A Brief Introduction to Data Augmentation

STAT 501: Statistical Literature and Problems

Shen Qu

Question

- Binary response regression models

Observed $Y_1, .., Y_n \sim Bern(p_i)$, i=1,..,n.

Covariates $X = X_1, ...X_p$

Desired $\beta = \beta_0, \beta_1, ..., \beta_p$

General Framework

The link function: $Pr(Y_i = 1|\beta) = H(\mathbf{x}_i^T \beta)$
H is a CDF

Frequentist's Method

Iteratively Reweighted Least Squares (IRLS) Algorithm
Fisher Scoring??? Newton-Raphson algorithm "gradient descent level II"???

Bayesian methods

v.s. Multi-layers or approximation
The latent variable $\mathbf{Z} = \mathbf{X}\beta + \varepsilon$

$$y_i|z_i = \begin{cases} 1 & \text{if } z_i > 0 \end{cases}$$

A Data-Augmentation schemes

A well-behaved Markov chain Monte Carlo (MCMC)

A underlying variable Z simulated from the proper distribution

Generating the missing data

Motivation

Assume pdf $f_X(x) : \mathbb{R}^p \to [0, \infty)$, function $g : \mathbb{R}^p \to \mathbb{R}$
want to estimate $E[g(x)]$.

$$E_{f_X}[g(x)] = \int_{\mathbb{R}^p} g(x) f_X(x) dx$$

When $E[g(x)]$ is hard to numerical integral or analytical approximate,
we can use simulation based methods.

Monte Carlo Sampling

Regardless of the distribution, if we have
$g(X_1), g(X_2), \ldots, g(X_m) \overset{iid}{\sim} f_X(x)$
then $\frac{1}{m} \sum_{i=1}^{m} g(X_i)$ is a good estimator for $E(g)$.

- Constructing a Markov chain, one iteration includes:

1. Draw $Y \sim f_{Y|X}(\mathring{u}|x)$.
2. Draw $X_{i+1} \sim f_{X|Y}(\mathring{u}|y)$.

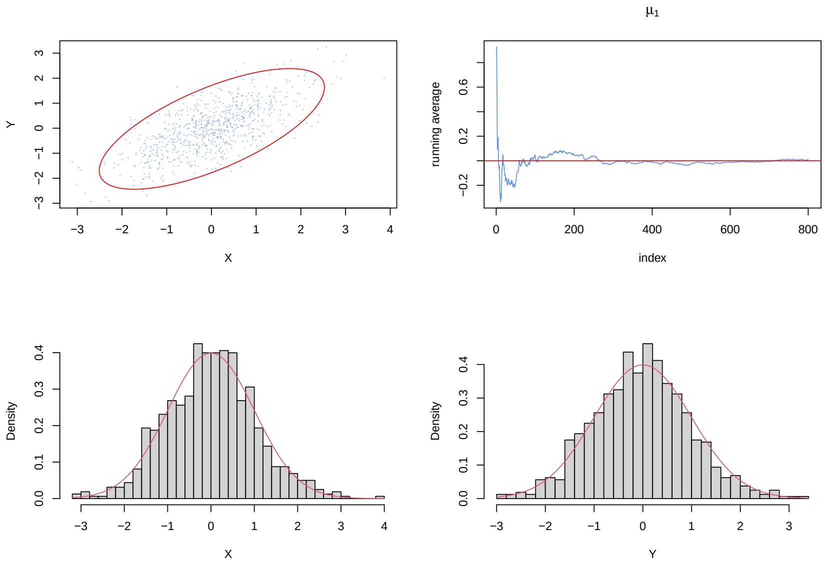
Repeat to simulate $f_X(x)$

Tanner and Wong (1987), Swendsen and Wang (1987)

Ex1: Bivariate Normal Density

The unknown true $X \sim N(0,1)$; $Y \sim N(0,1)$
We know $X, Y \sim N_2(0, 1, \frac{1}{\sqrt{2}})$; $f_X(x) = \int_{\mathbb{R}^q} f(x,y)dy$

1. Draw $(Y|X = x) \sim N(\frac{x}{\sqrt{2}}, \frac{1}{2})$.
2. Draw $(X|Y = y) \sim N(\frac{y}{\sqrt{2}}, \frac{1}{2})$.



μ₁
X
Y
running average
index
Density
X
Density
Y

Conditions and Properties

Harris ergodic, which satisfies three properties: irreducible, aperiodic, and recurrent.

