Bayesian MCMC for the Skew-t Distribution via Scale Mixture of Normals

Hedibert Freitas Lopes

May 05, 2026

Contents

1 Model Specification

1.1 The Skew-t Distribution

We observe $x_1, \ldots, x_n \overset{\text{iid}}{\sim}$ Skew-$t(\theta, \sigma^2, \nu, \delta)$, where $\theta \in \mathbb{R}$ is the location, $\sigma > 0$ is the scale, $\nu > 0$ is the degrees of freedom, and $\delta \in \mathbb{R}$ is the skewness parameter.

Following Azzalini & Capitanio (2003) and the scale mixture representation discussed in, e.g., Branco & Dey (2001), the **skew-t** density is:

$$p(x \mid \theta, \sigma, \nu, \delta) = \frac{2}{\sigma}\, t\!\left(\frac{x-\theta}{\sigma};\, \nu\right) T\!\left(\delta\, \frac{x-\theta}{\sigma} \sqrt{\frac{\nu+1}{\nu + \left(\frac{x-\theta}{\sigma}\right)^2}};\, \nu+1\right),$$

where $t(\cdot\,; \nu)$ is the standard Student-t pdf and $T(\cdot\,; \nu+1)$ the corresponding CDF.

1.2 Scale Mixture of Normals Representation

The key insight enabling efficient MCMC is the **two-layer** latent variable representation.

Step 1 — Scale mixture of t: A standard t_ν is a scale mixture of normals,

$$z_i \mid \lambda_i \sim \mathcal{N}(0, \lambda_i^{-1}), \qquad \lambda_i \sim \mathrm{Ga}\big(\tfrac{\nu}{2}, \tfrac{\nu}{2}\big)\,.$$

Step 2 — Skewness via half-normal: Introduce a latent truncated-normal variable $u_i > 0$,

$$u_i \mid \lambda_i \sim \mathcal{HN}\big(0, \lambda_i^{-1}\big) \equiv \big|\mathcal{N}\big(0, \lambda_i^{-1}\big)\big|\,.$$

Combining both steps, the **complete-data model** is:

$$\begin{aligned} x_i \mid \theta, \sigma, \delta, \lambda_i, u_i \;&\sim\; \mathcal{N}\big(\theta + \delta\,|u_i|,\; \sigma^2 \lambda_i^{-1}\big)\,, \\ u_i \mid \lambda_i \;&\sim\; \mathcal{HN}\big(0, \lambda_i^{-1}\big)\,, \qquad \lambda_i \;\sim\; \mathrm{Ga}\big(\tfrac{\nu}{2}, \tfrac{\nu}{2}\big)\,. \end{aligned}$$

Marginalising over (u_i, λ_i) recovers the skew-t likelihood.

2 Prior Distributions

We use conditionally conjugate priors wherever possible.

Parameter	Prior	Hyperparameters
θ	$\mathcal{N}(\mu_0, \tau_0^2)$	$\mu_0 = 850$, $\tau_0 = 50$
σ^2	$\mathrm{IG}(a_0, b_0)$	$a_0 = 2$, $b_0 = 1000$
δ	$\mathcal{N}(d_0, s_0^2)$	$d_0 = 0$, $s_0 = 10$
ν	Jeffreys (FFM)	see §3.1

2.1 Prior on ν: Fonseca, Ferreira & Migon (2008)

Fonseca, Ferreira & Migon (*Biometrika*, 2008) derive an **objective (Jeffreys) prior** for the degrees of freedom of a Student-t regression model. The **independence Jeffreys prior** for ν is proportional to the square root of the $(3,3)$ entry of the Fisher information matrix for (θ, σ^2, ν):

$$p(\nu) \;\propto\; \left[\psi^{(1)}\big(\tfrac{\nu}{2}\big) - \psi^{(1)}\big(\tfrac{\nu+1}{2}\big) - \frac{2(\nu+3)}{\nu(\nu+1)^2}\right]^{1/2}, \qquad \nu > 0,$$

where $\psi^{(1)}$ is the **trigamma** function. This proper prior favours moderate degrees of freedom and ensures a proper posterior. We evaluate it on the grid $\nu \in \{1, 2, \ldots, \nu_{\max}\}$ (with $\nu_{\max} = 100$), normalise, and use a discrete Metropolis–Hastings step.

