Bayesian modelling

Introduction

Léo Belzile, HEC Montréal

2025

Distribution and density function

Let $\boldsymbol{X} \in \mathbb{R}^d$ be a random vector with distribution function

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = \Pr(\boldsymbol{X} \leq \boldsymbol{x}) = \Pr(X_1 \leq x_1, \ldots, X_d \leq x_d).$$

If the distribution of $\boldsymbol{X}$ is absolutely continuous,

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = \int_{-\infty}^{x_d} \cdots \int_{-\infty}^{x_1} f_{\boldsymbol{X}}(z_1, \ldots, z_d) \mathrm{d}z_1 \cdots \mathrm{d}z_d,$$

where $f_{\boldsymbol{X}}(\boldsymbol{x})$ is the joint **density function**.

Mass function

By abuse of notation, we denote the mass function in the discrete case

$$0 \leq f_{\boldsymbol{X}}(\boldsymbol{x}) = \Pr(X_1 = x_1, \ldots, X_d = x_d) \leq 1.$$

The support is the set of non-zero density/probability total probability over all points in the support,

$$\sum_{\boldsymbol{x} \in \mathsf{supp}(\boldsymbol{X})} f_{\boldsymbol{X}}(\boldsymbol{x}) = 1.$$

Marginal distribution

The **marginal distribution** of a subvector $\boldsymbol{X}_{1:k} = (X_1, \ldots, X_k)^\top$ is

$$\begin{aligned} F_{\boldsymbol{X}_{1:k}}(\boldsymbol{x}_{1:k}) &= \Pr(\boldsymbol{X}_{1:k} \leq \boldsymbol{x}_{1:k}) \\ &= F_{\boldsymbol{X}}(x_1, \ldots, x_k, \infty, \ldots, \infty). \end{aligned}$$

Marginal density

The **marginal density** $f_{\boldsymbol{X}_{1:k}}(\boldsymbol{x}_{1:k})$ of an absolutely continuous subvector $\boldsymbol{X}_{1:k} = (X_1, \ldots, X_k)^\top$ is

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\boldsymbol{X}}(x_1, \ldots, x_k, z_{k+1}, \ldots, z_d) \mathrm{d}z_{k+1} \cdots \mathrm{d}z_d.$$

through integration from the joint density.

Conditional distribution

The conditional distribution function of $\boldsymbol{Y}$ given $\boldsymbol{X} = \boldsymbol{x}$, is

$$f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}; \boldsymbol{x}) = \frac{f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y})}{f_{\boldsymbol{X}}(\boldsymbol{x})}$$

for any value of $\boldsymbol{x}$ in the support of $\boldsymbol{X}$.

Conditional and marginal for contingency table

Consider a bivariate distribution for (Y_1, Y_2) supported on $\{1,2,3\} \times \{1,2\}$ whose joint probability mass function is given in Table 1

Table 1: Bivariate mass function with probability of each outcome for (Y_1, Y_2).

	$Y_1 = 1$	$Y_1 = 2$	$Y_1 = 3$	**row total**
$Y_2 = 1$	0.20	0.3	0.10	0.6
$Y_2 = 2$	0.15	0.2	0.05	0.4
col. total	0.35	0.5	0.15	1.0

Calculations for the marginal distribution

The marginal distribution of Y_1 is obtain by looking at the total probability for each row/column, e.g.,

$$\Pr(Y_1 = i) = \Pr(Y_1 = i, Y_2 = 1) + \Pr(Y_1 = i, Y_2 = 2).$$

- $\Pr(Y_1 = 1) = 0.35, \Pr(Y_1 = 2) = 0.5,$ $\Pr(Y_1 = 3) = 0.15.$
- $\Pr(Y_2 = 1) = 0.6$ and $\Pr(Y_2 = 2) = 0.4$