A sufficient condition for Harris ergodicity is

$$\mathcal{K} : k(x'|x) > 0 \quad \forall x', x \in \mathbf{X}$$

Definition

A Markov chain, $X = \{X_i\}_{i=0}^{\infty}$,with state space X. If the current state of the chain is $X = x$, then the density of the next state, X', is $k(x'|x)$. The Markov transition density (Mtd) is

$$k(x'|x) = \int_Y f_{X|Y}(x'|y) f_{Y|X}(y|x) dy$$

Check $k(x'|x)$ is a pdf:

$$\begin{aligned}
\int_X k(x'|x)dx' &= \int_X \left[\int_Y f_{X|Y}(x'|y)f_{Y|X}(y|x)dy\right] dx' \\
&= \int_Y f_{Y|X}(y|x) \left[\int_X f_{X|Y}(x'|y)dx'\right] dy \\
&= \int_Y f_{Y|X}(y|x)dy = 1
\end{aligned}$$

Invariant (stationarity)

f_X is an invariant density for K when

$$f_X(x') = \int_X k(x'|x)f_X(x)dx$$

Then the Markov chain is time homogeneous and the "recurrent" property holds. ???

Detailed balance

For all $x, x' \in \mathbf{X}$, let

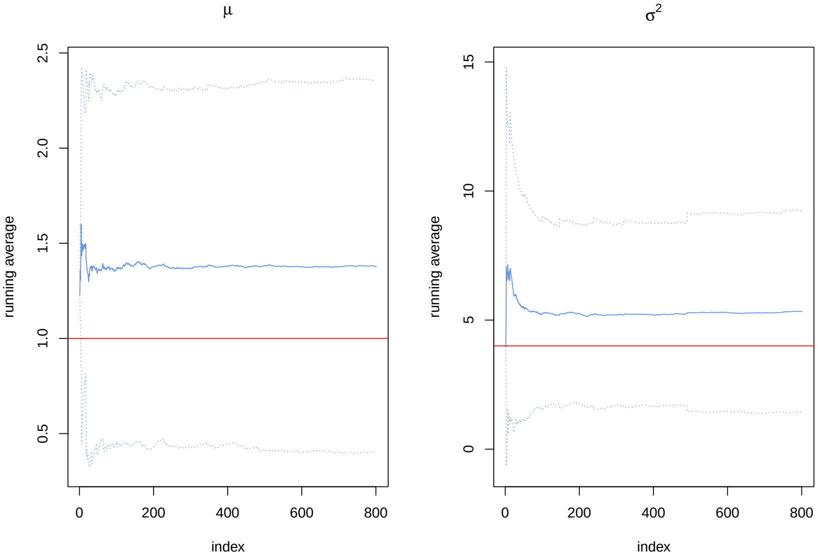
$$\delta(x', x) = k(x'|x)f_X(x) = \int \frac{f(x', z)}{\quad} \cdot \frac{f(z, x)}{\quad} f_X(x)dz$$

$p((\mu, \sigma^2), y|z) \propto \pi(\mu, \sigma^2) p(z, y|\mu, \sigma^2) = \frac{1}{\sigma^2} p(z, y|\mu, \sigma^2)$

1. Draw $(Y_i|\mu, \sigma^2, z) \sim Gamma(\frac{\nu+1}{2}, \frac{1}{2}(\frac{(z_i-\mu)^2}{\sigma^2} + \nu))$.
   $\hat{\mu} = \frac{1}{y_.} \sum_{j=1}^{m} z_j y_j$, $\hat{\sigma}^2 = \frac{1}{y_.} \sum_{j=1}^{m} y_j (z_j - \hat{\mu})^2$
2. Draw $(\sigma^2|y, z) \sim IG(\frac{m+1}{2}, \frac{y_. \hat{\sigma}^2}{2})$.
3. Draw $(\mu|\sigma^2, y, z) \sim N(\hat{\mu}, \frac{\sigma^2}{y_.})$.

EM algorithm (Dempster et al., 1977).