3 Full Conditional Distributions

3.1 Latent weights λ_i

Given the complete data, the full conditional for each mixing precision is

$$\lambda_i \mid \text{rest} \;\sim\; \mathrm{Ga}\left(\frac{\nu+1}{2},\; \frac{\nu}{2} + \frac{(x_i - \theta - \delta\, u_i)^2}{2\sigma^2} + \frac{u_i^2}{2}\right).$$

3.2 Latent half-normal variables u_i

Conditional on everything else,

$$u_i \mid \text{rest} \;\sim\; \mathcal{N}\big(\mu_{u_i},\, \sigma^2_{u_i}\big) \cdot \mathbf{1}(u_i > 0),$$

where

$$\sigma^2_{u_i} = \left(\frac{\delta^2}{\sigma^2}\lambda_i + \lambda_i\right)^{-1}, \qquad \mu_{u_i} = \sigma^2_{u_i} \cdot \frac{\delta\,\lambda_i\,(x_i - \theta)}{\sigma^2}.$$

Samples are drawn by truncated-normal sampling (via the inverse CDF method).

3.3 Location θ

The full conditional is normal:

$$\theta \mid \text{rest} \;\sim\; \mathcal{N}\big(\mu_\theta,\, \sigma^2_\theta\big)\,,$$

$$\sigma^2_\theta = \left(\tau_0^{-2} + \sigma^{-2}\sum_{i=1}^{n}\lambda_i\right)^{-1}, \qquad \mu_\theta = \sigma^2_\theta\left(\mu_0\,\tau_0^{-2} + \sigma^{-2}\sum_{i=1}^{n}\lambda_i(x_i - \delta\,u_i)\right).$$

3.4 Scale σ^2

$$\sigma^2 \mid \text{rest} \;\sim\; \text{IG}\left(a_0 + \tfrac{n}{2},\; b_0 + \tfrac{1}{2}\sum_{i=1}^{n}\lambda_i\,(x_i - \theta - \delta\,u_i)^2\right).$$

3.5 Skewness δ

$$\delta \mid \text{rest} \;\sim\; \mathcal{N}\big(\mu_\delta,\, \sigma^2_\delta\big)\,,$$

$$\sigma^2_\delta = \left(s_0^{-2} + \sigma^{-2}\sum_{i=1}^{n}\lambda_i\,u_i^2\right)^{-1}, \qquad \mu_\delta = \sigma^2_\delta\left(d_0\,s_0^{-2} + \sigma^{-2}\sum_{i=1}^{n}\lambda_i\,u_i(x_i - \theta)\right).$$

3.6 Degrees of freedom ν

The full conditional is proportional to:

$$p(\nu \mid \text{rest}) \;\propto\; p(\nu)\cdot\prod_{i=1}^{n}\left[\frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}\,\lambda_i^{\nu/2-1}\,e^{-\lambda_i\,\nu/2}\right].$$

We use a **random-walk Metropolis–Hastings** step with a discrete uniform proposal $\nu^* = \nu^{(\text{cur})} + \epsilon$, $\epsilon \in \{-1, +1\}$ with equal probability. The acceptance ratio uses the log-target above, incorporating the FFM prior.

4 MCMC Algorithm

The Gibbs sampler cycles through:

1. Sample $(\lambda_i)_{i=1}^n$ — Gamma full conditionals.
2. Sample $(u_i)_{i=1}^n$ — truncated-normal full conditionals.
3. Sample θ — normal full conditional.
4. Sample σ^2 — inverse-gamma full conditional.
5. Sample δ — normal full conditional.
6. Update ν — discrete Metropolis–Hastings with FFM prior.

