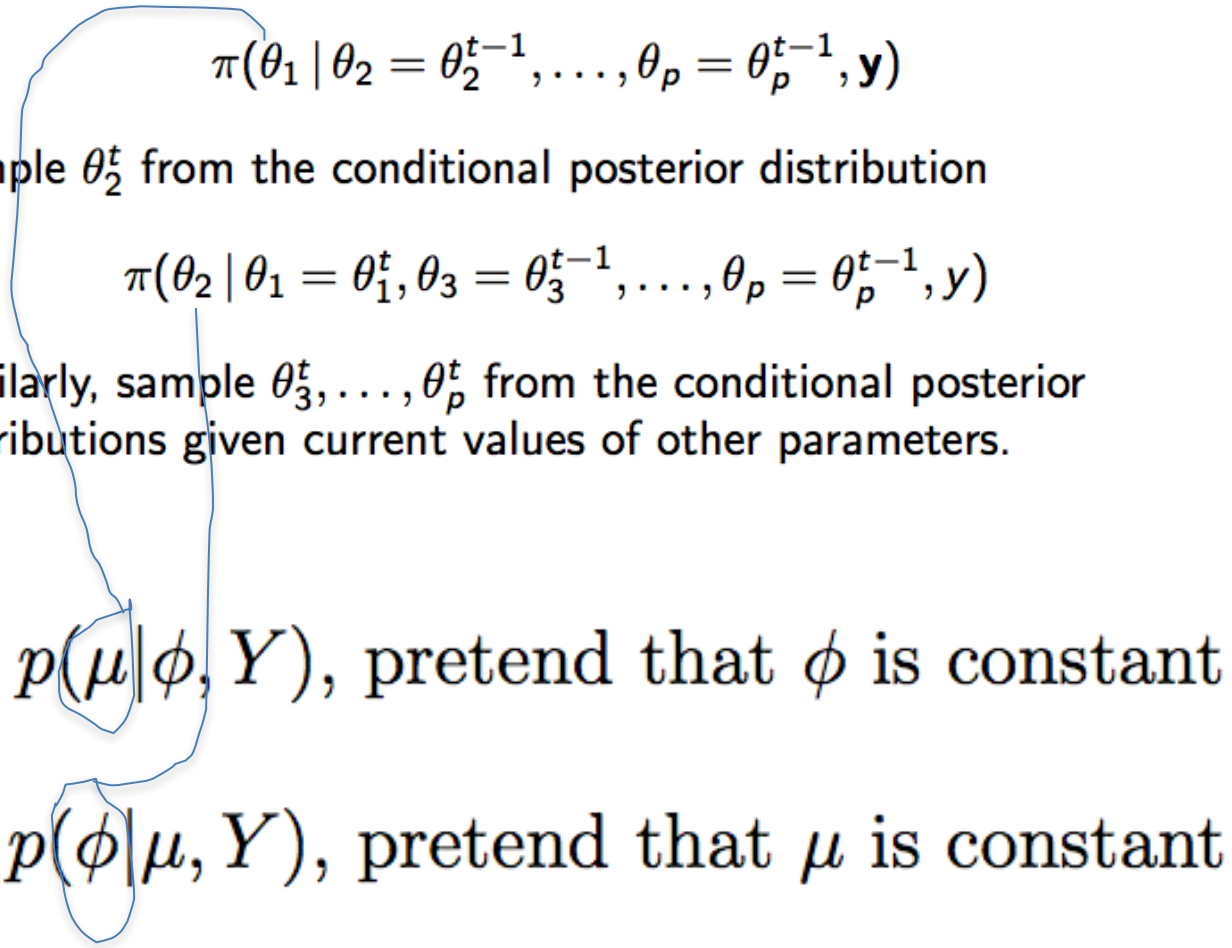
1. Sample θ_1^t from the conditional posterior distribution

$$\pi(\theta_1 \,|\, \theta_2 = \theta_2^{t-1}, \ldots, \theta_p = \theta_p^{t-1}, \mathbf{y})$$

2. Sample θ_2^t from the conditional posterior distribution

$$\pi(\theta_2 \,|\, \theta_1 = \theta_1^t, \theta_3 = \theta_3^{t-1}, \ldots, \theta_p = \theta_p^{t-1}, y)$$

3. Similarly, sample $\theta_3^t, \ldots, \theta_p^t$ from the conditional posterior distributions given current values of other parameters.