Conditional distribution

The conditional distribution

$$\Pr(Y_2 = i \mid Y_1 = 2) = \frac{\Pr(Y_1 = 2, Y_2 = i)}{\Pr(Y_1 = 2)},$$

so

$$\begin{aligned}\Pr(Y_2 = 1 \mid Y_1 = 2) &= 0.3/0.5 = 0.6\\ \Pr(Y_2 = 2 \mid Y_1 = 2) &= 0.4.\end{aligned}$$

Independence

Vectors $\boldsymbol{Y}$ and $\boldsymbol{X}$ are independent if

$$F_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y}) = F_{\boldsymbol{X}}(\boldsymbol{x})F_{\boldsymbol{Y}}(\boldsymbol{y})$$

for any value of $\boldsymbol{x}, \boldsymbol{y}$.

The joint density, if it exists, also factorizes

$$f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y}) = f_{\boldsymbol{X}}(\boldsymbol{x})f_{\boldsymbol{Y}}(\boldsymbol{y}).$$

If two subvectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, then the conditional density $f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y};\boldsymbol{x})$ equals the marginal $f_{\boldsymbol{Y}}(\boldsymbol{y})$.

Law of iterated expectation and variance

Let $\boldsymbol{Z}$ and $\boldsymbol{Y}$ be random vectors. The expected value of $\boldsymbol{Y}$ is

$$\mathsf{E}_{\boldsymbol{Y}}(\boldsymbol{Y}) = \mathsf{E}_{\boldsymbol{Z}}\left\{\mathsf{E}_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{Y})\right\}.$$

The **tower** property gives a law of iterated variance

$$\mathsf{Va}_{\boldsymbol{Y}}(\boldsymbol{Y}) = \mathsf{E}_{\boldsymbol{Z}}\left\{\mathsf{Va}_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{Y})\right\} + \mathsf{Va}_{\boldsymbol{Z}}\left\{\mathsf{E}_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{Y})\right\}.$$

Poisson distribution

The Poisson distribution has mass

$$f(x) = \mathsf{Pr}(Y = x) = \frac{\exp(-\lambda)\lambda^y}{\Gamma(y+1)}, \quad x = 0, 1, 2, \ldots$$

where $\Gamma(\cdot)$ denotes the gamma function.

The parameter λ of the Poisson distribution is both the expectation and the variance of the distribution, meaning

$$\mathsf{E}(Y) = \mathsf{Va}(Y) = \lambda.$$

Gamma distribution

A gamma distribution with shape $\alpha > 0$ and rate $\beta > 0$, denoted $Y \sim$ **gamma**(α, β), has density

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x), \qquad x \in (0, \infty),$$

where $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} \exp(-t)\mathrm{d}t$ is the gamma function.

Poisson with random scale

To handle overdispersion in count data, take

$$
\begin{aligned}
Y \mid \Lambda = \lambda &\sim \mathsf{Poisson}(\lambda) \\
\Lambda &\sim \mathsf{Gamma}(k\mu, k).
\end{aligned}
$$

The joint density of Y and Λ on $\mathbb{N} = \{0, 1, \ldots\} \times \mathbb{R}_{+}$ is

$$
\begin{aligned}
f(y, \lambda) &= f(y \mid \lambda) f(\lambda) \\
&= \frac{\lambda^{y} \exp(-\lambda)}{\Gamma(y+1)} \frac{k^{k\mu} \lambda^{k\mu - 1} \exp(-k\lambda)}{\Gamma(k\mu)}
\end{aligned}
$$

Conditional distribution

The conditional distribution of $\Lambda \mid Y = y$ can be found by considering only terms that are function of λ, whence

$$f(\lambda \mid Y = y) \stackrel{\lambda}{\propto} \lambda^{y+k\mu-1} \exp\{-(k+1)\lambda\}$$

so $\Lambda \mid Y = y \sim \mathsf{gamma}(k\mu + y, k + 1)$.

