Master solution of group tasks

Exercise 3

(Conjugate Bayes: analytical derivation - 8 Points) Assume that $y_1, \ldots, y_n$ are realizations (observations) generated by iid random variables which follow a $\mathcal{N}(m, \kappa^{-1})$ distribution. Moreover, assume that the prior of m follows a $\mathcal{N}(\mu, \lambda^{-1})$ distribution, where κ, μ and λ are fixed (known) constants. Derive analytically (with all constants) the following two distributions:

a) The prior predictive distribution of one future observation y assuming that no observations have been collected yet.

Solution: *Prior predictive distribution* is defined as follows

$$f(y) = \int_{-\infty}^{\infty} f(y \mid \theta) f(\theta) d\theta, \tag{1}$$

where $f(\theta)$ is the prior and $f(y \mid \theta)$ is the likelihood.

In our example,

$$\begin{aligned}
f(y) &= \int_{-\infty}^{\infty} f(y \mid m) f(m) dm \\
&= \int_{-\infty}^{\infty} \sqrt{\frac{\kappa}{2\pi}} \exp\left(-\frac{\kappa(y-m)^2}{2}\right) \sqrt{\frac{\lambda}{2\pi}} \exp\left(-\frac{\lambda(m-\mu)^2}{2}\right) dm \\
&= \frac{\sqrt{\kappa\lambda}}{2\pi} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left(\kappa(y^2 + m^2 - 2my) + \lambda(m^2 + \mu^2 - 2m\mu)\right)\right\} dm \\
&= \frac{\sqrt{\kappa\lambda}}{2\pi} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left(\kappa y^2 + \lambda\mu^2 - \frac{(\kappa y + \lambda\mu)^2}{\kappa + \lambda} + (\kappa + \lambda)\left(m - \frac{\kappa y + \lambda\mu}{\kappa + \lambda}\right)^2\right)\right\} dm \\
&= \frac{\sqrt{\kappa\lambda}}{2\pi} \exp\left\{-\frac{\lambda\kappa(y-\mu)^2}{2(\kappa+\lambda)}\right\} \sqrt{\frac{2\pi}{\kappa+\lambda}} \underbrace{\int_{-\infty}^{\infty} \sqrt{\frac{\kappa+\lambda}{2\pi}} \exp\left\{-\frac{\kappa+\lambda}{2}\left(m - \frac{\kappa y + \lambda\mu}{\kappa+\lambda}\right)^2\right\} dm}_{=1}
\end{aligned} \tag{2}$$

According to a property of the probability density functions the integral of a normal probability density function is 1. The remaining part of our equation

$$f(y) = \sqrt{\frac{\lambda\kappa}{2\pi(\kappa+\lambda)}} \exp\left\{-\frac{\lambda\kappa(y-\mu)^2}{2(\kappa+\lambda)}\right\}$$

Hence, the prior predictive distribution of one future observation y is $\mathcal{N}\left(\mu, \lambda^{-1} + \kappa^{-1}\right)$.

b) The posterior predictive distribution for one future observation y_{n+1} given $y_1, \ldots, y_n$ have been observed.

Solution:

According to the definition of the posterior predictive distribution in Held and Sabanés Bové, 2020, Section 9.3, we have

$$\begin{aligned} f(y_{n+1} \mid y_1, \ldots, y_n) &= \int_{-\infty}^{\infty} f(y_{n+1}, m \mid y_1, \ldots, y_n) dm \\ &= \int_{-\infty}^{\infty} f(y_{n+1} \mid m, y_1, \ldots, y_n) f(m \mid y_1, \ldots, y_n) dm \\ &\overbrace{=}^{\text{cond. ind.}} \int_{-\infty}^{\infty} \underbrace{f(y_{n+1} \mid m)}_{\text{likelihood}} \underbrace{f(m \mid y_1, \ldots, y_n)}_{\text{posterior density}} dm \end{aligned} \tag{3}$$

We have

$$f(y_{n+1} \mid m) = \sqrt{\frac{\kappa}{2\pi}} \exp\left(\frac{\kappa(y_{n+1} - m)^2}{2}\right)$$