A more general DA algorithm developed by Meng and van Dyk (1999)



μ
σ²
running average
running average
2.5
2.0
1.5
1.0
0.5
15
10
5
0
0
200
400
600
800
index
index

Solutions

- Simple
- Exact
- One layer

Probit Model

A Gibb's sampler:

$$(1)\mathbf{z}^*|\mathbf{y},\beta \sim N_{tr}(\mathbf{x}_i^T\beta, 1) \text{ truncated by 0 at} \begin{cases} \text{left} & y_i = 1 \\ \text{right} & y_i = 0 \end{cases}$$

$$(2)\beta|\mathbf{y},\mathbf{z}^* \sim N_p\left(\underbrace{(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{z}^*}_{m_\beta}, \underbrace{(\mathbf{x}^T\mathbf{x})^{-1}}_{v_\beta}\right)$$

Repeat (1) and (2) long enough. (Albert and Chib, 1993)

Because $\pi(\boldsymbol{\beta}, \boldsymbol{Z}|\boldsymbol{y}) = C\pi(\boldsymbol{\beta}) \prod_{i=1}^m \left[\mathbf{1}_{Z_i>0}\mathbf{1}_{y_i=1} + \mathbf{1}_{Z_i\leq 0}\mathbf{1}_{y_i=0}\right] \phi(Z_i)$

$\pi(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{Z}) = C\pi(\boldsymbol{\beta}) \prod_{i=1}^m \phi(Z_i; \mathbf{x}_i^T\boldsymbol{\beta}, 1)$

Simulate the observed data

Assume $\beta = \{-1, 0.5, 0.25\}$

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N_2(\begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix})$$

```
##     (Intercept)     x1    x2
## 1 0           1 0.4395 2.253
## 2 1           1 0.7698 1.971
## 3 1           1 2.5587 1.957
## 4 1           1 1.0705 3.369
```

```
## (Intercept)          x1          x2
##       1.000       1.034       2.146
```

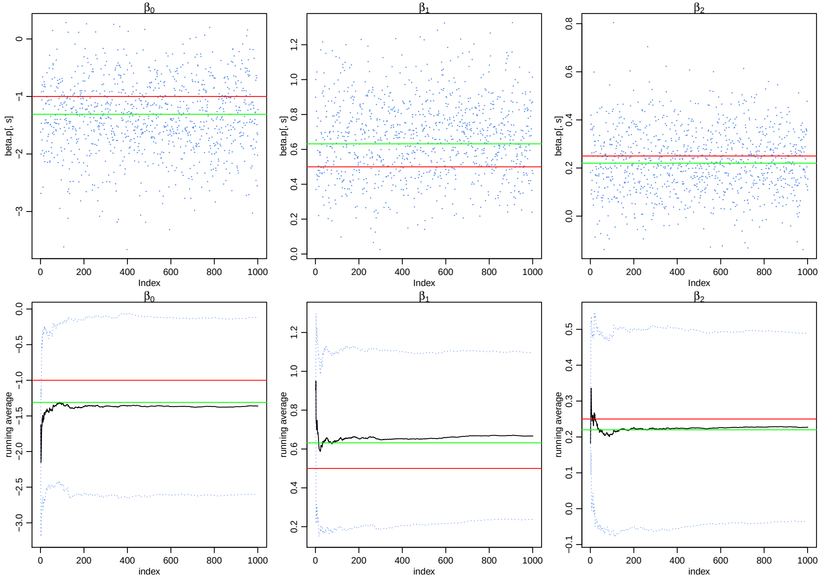
```
## Y
##  0  1
## 28 22
```

- Gibbs' results

	beta.p.mean	beta.p.median	beta.p.ll	beta.p.ul	beta.p.sd
Beta0	-1.362	-1.381	-2.655	-0.1813	0.6337
Beta1	0.6669	0.6714	0.2524	1.13	0.2191
Beta2	0.227	0.2199	-0.0332	0.483	0.1336

IRLS' results

	glm.p	glm.p.sd	2.5 %	97.5 %
Beta0	-1.31	0.5669	-2.503	-0.2209
Beta1	0.6318	0.2254	0.2095	1.092
Beta2	0.22	0.2117	-0.1916	0.6492



Logit Model

$P(Y_i = 1) = \frac{\exp(\mathbf{x}_i^T\beta)}{1+\exp(\mathbf{x}_i^T\beta)}$; $H^{-1}(p_i) = \ln\frac{p_i}{1-p_i} = \mathbf{x}_i^T\beta$ log odds of success