Marginal density of Poisson mean mixture

$$f(y) = \frac{f(y,\lambda)}{f(\lambda \mid y)} = \frac{\frac{\lambda^y \exp(-\lambda)}{\Gamma(y+1)} \frac{k^{k\mu}\lambda^{k\mu-1}\exp(-k\lambda)}{\Gamma(k\mu)}}{\frac{(k+1)^{k\mu+y}\lambda^{k\mu+y-1}\exp\{-(k+1)\lambda\}}{\Gamma(k\mu+y)}}$$

$$= \frac{\Gamma(k\mu+y)}{\Gamma(k\mu)\Gamma(y+1)} k^{k\mu}(k+1)^{-k\mu-y}$$

$$= \frac{\Gamma(k\mu+y)}{\Gamma(k\mu)\Gamma(y+1)} \left(1-\frac{1}{k+1}\right)^{k\mu} \left(\frac{1}{k+1}\right)^{y}$$

The marginal of Y is negative binomial with prob. of success $1/(k+1)$.

Likelihood

The **likelihood** $L(\boldsymbol{\theta})$ is a function of the parameter vector $\boldsymbol{\theta}$ that gives the 'density' of a sample under a postulated distribution, treating the observations as fixed,

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = f(\boldsymbol{y}; \boldsymbol{\theta}).$$

Likelihood for independent observations

If the joint density factorizes,

$$L(\boldsymbol{\theta};\boldsymbol{y}) = \prod_{i=1}^{n} f_i(y_i;\boldsymbol{\theta}) = f_1(y_1;\boldsymbol{\theta}) \times \cdots \times f_n(y_n;\boldsymbol{\theta}).$$

The corresponding log likelihood function for independent and identically distributions observations is

$$\ell(\boldsymbol{\theta};\boldsymbol{y}) = \sum_{i=1}^{n} \ln f(y_i;\boldsymbol{\theta})$$

Score

Let $\ell(\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$, be the log likelihood function. The gradient of the log likelihood, termed **score** is the p-vector

$$U(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

Information matrix

The **observed information matrix** is the hessian of the negative log likelihood,

$$j(\boldsymbol{\theta}; \boldsymbol{y}) = -\frac{\partial^2 \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top},$$

evaluated at the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$, so $j(\widehat{\boldsymbol{\theta}})$.

Expected information

Under regularity conditions, the **expected information**, also called **Fisher information** matrix, is

$$i(\boldsymbol{\theta}) = \mathsf{E}\left\{U(\boldsymbol{\theta};\boldsymbol{Y})U(\boldsymbol{\theta};\boldsymbol{Y})^\top\right\} = \mathsf{E}\left\{j(\boldsymbol{\theta};\boldsymbol{Y})\right\}$$

Note on information matrices

Information matrices are symmetric and provide information about the variability of $\widehat{\boldsymbol{\theta}}$.

The information of an iid sample of size n is n times that of a single observation

- information accumulates at a linear rate.

Example: random right-censoring

Consider a survival analysis problem for independent time-to-event data subject to (noninformative) random right-censoring. We observe

- failure times $Y_i (i = 1, \ldots, n)$ drawn from $F(\cdot; \boldsymbol{\theta})$ supported on $(0, \infty)$
- independent binary censoring indicators $C_i \in \{0, 1\}$, with 0 indicating right-censoring and $C_i = 1$ observed failure time.

Likelihood contribution with censoring

If individual observation i has not experienced the event at the end of the collection period, then the likelihood contribution is $\Pr(Y > y) = 1 - F(y; \boldsymbol{\theta})$, where y_i is the maximum time observed for Y_i. We write the log likelihood

$$\ell(\boldsymbol{\theta}) = \sum_{i:c_i=0} \log\{1 - F(y_i; \boldsymbol{\theta})\} + \sum_{i:c_i=1} \log f(y_i; \boldsymbol{\theta})$$

Censoring and exponential data

Suppose for simplicity that $Y_i \sim \text{expo}(\lambda)$ and let $m = c_1 + \cdots + c_n$ denote the number of observed failure times. Then, the log likelihood and the Fisher information are

$$\ell(\lambda) = \lambda \sum_{i=1}^{n} y_i + \log \lambda m$$
$$i(\lambda) = m/\lambda^2$$

and the right-censored observations for the exponential model do not contribute to the information.