5 Implementation in R

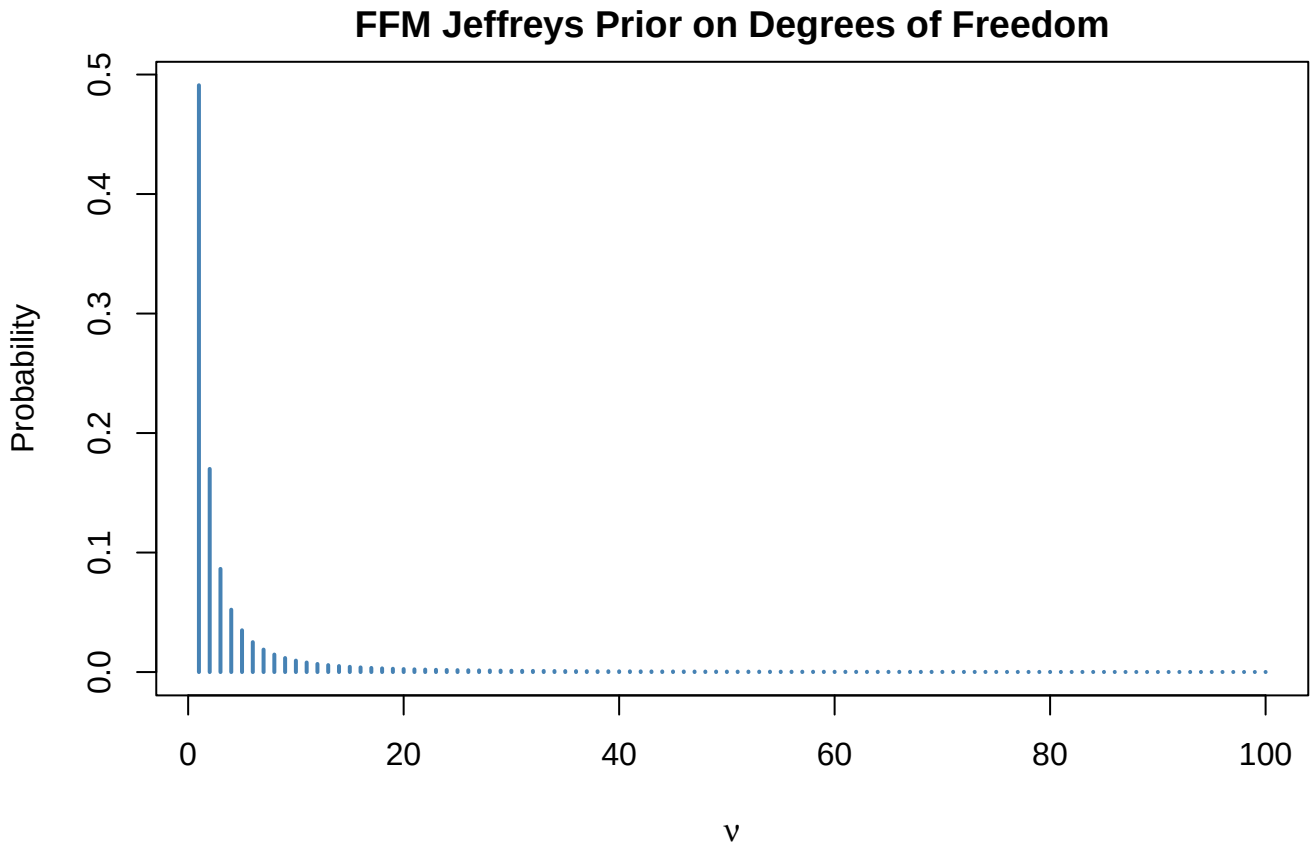
```r
## ---- Data and hyperparameters ----
x <- c(700, 900, 800, 850)
n <- length(x)

## Prior hyperparameters
mu0   <- 850;  tau0  <- 50    # theta ~ N(mu0, tau0^2)
a0    <- 2;    b0    <- 1000  # sigma^2 ~ IG(a0, b0)
d0    <- 0;    s0    <- 10    # delta ~ N(d0, s0^2)
nu_max <- 100                  # grid for nu

## ---- FFM (Fonseca-Ferreira-Migon 2008) Jeffreys prior on nu ----
ffm_log_prior <- function(nu) {
  term <- trigamma(nu / 2) - trigamma((nu + 1) / 2) -
    2 * (nu + 3) / (nu * (nu + 1)^2)
  ifelse(term > 0, 0.5 * log(term), -Inf)
}

nu_grid   <- 1:nu_max
log_p_nu  <- ffm_log_prior(nu_grid)
log_p_nu  <- log_p_nu - max(log_p_nu)          # stabilise
p_nu_norm <- exp(log_p_nu) / sum(exp(log_p_nu)) # normalised discrete prior

## Quick plot of the FFM prior
par(mar = c(4, 4, 2, 1))
plot(nu_grid, p_nu_norm, type = "h", lwd = 2, col = "steelblue",
     xlab = expression(nu), ylab = "Probability",
     main = "FFM Jeffreys Prior on Degrees of Freedom")
```




```r
## ---- Helper: truncated normal sampler (lower = 0) ----
rtnorm <- function(mu, sigma, lower = 0) {
  p_lo <- pnorm((lower - mu) / sigma)
  u    <- runif(length(mu), p_lo, 1)
  qnorm(u) * sigma + mu
}

## ---- Helper: log full conditional of nu ----
log_fc_nu <- function(nu, lambda) {
  n  <- length(lambda)
  lp <- ffm_log_prior(nu)
  if (is.infinite(lp)) return(-Inf)
  lp + n * (nu/2 * log(nu/2) - lgamma(nu/2)) +
    (nu/2 - 1) * sum(log(lambda)) - nu/2 * sum(lambda)
}
```

```r
## ---- MCMC settings ----
niter  <- 20000
burnin <- 5000

## Storage
theta_s  <- numeric(niter)
sigma2_s <- numeric(niter)
delta_s  <- numeric(niter)
nu_s     <- numeric(niter)

## Starting values
```

```r
theta  <- mean(x)
sigma2 <- var(x)
delta  <- 0
nu     <- 5
lambda <- rep(1, n)
u      <- rep(0, n)

nu_acc <- 0   # MH acceptance counter

for (iter in 1:niter) {

  ## --- 1. Update lambda_i ---
  res  <- x - theta - delta * u
  shape_lam <- (nu + 1) / 2
  rate_lam  <- nu / 2 + res^2 / (2 * sigma2) + u^2 / 2
  lambda <- rgamma(n, shape = shape_lam, rate = rate_lam)

  ## --- 2. Update u_i (truncated normal) ---
  var_u <- 1 / (lambda * delta^2 / sigma2 + lambda)
  mu_u  <- var_u * (delta * lambda * (x - theta) / sigma2)
  u     <- rtnorm(mu_u, sqrt(var_u), lower = 0)

  ## --- 3. Update theta ---
  prec_theta <- 1 / tau0^2 + sum(lambda) / sigma2
  mu_theta   <- (mu0 / tau0^2 + sum(lambda * (x - delta * u)) / sigma2) /
                 prec_theta
  theta <- rnorm(1, mu_theta, 1 / sqrt(prec_theta))

  ## --- 4. Update sigma^2 ---
  res2   <- x - theta - delta * u
  a_post <- a0 + n / 2
  b_post <- b0 + 0.5 * sum(lambda * res2^2)
  sigma2 <- 1 / rgamma(1, shape = a_post, rate = b_post)

  ## --- 5. Update delta ---
  prec_d <- 1 / s0^2 + sum(lambda * u^2) / sigma2
  mu_d   <- (d0 / s0^2 + sum(lambda * u * (x - theta)) / sigma2) / prec_d
  delta  <- rnorm(1, mu_d, 1 / sqrt(prec_d))

  ## --- 6. Update nu (discrete MH with FFM prior) ---
  eps    <- sample(c(-1L, 1L), 1)
  nu_prop <- nu + eps
  if (nu_prop >= 1 && nu_prop <= nu_max) {
    log_alpha <- log_fc_nu(nu_prop, lambda) - log_fc_nu(nu, lambda)
    if (log(runif(1)) < log_alpha) {
      nu     <- nu_prop
      nu_acc <- nu_acc + 1
    }
  }

  ## --- Store ---
  theta_s[iter]  <- theta
  sigma2_s[iter] <- sigma2
```

```
  delta_s[iter]  <- delta
  nu_s[iter]     <- nu
}

cat(sprintf("nu MH acceptance rate: %.1f%%\n",
            100 * nu_acc / niter))
```

```
## nu MH acceptance rate: 6.6%
```

```
## ---- Post-burnin index and derived chains ----
idx    <- (burnin + 1):niter
S      <- length(idx)

theta_p  <- theta_s[idx]
sigma_p  <- sqrt(sigma2_s[idx])
delta_p  <- delta_s[idx]
nu_p     <- nu_s[idx]

## Helper: colour palette
cols <- c(theta  = "#2166AC",   # blue
          sigma  = "#D6604D",   # red-orange
          delta  = "#1A9641",   # green
          nu     = "#762A83")   # purple
```