$(1)\quad \omega_i^*|\mathbf{x}_i, \beta \sim PG(n_i, |\mathbf{x}_i^T\beta|)$

Introduce Pólya–Gamma

$$f(\omega|\beta) = \cosh(\frac{1}{2}\mathbf{x}^T\boldsymbol{\beta})\exp\left[-\frac{1}{2}(\mathbf{x}^T\boldsymbol{\beta})^2\omega\right]g(\omega)$$

where $\cosh(c) = \frac{1}{2}(e^c + e^{-c}) = \frac{1+e^2c}{2e^c}$, $g(\omega)$ is free of β s.t.

$$g(\omega) = \sum_n^\infty (-1)^n \frac{2n+1}{\sqrt{2\pi\omega^3}} \exp[-\frac{(2n+1)^2}{8\omega}]\mathbb{I}_{(0,\infty)}(\omega)$$

Since y is observed data, $\pi(\omega|\beta, y) = f(\omega|\beta)$.

Assign the prior of β is $\sim N_p(\mathbf{b}, \mathbf{B})$

$$\begin{aligned}
\pi(\boldsymbol{\beta}, \boldsymbol{\omega}|\boldsymbol{y}) &= \frac{\pi(\beta)}{C(y)} \prod_{i=1}^{N} [p(y_i|\beta)] f(\omega_i|\beta) \\
&= \frac{\pi(\beta)}{C(y)} \prod_{i=1}^{N} \left[\frac{e^{y_i \mathbf{x}_i^T \beta}}{1 + e^{\mathbf{x}_i^T \beta}} \right] \cosh\left(\frac{1}{2} \boldsymbol{x_i^T} \beta\right) \exp\left[-\frac{1}{2} (\mathbf{x}_i^T \beta)^2 \omega_i \right] g(\omega_i) \\
&\propto \phi_p(\mathbf{b}, \mathbf{B}) \prod_{i=1}^{N} \frac{e^{y_i \mathbf{x}_i^T \beta}}{1 + e^{\mathbf{x}_i^T \beta}} \cdot \frac{1 + e^{\mathbf{x}_i^T \beta}}{2 e^{\frac{1}{2} \mathbf{x}_i^T \beta}} \exp\left[-\frac{1}{2} (\mathbf{x}_i^T \beta)^2 \omega \right] \\
&= 2^{-n} \phi_p(\mathbf{b}, \mathbf{B}) \prod_{i=1}^{N} \exp\left[(y_i - \frac{1}{2}) \mathbf{x}_i^T \beta - \frac{1}{2} (\mathbf{x}_i^T \beta)^2 \omega_i \right]
\end{aligned}$$

When ω_i is known, $\pi(\boldsymbol{\beta}, \boldsymbol{\omega}|\boldsymbol{y}) = \pi(\beta|\boldsymbol{\omega}, \boldsymbol{y})$.

Let $\boldsymbol{\Omega} = diag_n(\omega_{\mathbf{i}})$; $\boldsymbol{\kappa} = \mathbf{y_{1:n}} - \frac{1}{2}$

Then we get to conditional pmfs of $\pi(\beta|\boldsymbol{\omega}, \boldsymbol{y})$

$$
\begin{aligned}
\pi(\beta|\boldsymbol{\omega}, \boldsymbol{y}) &\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta'(x'^{\cdot}x + B^{-1})\beta - 2\beta'(x'\kappa + B^{-1}b)}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\mathbf{v}_{\omega}^{-1}[\boldsymbol{\beta' I\beta - 2\beta' v_{\omega}(x'\kappa + B^{-1}b)}]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\boldsymbol{(\beta - m_{\omega})' v_{\omega}^{-1}(\beta - m_{\omega})}\right\}
\end{aligned}
$$

where

$$
\mathbf{m}_{\omega} = \underset{(\mathbf{p\cdot p})}{\mathbf{v}_{\omega}}\,(\underset{(\mathbf{p\cdot n})}{\mathbf{x^T}}(\mathbf{y_{1:n}} - \tfrac{\mathbf{1}}{\mathbf{2}}) + \mathbf{B^{-1}b});\ \mathbf{v}_{\omega} = (\mathbf{x^{T\cdot}x} + \mathbf{B^{-1}})^{\mathbf{-1}}.
$$