We have derived the posterior distribution in Worksheet 02, exercise 3.

$$m \mid y_1, \ldots, y_n \sim \mathcal{N}\left(\frac{\kappa \bar{y} n + \lambda \mu}{\kappa n + \lambda}, \frac{1}{\kappa n + \lambda}\right).$$

Define

$$\mu_{\text{post}} = \frac{\kappa n \bar{y} + \lambda \mu}{n\kappa + \lambda},$$

and

$$\lambda_{\text{post}} = n\kappa + \lambda.$$

Hence,

$$f(y_{n+1} \mid y_1, \ldots, y_n) = \int_{-\infty}^{\infty} \sqrt{\frac{\kappa}{2\pi}} \exp\left(\frac{\kappa(y_{n+1} - m)^2}{2}\right) \sqrt{\frac{\lambda_{\text{post}}}{2\pi}} \exp\left(-\frac{\lambda_{\text{post}}}{2}(m - \mu_{\text{post}})^2\right) dm.$$

A rearrangement of the terms leads to the posterior predictive distribution:

$$y_{n+1} \mid y_1, \ldots, y_n \sim \mathcal{N}\left(\mu_{\text{post}}, \lambda_{\text{post}}^{-1} + \kappa^{-1}\right), \tag{4}$$

with

$$\mu_{\text{post}} = \frac{\kappa n \bar{y} + \lambda \mu}{n\kappa + \lambda},$$

and

$$\lambda_{\text{post}} = n\kappa + \lambda.$$

Exercise 4

(Conjugate Bayesian analysis in practice - 6 Points) Apply analytical formulas derived in Exercise 3 above to the vector of `Height` (cm) measurements 166, 168, 168, 177, 160, 170, 172, 159, 175, 164, 175, 167, 164 of 13 Swiss females. Assume that $y_1, \ldots, y_n$ are observations generated by $\mathcal{N}(m, \kappa^{-1})$ distribution with $\kappa = 1/900$. Moreover, assume a $\mathcal{N}(\mu, \lambda^{-1})$ prior for m with $\mu = 161$ and $\lambda = 1/70$.

a) Plot the prior predictive distribution for one observation y and compute its expectation and standard deviation. Estimate $P[y > 200]$ for one future observation of `Height`.

Solution:

```
Height <- c(166, 168, 168, 177, 160, 170, 172, 159, 175, 164, 175, 167, 164)
n <- length(Height)
y_bar <- mean(Height)
kappa <- 1 / 900
mu <- 161
lambda <- 1 / 70
```

The expectation of the prior predictive distribution for one future observation is $\mu = 161$ and the standard deviation is $\sqrt{\lambda^{-1} + \kappa^{-1}} = 31.14$.

```
(mean_prior_pred <- mu) # expectation of the prior predictive distribution

## [1] 161

(sd_prior_pred <- sqrt(lambda^(-1) + kappa^(-1))) # standard deviation

## [1] 31.14482

# of the prior predictive distribution
```

Figure 1 shows the prior predictive distribution in red.

$P[y > 200]$ is

```
(prior_pred_200 <- 1 - pnorm(200,
  mean = mean_prior_pred,
  sd = sd_prior_pred
))

## [1] 0.1052459
```

b) Plot the posterior predictive distribution for one observation y_{n+1} given that $y_1, \ldots, y_n$ have been observed and compute its expectation and standard deviation. Estimate $P[y_{n+1} > 200 \mid y_1, \ldots, y_n]$ for one future observation of y_{n+1} of `Height`.

Solution:

The posterior predictive distribution is given by the formula (4), then the expectation of the posterior predictive distribution for one future observation is $\mu_{\text{post}} = 164.558$ and the standard deviation is $\sqrt{(n\kappa + \lambda)^{-1} + \kappa^{-1}} = 30.57$.

```
(mean_post_pred <- (kappa * n * y_bar + lambda * mu) / (n * kappa + lambda)) # expe

## [1] 164.558

# of the posterior predictive distribution
(sd_post_pred <- sqrt((lambda + n * kappa)^(-1) + kappa^(-1))) # standard deviatior

## [1] 30.57461

# of the posterior predictive distribution (n*kappa + lambda)^(-1)
mean_post <- mean_post_pred
sd_post <- sqrt((lambda + n * kappa)^(-1))
```

$P[y_{n+1} > 200 \mid y_1, \ldots, y_n]$ is

```
(posterior_pred_200 <- 1 - pnorm(200, mean = mean_post_pred, sd = sd_post_pred))

## [1] 0.123188
```

c) Compare the results obtained for predictive distribution with those obtained for the posterior in Exercise 4 of Worksheet 02. Discuss how much posterior, prior predictive, and posterior predictive distributions differ.