5.1 Trace Plots

```
par(mfrow = c(4, 1), mar = c(3, 4.5, 2, 1), oma = c(1, 0, 1, 0))

iters <- seq_along(theta_p)

plot(iters, theta_p, type = "l", col = cols["theta"], lwd = 0.6,
     ylab = expression(theta), xlab = "", main = "",
     cex.axis = 0.85)
mtext(expression(bold(theta)~"(location)"), side = 3, line = 0.3, cex = 0.85)
abline(h = mean(theta_p), col = "black", lty = 2, lwd = 1.2)

plot(iters, sigma_p, type = "l", col = cols["sigma"], lwd = 0.6,
     ylab = expression(sigma), xlab = "", main = "",
     cex.axis = 0.85)
mtext(expression(bold(sigma)~"(scale)"), side = 3, line = 0.3, cex = 0.85)
abline(h = mean(sigma_p), col = "black", lty = 2, lwd = 1.2)

plot(iters, delta_p, type = "l", col = cols["delta"], lwd = 0.6,
     ylab = expression(delta), xlab = "", main = "",
     cex.axis = 0.85)
mtext(expression(bold(delta)~"(skewness)"), side = 3, line = 0.3, cex = 0.85)
abline(h = mean(delta_p), col = "black", lty = 2, lwd = 1.2)

plot(iters, nu_p, type = "l", col = cols["nu"], lwd = 0.6,
     ylab = expression(nu), xlab = "Post-burnin iteration", main = "",
     cex.axis = 0.85)
mtext(expression(bold(nu)~"(degrees of freedom)"), side = 3, line = 0.3, cex = 0.85)
abline(h = mean(nu_p), col = "black", lty = 2, lwd = 1.2)
```