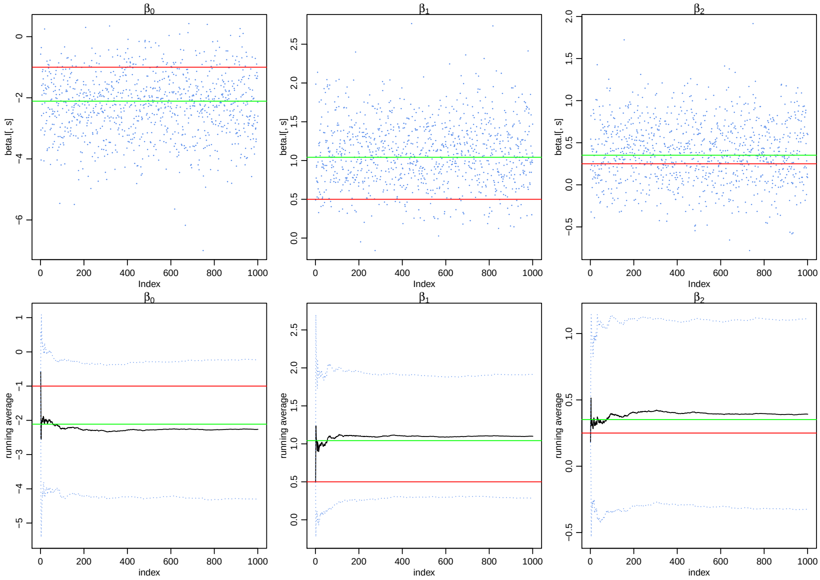
Then we confirm that $\beta|\boldsymbol{\omega}, \boldsymbol{y} \sim N_p(m_{\omega}, V_{\omega})$

Gibbs' results (divided by $\pi/\sqrt{3}$)

	beta.l.mean	beta.l.median	beta.l.sd	beta.l.ll	beta.l.ul
Beta0	-1.248	-1.228	0.5715	-2.392	-0.197
Beta1	0.6064	0.5987	0.2289	0.1728	1.08
Beta2	0.2163	0.2116	0.2017	-0.1735	0.6211

IRLS' results (divided by $\pi/\sqrt{3}$)

	glm.l	glm.l.sd	2.5 %	97.5 %
Beta0	-1.165	0.531	-2.306	-0.1874
Beta1	0.5744	0.2165	0.184	1.047
Beta2	0.1939	0.1935	-0.1768	0.5969



β_0
β_1
β_2
beta.l[, s]
Index
running average
index

Marginal effects

	irls.p	irls.l	gibbs.p	gibbs.l
Beta0	-0.438	-0.4268	-0.4483	-0.4481
Beta1	0.2112	0.2105	0.2195	0.2178
Beta2	0.0735	0.071	0.0747	0.0777

Confidence Interval

	i.p.l	i.l.l	g.p.l	g.l.l	i.p.u	i.l.u	g.p.u	g
Beta0	-0.8367	-0.845	-0.8741	-0.8591	-0.0739	-0.0687	-0.0597	0
Beta1	0.07	0.0674	0.0831	0.062	0.3649	0.3836	0.3718	0
Beta2	-0.0641	-0.0648	-0.0109	-0.0623	0.217	0.2188	0.159	0

literature

Jun S. Liu & Ying Nian Wu (1999) Parameter Expansion for Data Augmentation, Journal of the American Statistical Association, 94:448, 1264-1274, DOI: 10.1080/01621459.1999.10473879

- Gelman, A. (2014). Bayesian data analysis (Third edition.). CRC Press.

11.7 Bibliographic note

Tanner and Wong (1987) introduced the idea of iterative simulation to many statisticians, using the special case of 'data augmentation' to emphasize the analogy to the EM algorithm (see Section 13.4).

Auxiliary variables

12.1 Efficient Gibbs samplers

Gibbs sampler computations can often be simplified or convergence accelerated by adding auxiliary variables, for example indicators for mixture distributions, as described in Chapter 22. The idea of adding variables is also called data augmentation and is often a useful conceptual and computational tool, both for the Gibbs sampler and for the EM algorithm (see Section 13.4).

12.7 Bibliographic note

For the relatively simple ways of improving simulation algorithms

Imai, K., and van Dyk, D. A. (2005). A Bayesian analysis of the multinomial probit model using marginal data augmentation. Journal of Econometrics. 124, 311–334. https://doi.org/10.1016/j.jeconom.2004.02.002

Rubin, D. B. (1987b). A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. Discussion of Tanner and Wong (1987). Journal of the American Statistical Association 82, 543–546.

Tanner, M. A., and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). Journal of the American Statistical Association 82, 528–550.