Solution:

The results derived above and in Worksheet 02 Exercise 4 show that the mean parameters for the prior and prior predictive distributions are the same, whereas the variances differ. The variance of the prior predictive distribution is larger than the variance of the prior distribution by the variance of the data. We can state similar arguments also for the posterior distribution and the posterior predictive distribution. They have the same mean parameters but the variances differ. The posterior predictive distribution has larger variance because the variance of the data is added to the variance of the posterior distribution. Moreover, the posterior predictive probability $P[y_{n+1} > 200 \mid y_1, \ldots, y_n] = 0.123$ is much larger than the posterior probability $P[m > 200 \mid y_1, \ldots, y_n] = 9.43e - 10$ (Figure 1).

Exercise 5

(The change-of-variables formula - 8 Points) Consider $X \sim \mathrm{G}(a, b)$ and

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx) \tag{5}$$

Solution:

- Consider the random variable $Y = \frac{1}{X}$. Derive the density of Y.

  Let $g(x)$ be a strictly monotonic function with inverse having a continuous derivative. Consider a transformed variable $Y = g(X)$. Then we have the change-of-variables formula

  $$f_Y(y) = f_X\left(g^{-1}(y)\right) \left| \frac{dg^{-1}(y)}{dy} \right|. \tag{6}$$

  $y = \frac{1}{x} = g(x)$, then $x = \frac{1}{y} = g^{-1}(y)$ and $\frac{dg^{-1}(y)}{dy} = -\frac{1}{y^2}$.

  By Equation (6)

  $$f(y) = \frac{b^a}{\Gamma(a)} \left(\frac{1}{y}\right)^{a-1} \exp\left\{-\frac{b}{y}\right\} \left| -\frac{1}{y^2} \right| = \frac{b^a}{\Gamma(a)} y^{-(a+1)} \exp\left\{-\frac{b}{y}\right\}, \tag{7}$$

  which is the density function of the Inverse Gamma (IG) distribution.

- Consider the random variable $Z = \sqrt{Y} = \sqrt{1/X}$. Derive the density of Z.

  $Z = \sqrt{\frac{1}{x}} = g(x)$, then $x = \frac{1}{z^2} = g^{-1}(z)$ and $\frac{dg^{-1}(z)}{dz} = -\frac{2}{z^3}$.

  By Equation (6)

  $$f(z) = \frac{b^a}{\Gamma(a)} \left(\frac{1}{z^2}\right)^{a-1} \exp\left\{-\frac{b}{z^2}\right\} \left| -\frac{2}{z^3} \right| = \frac{2b^a}{\Gamma(a)} z^{-(2a+1)} \exp\left\{-\frac{b}{z^2}\right\} \tag{8}$$

  which is the density function of the Square root Inverse Gamma (SIG) distribution. Figure 2 shows densities of G, IG and SIG distributions. Figure 3 shows that both IG and SIG distributions assign probability 0 to values close to 0.

Exercise 6

(Monte Carlo: transformations of random variables - 8 Points)

a) Generate a Monte Carlo sample (i.i.d realizations of X) of size $M = 1000$ using `rgamma()` function in R and by assuming `set.seed(44566)`.

Solution:

```
M <- 1000
set.seed(44566)
G_sample <- rgamma(M, shape = a, rate = b)
```

We parametrize the Gamma distribution in Equation (5) by the `shape = a` and `rate = b`.

b) Given the sample in (a) generate a MC sample from the IG distribution $Y = 1/X$.