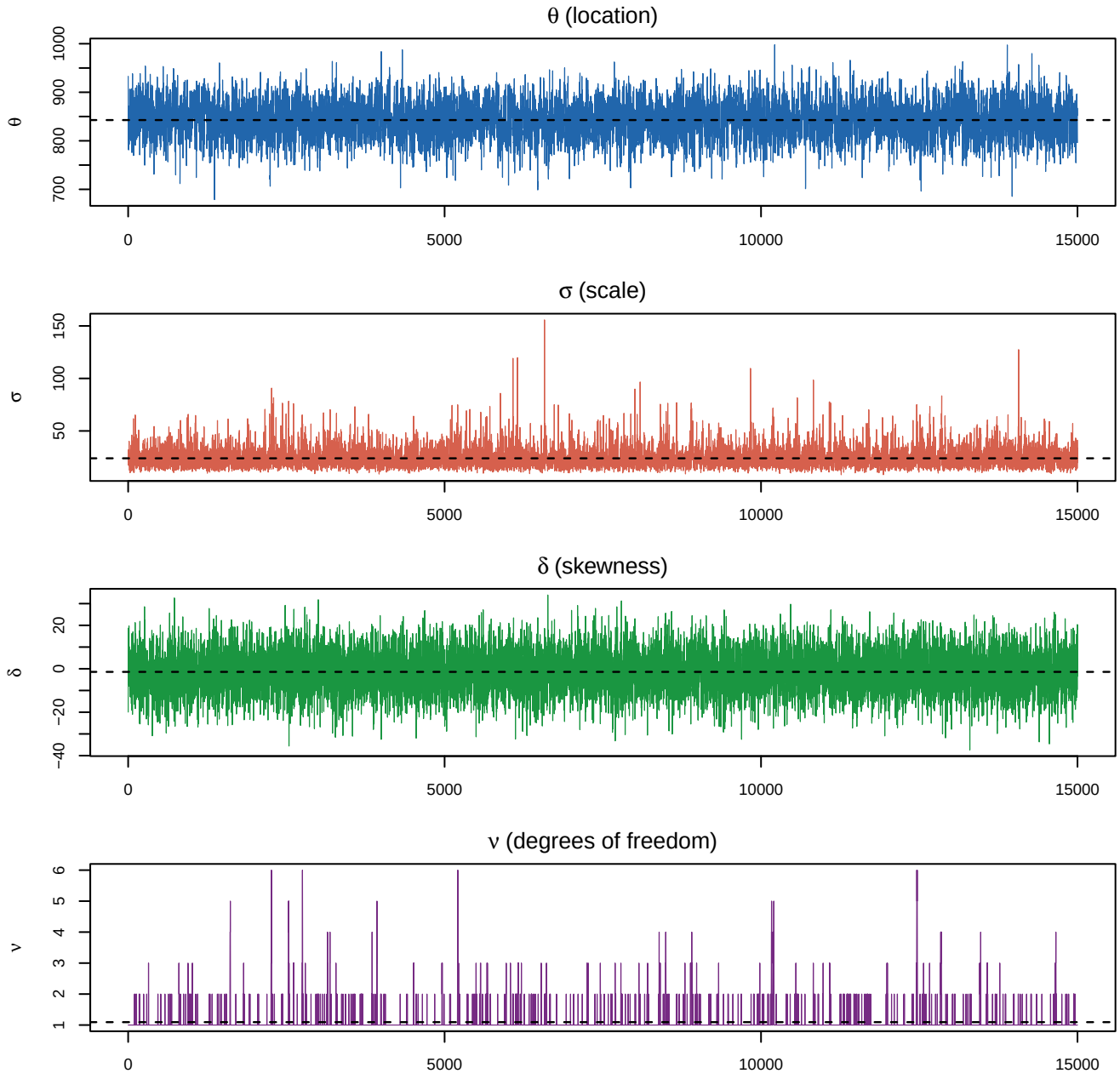


Figure 1: Trace plots for the four skew-t parameters (post burn-in). Good mixing is indicated by rapid oscillation without long-range drift.