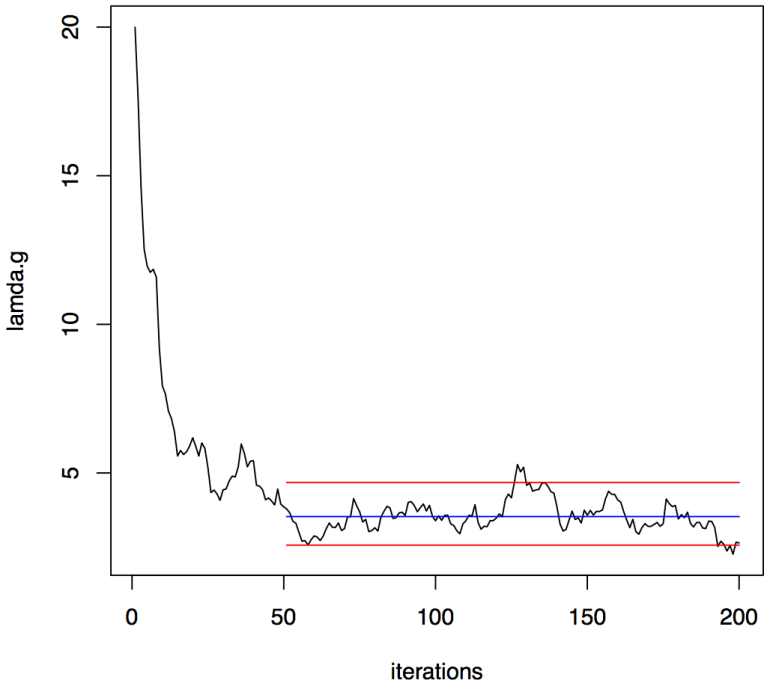
Gibbs Sampling (continued)

- Under mild regularity conditions, samples converge to stationary distribution $\pi(\theta \mid \mathbf{y})$
- At the start of the sampling, the samples are not from the posterior distribution $\pi(\theta \mid \mathbf{y})$.
- It is necessary to discard the initial samples as a *burn-in* to allow convergence
- In simple models, convergence typically occurs quickly & burn-in of 100 iterations should be sufficient

See Casella & George, 1992 linked below if you're interested in *why* it works
http://www.stat.ufl.edu/archived/casella/OlderPapers/ExpGibbs.pdf

HW4

1. Suppose we have a model $y_i \sim \text{Pois}(\phi_i \lambda)$, with $\phi_i \sim \text{Ga}(\nu/2, \nu/2)$ & $\lambda \sim \text{Ga}(a, b)$ Use a=5, b=2 for your HW
2. Derive the full conditional posterior distributions of each ϕ_i and of λ
3. Simulate data with $\nu = 0.5$, $\lambda = 5$ and $n = 100$
4. Implement the Gibbs sampler & show a trace plot of samples of λ
5. Show an MCMC estimate of the posterior mean & 95% credible interval on this plot along with true value of λ - comment



Gibbs Sampling

Not only useful for approximating posterior distribution, but also useful for imputing missing data

Say, you sent out a questionnaire and ask 100 professors about their body weight (bw)

55 responded, 45 didn't

If this is the only piece of information you've collected, it would be hard to know whether your samples (those that responded) are representative.