Solution:

```
IG_sample <- 1 / G_sample
```

c) Given the sample in (a) or (b) generate a MC sample from the Square root Inverse Gamma distribution $Z = \sqrt{Y} = \sqrt{1/X}$.

Solution:

```
SIG_sample <- sqrt(1 / G_sample)
```

1. Plot a traceplot of the MC sample
2. Plot a histogram of the MC sample with the overlayed true density curve from Exercise 5 above;

3. Summarize the MC sample by computing the sample mean and the sample median.

```
library(knitr)
df_mean_median <- matrix(NA, nrow = 3, ncol = 2)
rownames(df_mean_median) <- c("G", "IG", "SIG")
colnames(df_mean_median) <- c("mean", "median")
df_mean_median <- data.frame(df_mean_median)
df_mean_median$mean <- c(
  mean(G_sample), mean(IG_sample),
  mean(SIG_sample)
)
df_mean_median$median <- c(
  median(G_sample), median(IG_sample),
  median(SIG_sample)
)
kable(df_mean_median)
```

	mean	median
G	3.9667004	3.2625686
IG	0.6143637	0.3065072
SIG	0.6672328	0.5536309

Question: What is the relation between the sample medians of X, Y, and Z? As we can see below 1 divided by the median of the Gamma sample is equal to the median of the Inverse Gamma sample. Plus, the square root of the reciprocal of the median of the Gamma sample equals to the median of the SIG sample. The equality holds because of the invariance property of the median.

Question: Does this relation also apply to sample means of X, Y, and Z? Why not? As it can be seen below, similar relation doesn't hold for the means of the samples because mean is not invariant to one-to-one transformations. For more details see Jensen's inequality in the book by Held and Sabanés Bové, 2020, in Appendix A.3.7, on page 354. Transformation functions are convex in both cases, which means that the expectation of the transformation is larger than the transformation of the expectation. This is the behavior of the means we see in Table 1.

	1/G	IG	$\sqrt{1/G}$	SIG
median	0.306507	0.3065072	0.5536307	0.5536309
mean	0.2520987	0.6143637	0.5020943	0.6672328

Table 1: Median and mean for the inverse of gamma, inverse gamma, and square root of inverse of gamma and square root inverse gamma distributions.

```
par(pty = "s")
curve(dnorm(x, mean = mean_post_pred, sd = sd_post_pred),
  xlab = expression(m), ylab = "density", ylim = c(0, 0.07),
  from = 70, to = 250, lwd = 2, col = "black"
)
curve(dnorm(x, mean = mean_prior_pred, sd = sd_prior_pred),
  xlab = expression(m), ylab = "density",
  from = 70, to = 250, lwd = 2, add = TRUE, col = "red"
)
curve(dnorm(x, mean = mean_post, sd = sd_post),
  xlab = expression(m), ylab = "density",
  from = 70, to = 250, lwd = 2, add = TRUE, col = "blue"
)
abline(v = 200, lty = 2, lwd = 2, col = "darkgrey")
legend("topleft",
  legend = c("prior-pred", "post-pred", "post"),
  col = c("red", "black", "blue"), lwd = 2, bty = "n", cex = .7
)
```