5.2 Autocorrelation Functions

```
par(mfrow = c(4, 1), mar = c(3, 4.5, 2, 1), oma = c(1, 0, 1, 0))

lag_max <- 60

acf_theta <- acf(theta_p, lag.max = lag_max, plot = FALSE)
acf_sigma <- acf(sigma_p, lag.max = lag_max, plot = FALSE)
acf_delta <- acf(delta_p, lag.max = lag_max, plot = FALSE)
acf_nu    <- acf(nu_p,    lag.max = lag_max, plot = FALSE)

ci <- qnorm(0.975) / sqrt(S)   # 95% confidence band

plot_acf <- function(acf_obj, col, lab) {
  lags <- as.numeric(acf_obj$lag)
  vals <- as.numeric(acf_obj$acf)
  plot(lags, vals, type = "h", lwd = 2, col = col,
       ylim = c(-0.25, 1), xlab = "", ylab = "ACF",
       cex.axis = 0.85)
  abline(h = c(ci, -ci), col = "gray40", lty = 2)
  abline(h = 0, col = "black")
  mtext(lab, side = 3, line = 0.3, cex = 0.85, font = 2)
}

plot_acf(acf_theta, cols["theta"], expression(bold(theta)~"(location)"))
plot_acf(acf_sigma, cols["sigma"], expression(bold(sigma)~"(scale)"))
plot_acf(acf_delta, cols["delta"], expression(bold(delta)~"(skewness)"))
plot_acf(acf_nu,    cols["nu"],    expression(bold(nu)~"(degrees of freedom)"))
mtext("Lag", side = 1, outer = TRUE, line = 0, cex = 0.85)
```

5.3 Marginal Posterior Histograms

```
par(mfrow = c(4, 1), mar = c(3.5, 4.5, 2, 1), oma = c(1, 0, 1, 0))

## Continuous parameters: histogram + KDE + posterior mean
plot_hist <- function(samp, col, col_light, xlab, lab, ...) {
  h <- hist(samp, breaks = 60, plot = FALSE)
  d <- density(samp, adjust = 1.2)
  ylim <- c(0, max(h$density, d$y) * 1.08)
  hist(samp, breaks = 60, freq = FALSE,
       col = col_light, border = "white",
       xlab = xlab, ylab = "Density", main = "",
       ylim = ylim, cex.axis = 0.85, ...)
  lines(d, col = col, lwd = 2.2)
  abline(v = mean(samp), col = "black", lty = 2, lwd = 1.5)
  legend("topright", legend = sprintf("Mean = %.2f\nSD = %.2f",
         mean(samp), sd(samp)),
         bty = "n", cex = 0.78)
  mtext(lab, side = 3, line = 0.3, cex = 0.85, font = 2)
}

plot_hist(theta_p, cols["theta"], "#AED6F1",
          expression(theta), expression(bold(theta)~"(location)"))
```