Information for the Gaussian distribution

Consider $Y \sim$ **Gauss**(μ, τ^{-1}), parametrized in terms of precision τ. The likelihood contribution for an n sample is, up to proportionality,

$$\ell(\mu, \tau) \propto \frac{n}{2}\log(\tau) - \frac{\tau}{2}\sum_{i=1}^{n}(Y_i^2 - 2\mu Y_i + \mu^2)$$

Gaussian information matrices

The observed and Fisher information matrices are

$$j(\mu, \tau) = \begin{pmatrix} n\tau & -\sum_{i=1}^{n}(Y_i - \mu) \\ -\sum_{i=1}^{n}(Y_i - \mu) & \frac{n}{2\tau^2} \end{pmatrix},$$

$$i(\mu, \tau) = n \begin{pmatrix} \tau & 0 \\ 0 & \frac{1}{2\tau^2} \end{pmatrix}$$

Since $\mathsf{E}(Y_i) = \mu$, the expected value of the off-diagonal entries of the Fisher information matrix are zero.

Example: first-order autoregressive process

Consider an $\mathsf{AR}(1)$ model of the form

$$Y_t = \mu + \phi(Y_{t-1} - \mu) + \varepsilon_t,$$

where

- ϕ is the lag-one correlation,
- μ the global mean and
- ε_t is an iid innovation with mean zero and variance σ^2.

If $|\phi| < 1$, the process is stationary, and the variance does not increase with t.

Markov property and likelihood decomposition

The Markov property states that the current realization depends on the past, $Y_t \mid Y_1, \ldots, Y_{t-1}$, only through the most recent value Y_{t-1}. The log likelihood thus becomes

$$\ell(\boldsymbol{\theta}) = \ln f(y_1) + \sum_{i=2}^{n} f(y_i \mid y_{i-1}).$$

Marginal of AR(1)

The $\mathsf{AR}(1)$ stationarity process has unconditional moments

$$\mathsf{E}(Y_t) = \mu, \qquad \mathsf{Var}(Y_t) = \sigma^2/(1-\phi^2).$$

The $\mathsf{AR}(1)$ process is first-order Markov since the conditional distribution $f(Y_t \mid Y_{t-1}, \ldots, Y_{t-p})$ equals $f(Y_t \mid Y_{t-1})$.

Log likelihood of AR(1)

If innovations are Gaussian, we have

$$Y_t \mid Y_{t-1} = y_{t-1} \sim \mathsf{Gauss}\{\mu(1-\phi) + \phi y_{t-1}, \sigma^2\}, \qquad t >$$

so the log-likelihood is

$$\ell(\mu, \phi, \sigma^2) = -\frac{n}{2}\log(2\pi) - n\log\sigma + \frac{1}{2}\log(1-\phi^2)$$
$$- \frac{(1-\phi^2)(y_1-\mu)^2}{2\sigma^2} - \sum_{i=2}^{n} \frac{(y_t - \mu(1-\phi) - \phi y_{t-1})^2}{2\sigma^2}$$

Moments

By the laws of iterated expectation and iterative variance,

$$
\begin{aligned}
\mathsf{E}(Y) &= \mathsf{E}_{\Lambda}\{\mathsf{E}(Y \mid \Lambda\}\\
&= \mathsf{E}(\Lambda) = \mu\\
\mathsf{Va}(Y) &= \mathsf{E}_{\Lambda}\{\mathsf{Va}(Y \mid \Lambda)\} + \mathsf{Va}_{\Lambda}\{\mathsf{E}(Y \mid \Lambda)\}\\
&= \mathsf{E}(\Lambda) + \mathsf{Va}(\Lambda)\\
&= \mu + \mu/k.
\end{aligned}
$$

The marginal distribution of Y, unconditionally, has a variance which exceeds its mean.