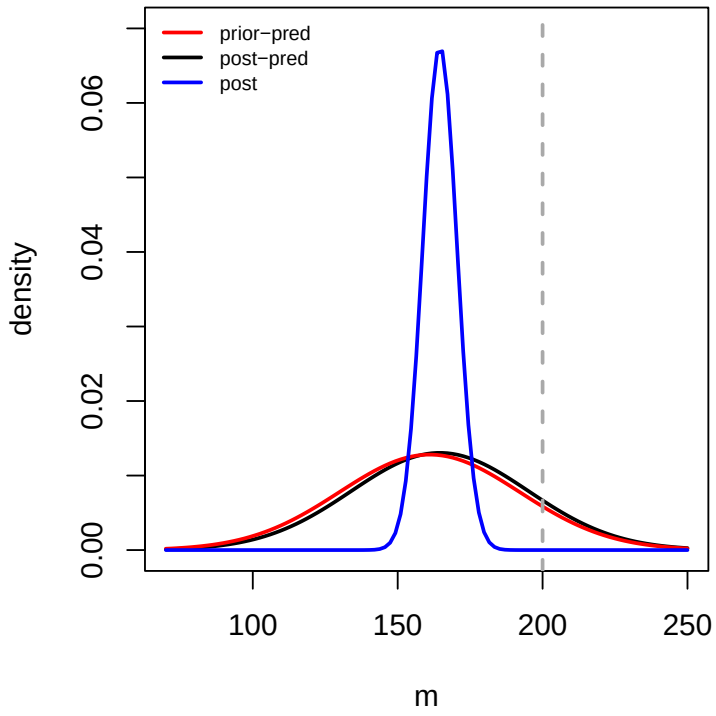

Figure 1: Prior predictive distribution, posterior predictive distribution for one future observation and posterior distribution. The dashed line marks `Height` = 200.

```r
a <- 1.6
b <- 0.4

# gamma
G_density <- function(x, a, b) b^a / gamma(a) * x^(a - 1) * exp(-b * x)
# inverse gamma
IG_density <- function(x, a, b) b^a / gamma(a) * x^(-(a + 1)) * exp(-b / x)
# square root inverse gamma
SIG_density <- function(x, a, b) 2 * b^a / gamma(a) * x^(-(2 * a + 1)) * exp(-b / (x^2))


curve(G_density(x, a, b), 0, 3, lwd = 2, ylim = c(0, 2.5), ylab = "density")
curve(IG_density(x, a, b), 0, 3, lwd = 2, col = "red", add = TRUE)
curve(SIG_density(x, a, b), 0, 3, lwd = 2, col = "blue", add = TRUE)
legend("topright",
  legend = c("G", "IG", "SIG"), col = c("black", "red", "blue"),
  lwd = 2, bty = "n"
)
```