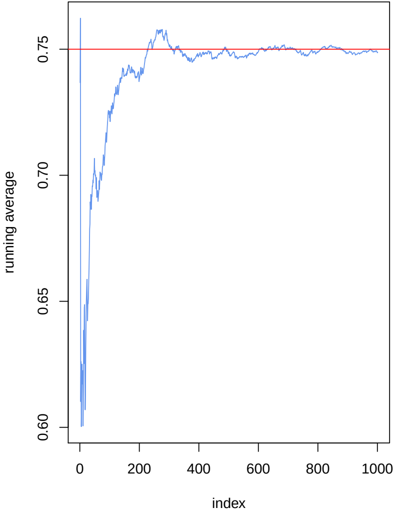
van Dyk, D. A., and Meng, X. L. (2001). The art of data augmentation (with discussion). Journal of Computational and Graphical Statistics 10, 1–111.

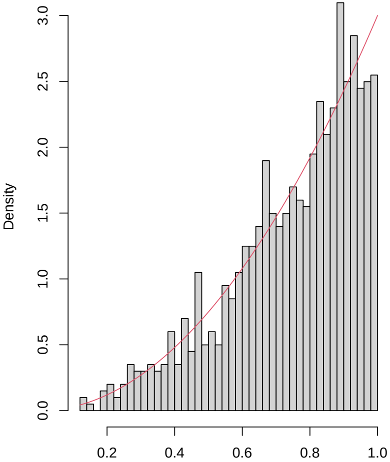
Ex2: Simple Slice Sampler (Neal, 2003)

$(f_X(x) = 3x^2 I_{(0,1)}(x);\ E[X] = \int_0^1 x f_X(x) dx = 0.75)$

1. Draw $(Y|X = x) \sim Unif(0, x)$ and $U \sim Unif(0, 1)$.
2. Update $(X|Y = y, U = u) = \sqrt{u(1 - y^2) + y^2}$.

```
g.ex2<-function (n,step=20){
  x <- .5
  y <- .5
  X <- NA
  for (i in 1:n) {
    y <- runif(1,0,x)
    u <- runif(1,0,1)
 X[i] <- x <- sqrt(u*(1-y^2)+y^2)
      }
 thinned=seq(round(n*0.2),n,step)
 X[thinned]
}
```



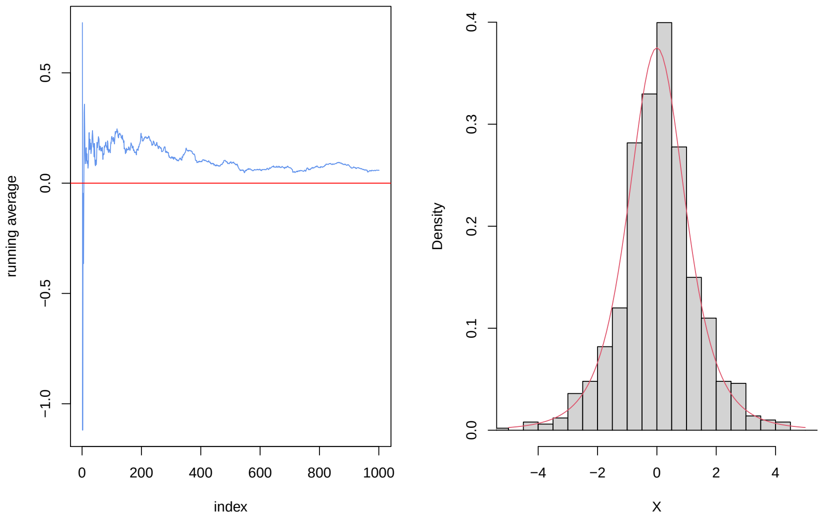


Ex3: t: Normal-Gamma

$(X \sim t_4,\ f_X(x) = \frac{3}{8}(1 + \frac{x^2}{4})^{-\frac{5}{2}})$

$[f(x,y) = \frac{4}{\sqrt{2\pi}} y^{\frac{3}{2}} \exp\{-y(\frac{x^2}{2} + 2)\} I_{(0,\infty)}(y)\ ]$

1. Draw $(Y|X = x) \sim Gamma(\frac{5}{2}, \frac{x^2}{2} + 2)$.
2. Draw $(X|Y = y) \sim N(0, y^{-1})$


running average
0.5
0.0
−0.5
−1.0
0
200
400
600
800
1000
index
Density
0.4
0.3
0.2
0.1
0.0
−4
−2
0
2
4
X