Monte Carlo methods

Suppose we can simulate B i.i.d. variables with the same distribution, $X_b \sim F\,(b = 1, \ldots, B)$.

We want to compute $\mathsf{E}\{g(X)\} = \mu_g$ for some functional $g(\cdot)$

- $g(x) = x$ (posterior mean)
- $g(x) = \mathsf{I}(x \in A)$ (probability of event)
- etc.

Monte Carlo methods

We substitute expected value by sample average

$$\widehat{\mu}_g = \frac{1}{B}\sum_{b=1}^{B} g(X_b), \qquad X_b \sim F$$

- law of large number guarantees convergence of $\widehat{\mu}_g \to \mu_g$ if the latter is finite.
- Under finite second moments, central limit theorem gives

$$\sqrt{B}(\widehat{\mu}_g - \mu_g) \sim \mathsf{No}(0, \sigma_g^2).$$

Ordinary Monte Carlo

We want to have an estimator as precise as possible.

- but we can't control the variance of $g(X)$, say σ_g^2
- the more simulations B, the lower the variance of the mean.
- sample average for i.i.d. data has variance σ_g^2/B
- to reduce the standard deviation by a factor 10, we need 100 times more draws!

Remember: the answer is **random**.

Example: functionals of gamma distribution

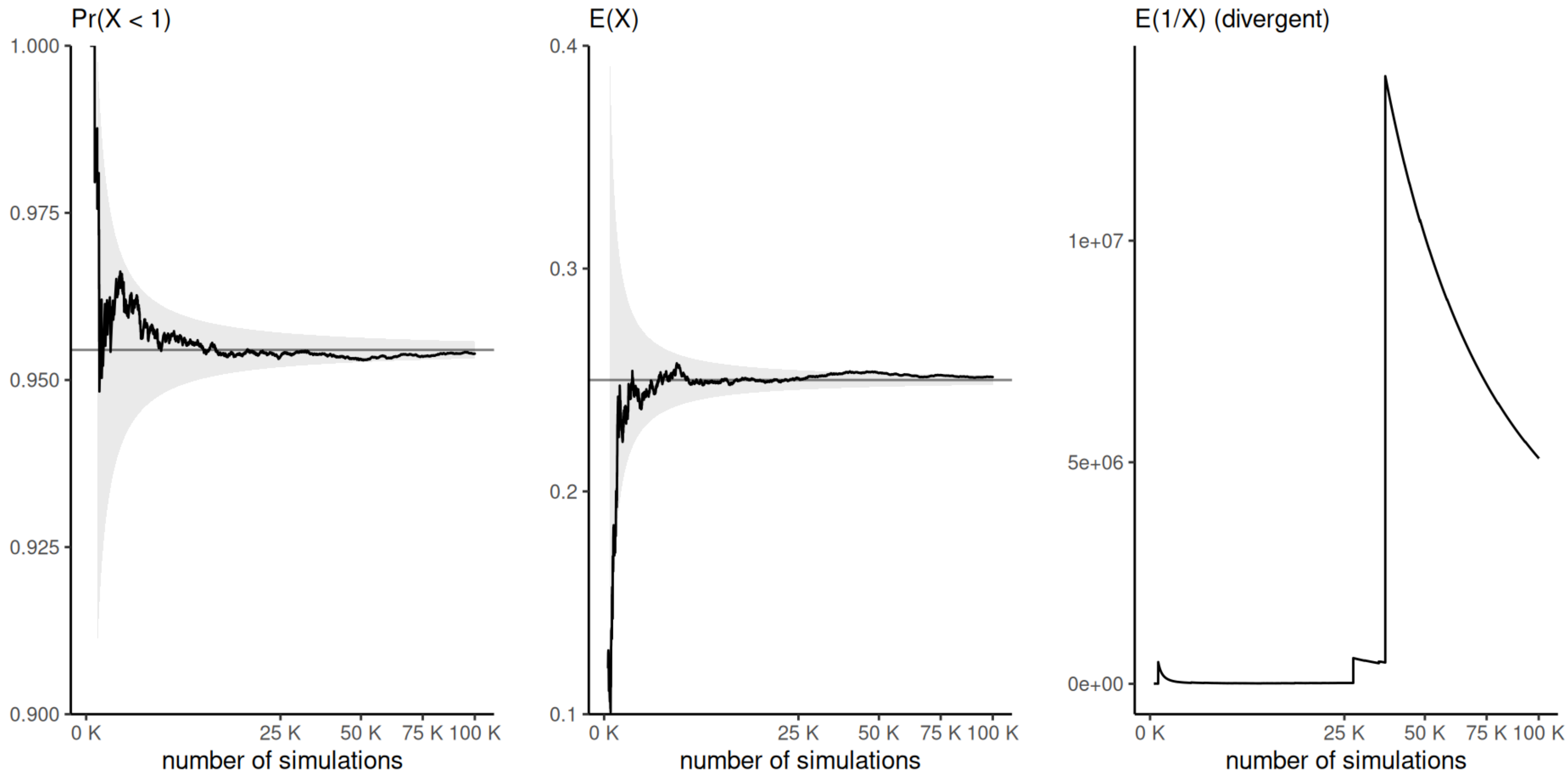

Figure 1: Running mean trace plots for $g(x) = \mathrm{I}(x < 1)$ (left), $g(x) = x$ (middle) and $g(x) = 1/x$ (right) for a Gamma distribution with shape 0.5 and rate 2, as a function of the Monte Carlo sample size.

Simulation algorithms: inversion method

If F is an absolutely continuous distribution function, then

$$F(X) \sim \mathsf{U}(0, 1).$$