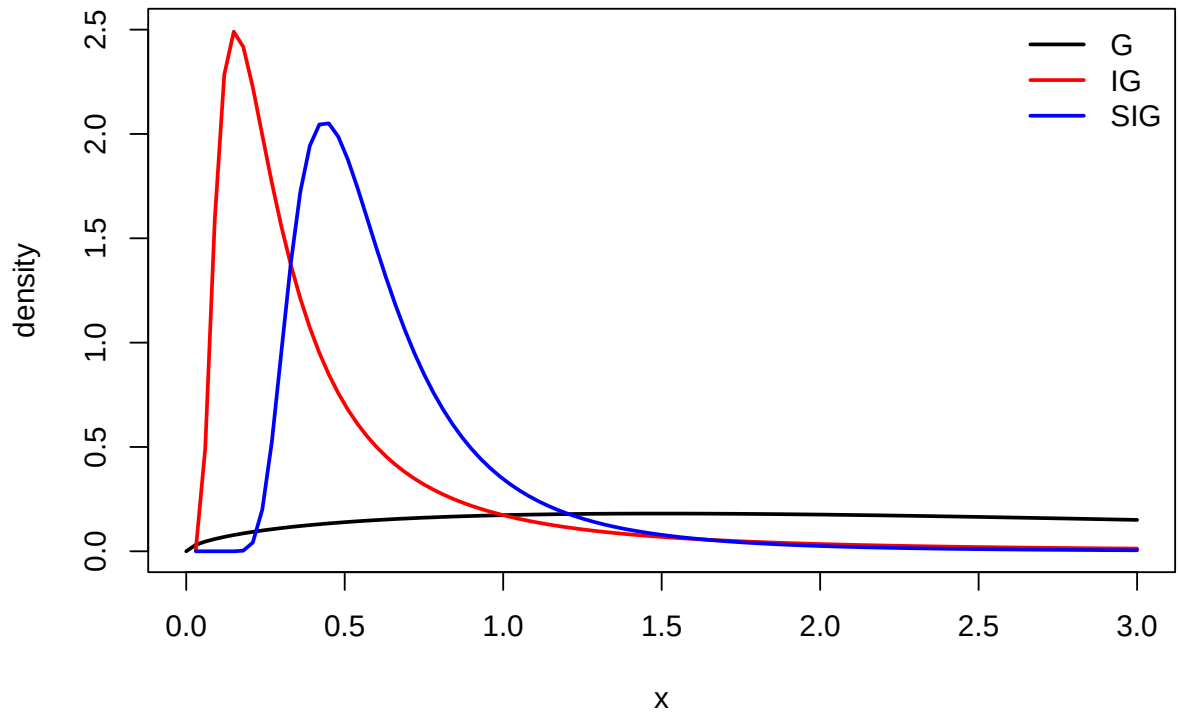

Figure 2: Density functions for G(1.6, 0.4), IG(1.6, 0.4) and SIG(1.6, 0.4) distributions.

```r
curve(IG_density(x, a, b), 0, 0.5, lwd = 2, col = "red", ylab = "density")
curve(SIG_density(x, a, b), 0, 0.5, lwd = 2, col = "blue", add = TRUE)
legend("topright",
  legend = c("IG", "SIG"), col = c("red", "blue"),
  lwd = 2, bty = "n"
)
```

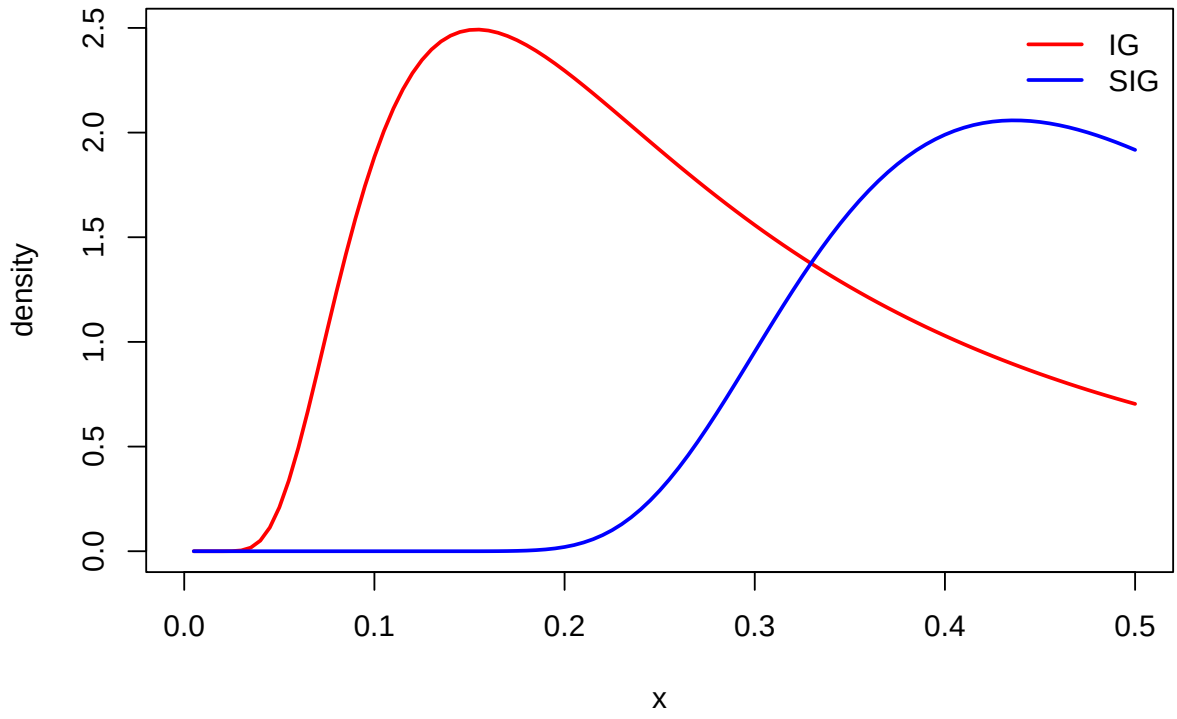

Figure 3: Density functions for IG(1.6, 0.4) and SIG(1.6, 0.4) distributions in the domain (0, 0.5).

```r
par(mfrow = c(3, 1), cex = 0.9)
plot(seq_along(G_sample), G_sample,
  type = "l", xlab = "M", ylab = "",
  main = "Traceplot X", cex = 0.5, ylim = c(0, 20), las = 1
)
plot(seq_along(IG_sample), IG_sample,
  type = "l", xlab = "M", ylab = "",
  main = "Traceplot Y", cex = 0.5, ylim = c(0, 20), las = 1
)
plot(seq_along(SIG_sample), SIG_sample,
  type = "l", xlab = "M", ylab = "",
  main = "Traceplot Z", cex = 0.5, ylim = c(0, 20), las = 1
)
```