But, if you collected other information (age, gender, # of times they exercise/week, medical history), you might have some good guesses about the mechanisms for missing data and therefore can try to impute the missing data (see data matrix on board)

MCMC applied in imputing missing data

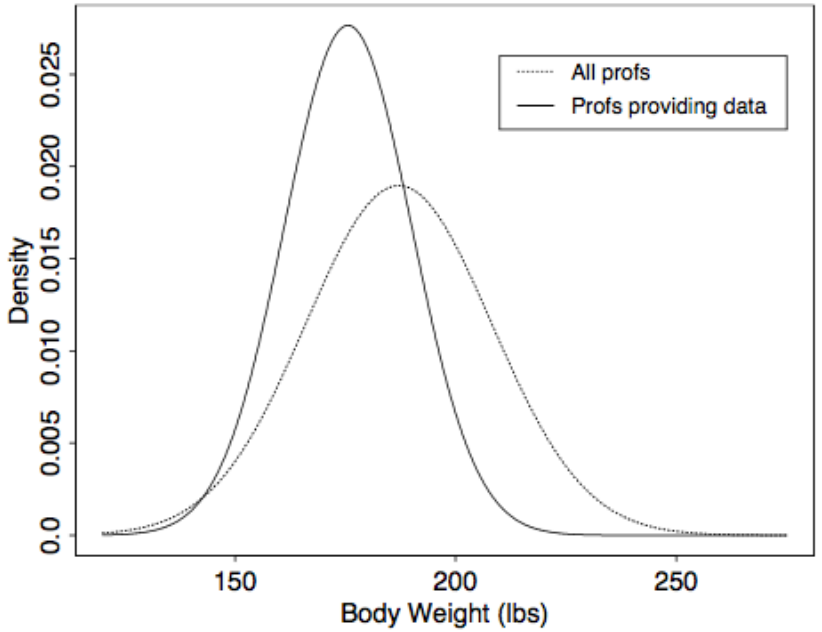
- Missing data are extremely common in most application areas
- Many methods break down in the presence of missing data
- Commonly 'complete case' analyses are done that simply discard all individuals having any missing values
- Alternatively, one can 'impute' missing values & then run the analysis as if the imputed values were observed

- There are many ad hoc methods available for imputation
- Sometimes people just fill in the mean of the observed values for the same variable
- What's wrong with this conceptually?
- Ignores uncertainty in imputation & may result in biased estimation

General notation

$$Y_i \overset{iid}{\sim} f_Y(\theta) \quad \text{for } i = 1, \ldots, n$$

$$M_i = \begin{cases} 1 \text{ if } Y_i \text{ is missing} \\ 0 \text{ if } Y_i \text{ is observed} \end{cases}$$



Interest: Inference on θ

- Let $y_1, \ldots, y_{100}$ denote the body weights for the 100 profs surveyed
- Let $m_1, \ldots, m_{100}$ denote 0/1 indicators that the survey was returned
- **Selection Model Likelihood**:

$$\left[(2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\} \right]$$

$$\left[\prod_{i=1}^{n} h(y_i; \psi)^{m_i} \left\{ 1 - h(y_i; \psi) \right\}^{1-m_i} \right],$$

where $h(y_i; \psi) = \Pr(M = 1 \mid Y = y_i)$.

For example, the bigger their weight, the more likely, M=1 (not responding, i.e., missing)

Alternatively, You can use a pattern-mixture likelihood

- Likelihood is as follows

$$\left[(2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu-m_i\alpha)^2\right\}\right]$$
$$\left[\prod_{i=1}^{n}\pi^{m_i}(1-\pi)^{1-m_i}\right].$$

- The normal mean depends on whether or not the observation is missing, which is not the case for the selection model

In addition to missing Y, you could also have missing data on the covariates X

Data augmentation with MCMC

- Choose a prior for θ, τ, ψ & consider unknown X's as latent data
- Apply the following MCMC algorithm:
  1. Impute missing X's by sampling from their full conditional distribution.
  2. Conditional on completed data, follow standard Gibbs sampling (or other) steps for updating the parameters θ, τ, ψ.
  3. Repeat steps 1-2 for a large number of iterations.
- Often, the conditional distributions required for implementation of this algorithm follow a simple form.