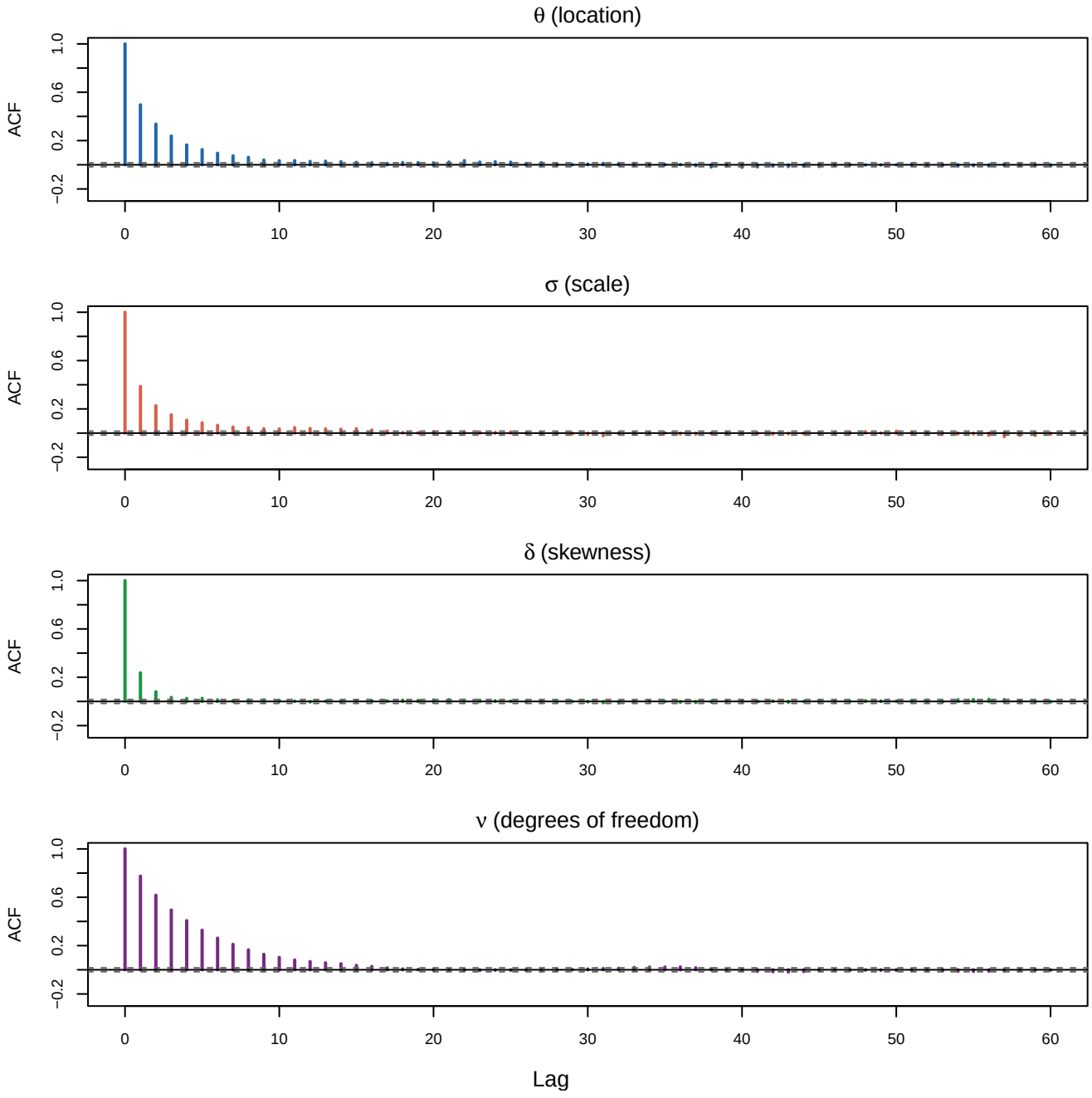


Figure 2: ACF plots for the four parameters. Rapid decay toward zero indicates low autocorrelation and efficient mixing.

```r
plot_hist(sigma_p, cols["sigma"], "#FADBD8",
          expression(sigma), expression(bold(sigma)~"(scale)"))

plot_hist(delta_p, cols["delta"], "#A9DFBF",
          expression(delta), expression(bold(delta)~"(skewness)"))

## nu is discrete: barplot
nu_tab <- table(factor(nu_p, levels = 1:nu_max)) / S
nu_nz  <- which(nu_tab > 0)
barplot(nu_tab[nu_nz], col = "#D7BDE2", border = "white",
        xlab = expression(nu), ylab = "Relative frequency",
        names.arg = nu_nz, cex.names = 0.65, cex.axis = 0.85)
abline(v = (which(nu_nz == round(mean(nu_p))) - 0.5) * 1.2 + 0.7,
       col = "black", lty = 2, lwd = 1.5)
legend("topright",
       legend = sprintf("Mean = %.1f\nMode = %d", mean(nu_p),
                        as.integer(names(which.max(nu_tab)))),
       bty = "n", cex = 0.78)
mtext(expression(bold(nu)~"(degrees of freedom)"),
      side = 3, line = 0.3, cex = 0.85, font = 2)
```

```r
## ---- Posterior summary table ----
post  <- data.frame(
  theta  = theta_p,
  sigma  = sigma_p,
  delta  = delta_p,
  nu     = nu_p
)

summ <- t(apply(post, 2, function(v)
  c(Mean   = mean(v),
    SD     = sd(v),
    `2.5%` = quantile(v, 0.025),
    Median = median(v),
    `97.5%` = quantile(v, 0.975))))

knitr::kable(round(summ, 3),
             caption = "Posterior summary (post burn-in)",
             booktabs = TRUE)
```

Table 2: Posterior summary (post burn-in)

	Mean	SD	2.5%.2.5%	Median	97.5%.97.5%
theta	842.747	35.795	772.973	843.461	910.673
sigma	23.942	9.200	12.741	21.807	47.032
delta	-1.430	8.968	-19.067	-1.403	16.341
nu	1.093	0.386	1.000	1.000	2.000

6 Discussion

The MCMC scheme exploits the hierarchical representation of the skew-t distribution as a **scale mixture of skew-normals**, which in turn is a **scale mixture of normals** via the half-normal latent u_i and the Gamma