The inversion method consists in applying the quantile function F^{-1} to $U \sim \mathsf{U}(0, 1)$, viz.

$$F^{-1}(U) \sim X.$$

Inversion method for truncated distributions

Consider a random variable Y with distribution function F.

If X follows the same distribution as Y, but restricted over the interval $[a, b]$, then

$$\Pr(X \leq x) = \frac{F(x) - F(a)}{F(b) - F(a)}, \qquad a \leq x \leq b,$$

Therefore,

$$F^{-1}[F(a) + \{F(b) - F(a)\}U] \sim X$$

Simulation algorithms: accept-reject

- **Target**: sample from density $p(x)$ (hard to sample from)
- **Proposal**: find a density $q(x)$ with nested support, $\mathrm{supp}(p) \subseteq \mathrm{supp}(q)$, such that

$$\frac{p(x)}{q(x)} \leq C, \quad C \geq 1.$$

Rejection sampling algorithm

1. Generate X from proposal with density $q(x)$.
2. Compute the ratio $R \leftarrow p(X)/q(X)$.
3. If $CU \leq R$ for $U \sim \mathsf{U}(0,1)$, return X, else go back to step 1.

Remarks on rejection sampling

- Acceptance rate is $1/C$
  - we need on average C draws from q to get one from p
- q must be more heavy-tailed than p
  - e.g., $q(x)$ Student-t for $p(x)$ Gaussian
- q should be cheap and easy to sample from!

Designing a good proposal density

Good choices must satisfy the following constraints:

- pick a family $q(x)$ so that

$$C = \sup_x \frac{p(x)}{q(x)}$$

  is as close to 1 as possible.

- you can use numerical optimization with $f(x) = \log p(x) - \log q(x)$ to find the mode $x^\star$ and the upper bound $C = \exp f(x^\star)$.