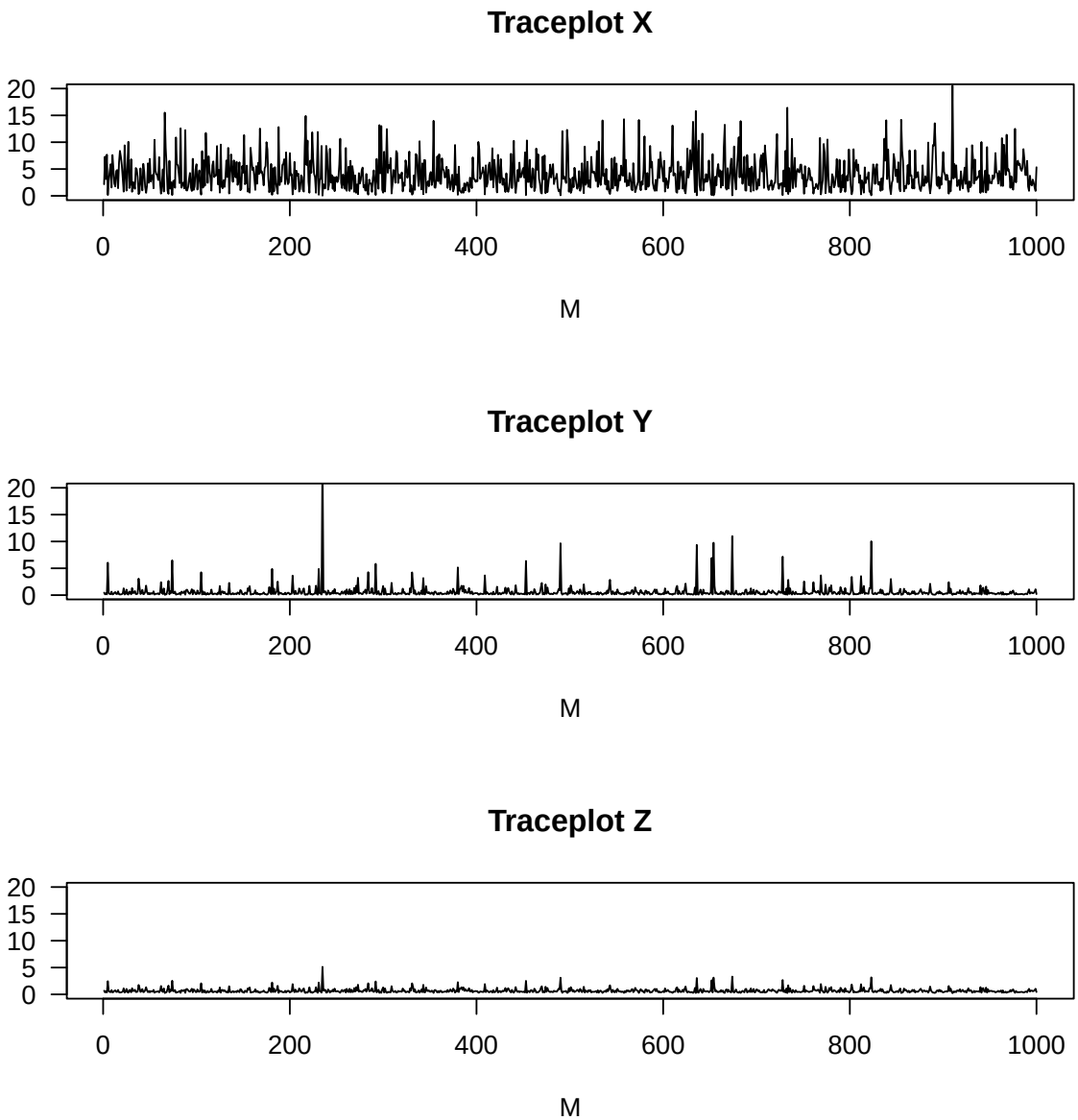

Figure 4: The traceplots for the samples of X (G), Y (IG) and Z (SIG).

```r
par(mfrow = c(1, 3), cex = 0.9)
hist(G_sample,
  probability = TRUE, ylim = c(0, 1), xlim = c(0, 20),
  breaks = 20, main = ""
)
curve(G_density(x, a, b), 0, 20, lwd = 2, add = TRUE)

hist(IG_sample,
  probability = TRUE, ylim = c(0, 1), xlim = c(0, 20),