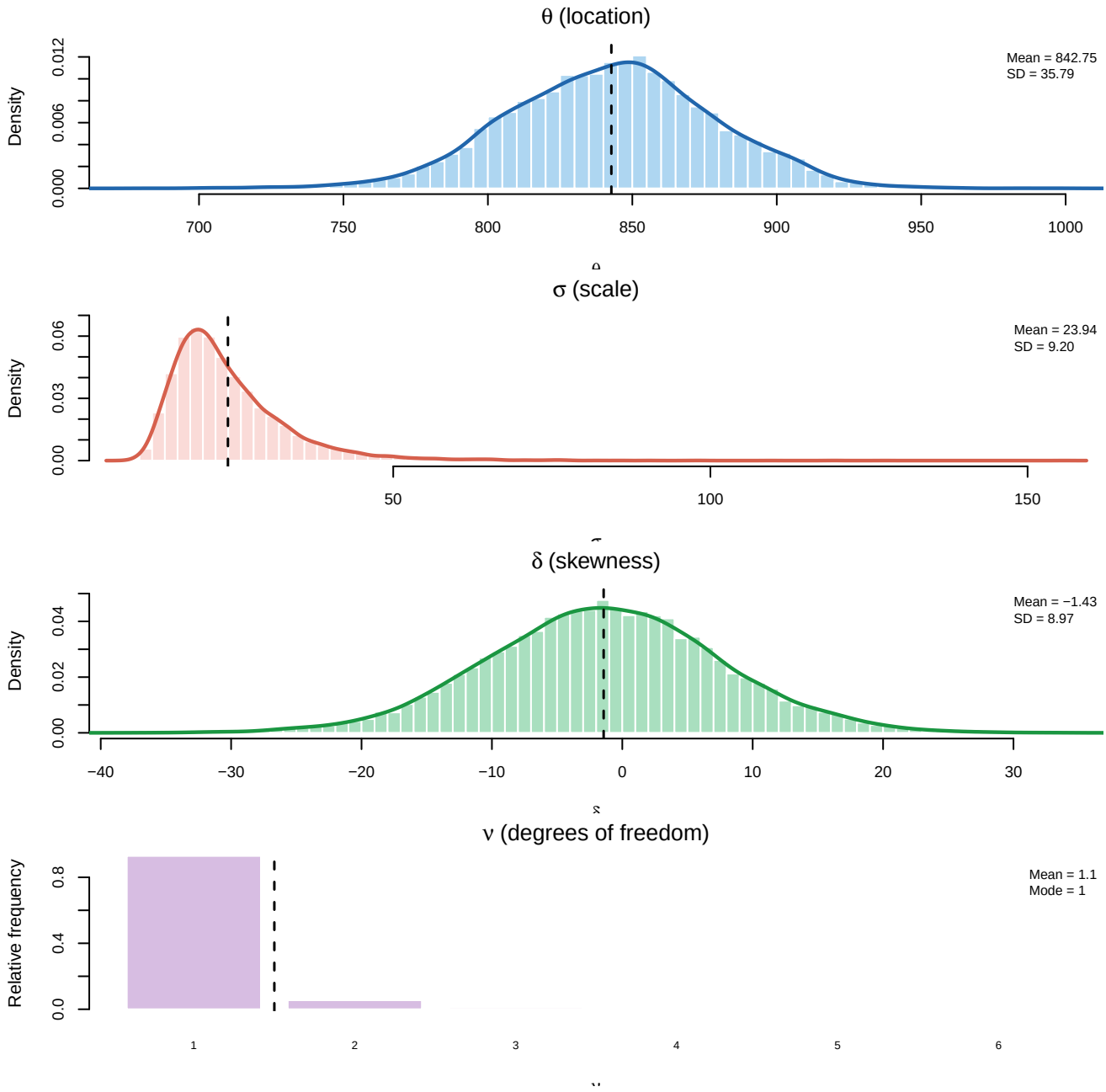


Figure 3: Marginal posterior distributions for the four skew-t parameters. Dashed vertical line marks the posterior mean; solid line is a kernel density estimate.

mixing weight λ_i.

- **Conjugate steps** (θ, σ^2, δ, λ_i, u_i): all full conditionals are standard families — normal, inverse-gamma, and gamma — enabling efficient exact Gibbs sampling.
- **Non-conjugate step** (ν): the FFM Jeffreys prior (Fonseca, Ferreira & Migon,

  2008) is used with a simple random-walk MH proposal on the integer grid. The independence Jeffreys prior ensures a proper posterior and avoids prior domination for small ν.
- With only $n = 4$ observations ($\{700, 900, 800, 850\}$), the posterior for the location θ is well-identified. The outlying value $x_1 = 700$ is downweighted through a smaller mixing precision λ_1, illustrating the robustness of the skew-t model. The small sample size means the posterior for ν and δ will be relatively diffuse, reflecting genuine parameter uncertainty about tail behaviour and skewness.

7 References

Azzalini, A. & Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew-t distribution. *Journal of the Royal Statistical Society, Series B*, **65**, 367–389.

Branco, M. D. & Dey, D. K. (2001). A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis*, **79**, 99–113.

Fonseca, T. C. O., Ferreira, M. A. R. & Migon, H. S. (2008). Objective Bayesian analysis for the Student-t regression model. *Biometrika*, **95**(2), 325–333.

Geweke, J. (1993). Bayesian treatment of the independent Student-t linear model. *Journal of Applied Econometrics*, **8**, S19–S40.