Accept-reject illustration

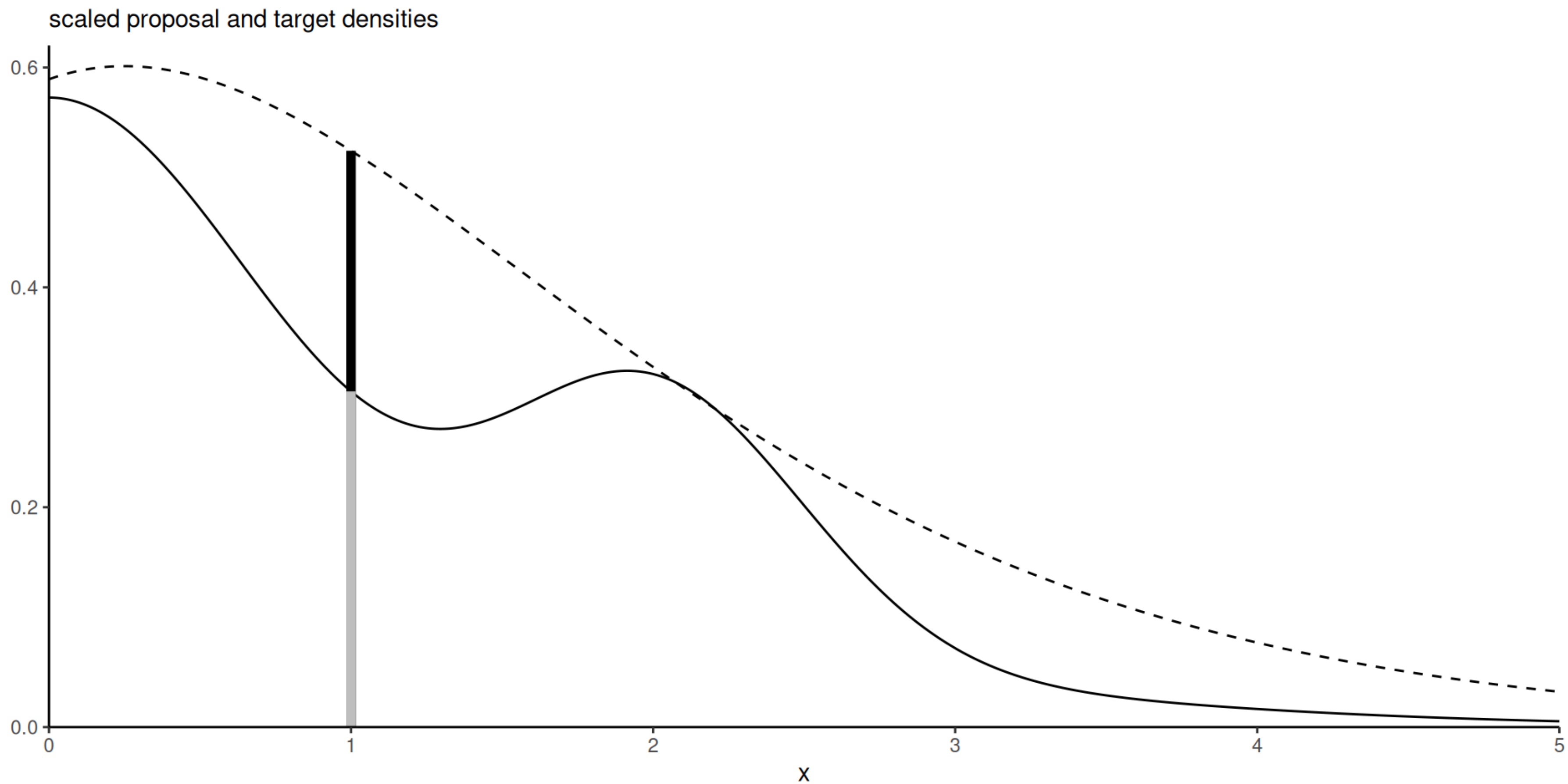

Figure 2: Target density (full) and scaled proposal density (dashed): the vertical segment at $x = 1$ shows the percentage of acceptance for a uniform slice under the scaled proposal, giving an acceptance ratio of 0.58.