  breaks = 20, main = ""
)
curve(IG_density(x, a, b), 0, 20, lwd = 2, add = TRUE)

hist(SIG_sample,
  probability = TRUE, ylim = c(0, 1), xlim = c(0, 20),
  breaks = 20, main = ""
)
curve(SIG_density(x, a, b), 0, 20, lwd = 2, add = TRUE)
```

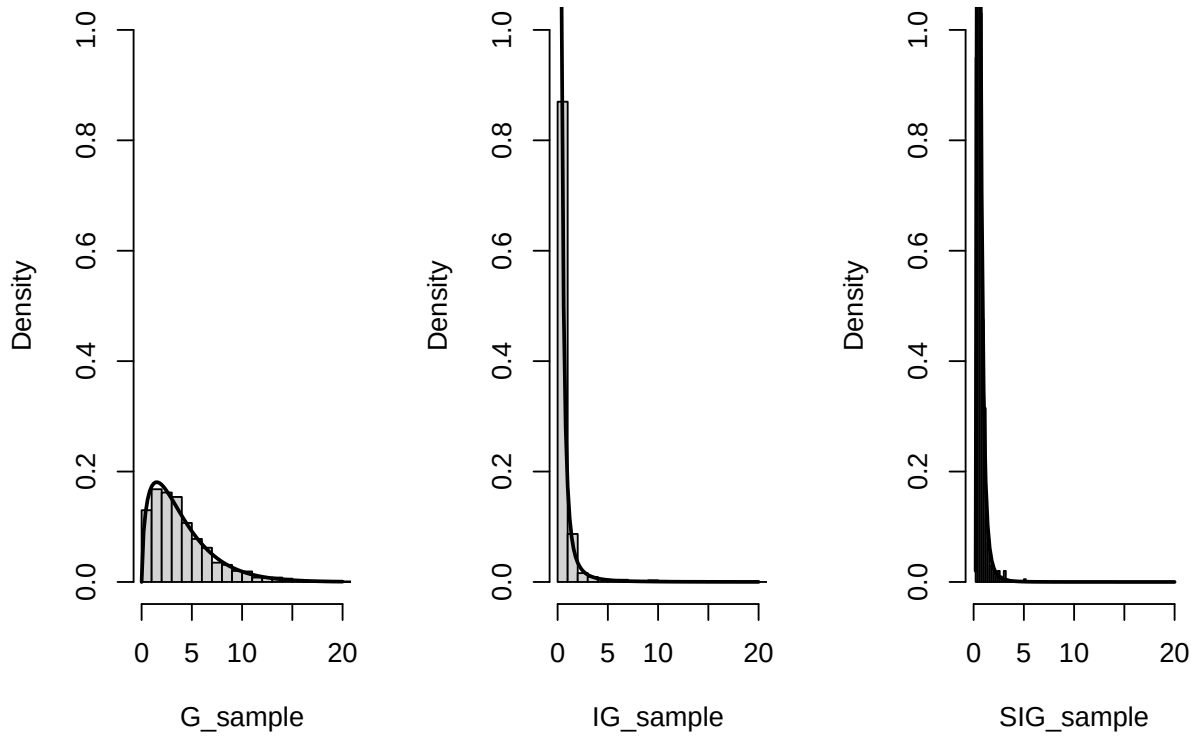

Figure 5: The histograms for the samples of X (G), Y (IG) and Z (SIG).