Truncated Gaussian via accept-reject

Consider sampling $Y \sim \mathsf{No}(\mu, \sigma^2)$, but truncated in the interval (a, b). The target density is

$$p(x; \mu, \sigma, a, b) = \frac{1}{\sigma}\frac{\phi\left(\frac{x-\mu}{\sigma}\right)}{\Phi(\beta) - \Phi(\alpha)}.$$

for $\alpha = (a-\mu)/\sigma$ and $\beta = (b-\mu)/\sigma$. where $\phi(\cdot)$, $\Phi(\cdot)$ are respectively the density and distribution function of the standard Gaussian distribution.

Accept-reject (crude version)

1. Simulate $X \sim \mathsf{No}(\mu, \sigma^2)$
2. reject any draw if $X < a$ or $X > b$.

The acceptance rate is $C^{-1} = \{\Phi(\beta) - \Phi(\alpha)\}$

```r
# Standard Gaussian truncated on [0,1]
candidate <- rnorm(1e5)
trunc_samp <- candidate[candidate >= 0 & candidate <= 1]
# Acceptance rate
length(trunc_samp)/1e5
```

```
[1] 0.34242
```

```r
# Theoretical acceptance rate
pnorm(1)-pnorm(0)
```

```
[1] 0.3413447
```

Accept-reject for truncated Gaussian

Since the Gaussian is a location scale family, the inversion method gives

$$X \sim \mu + \sigma\Phi^{-1}\left[\Phi(\alpha) + \{\Phi(\beta) - \Phi(\alpha)\}U\right]$$

We however need to evaluate Φ numerically (no closed-form expression).

The method fails for *rare event* simulation because the computer returns

- $\Phi(x) = 0$ for $x \leq -39$
- $\Phi(x) = 1$ for $x \geq 8.3$,

implying that $a \leq 8.3$ for this approach to work (Botev & L'Écuyer, 2017).

Simulating tails of Gaussian variables

We consider simulation from a standard Gaussian truncated above $a > 0$

Write the density of the truncated Gaussian as (Devroye, 1986, p. 381)

$$f(x) = \frac{\exp(-x^2/2)}{\int_a^{\infty} \exp(-z^2/2)\mathrm{d}z} = \frac{\exp(-x^2/2)}{c_1}.$$

Note that, for $x \geq a$,

$$c_1 f(x) \leq \frac{x}{a}\exp\left(-\frac{x^2}{2}\right) = a^{-1}\exp\left(-\frac{a^2}{2}\right)g(x);$$

where $g(x)$ is the density of a Rayleigh variable shifted by a.[1]

1. The constant $C = \exp(-a^2/2)(c_1 a)^{-1}$ approaches 1 quickly as $a \to \infty$ (asymptotically optimality).

Accept-reject: truncated Gaussian with Rayleigh

The shifted Rayleigh has distribution function

$$G(x) = 1 - \exp\{(a^2 - x^2)/2\}, x \geq a.$$

Marsaglia algorithm

1. Generate a shifted Rayleigh above a, $X \leftarrow \{a^2 - 2\log(U)\}^{1/2}$ for $U \sim \mathsf{U}(0,1)$
2. Accept X if $XV \leq a$, where $V \sim \mathsf{U}(0,1)$.

For sampling on $[a,b]$, propose from a Rayleigh truncated above at b (Botev & L'Écuyer, 2017).

```r
a <- 8.3
niter <- 1000L
X <- sqrt(a^2 + 2*rexp(niter))
samp <- X[runif(niter)*X <= a]
```

References

Botev, Z., & L'Écuyer, P. (2017). Simulation from the normal distribution truncated to an interval in the tail. *Proceedings of the 10th EAI International Conference on Performance Evaluation Methodologies and Tools on 10th EAI International Conference on Performance Evaluation Methodologies and Tools*, 23–29. https://doi.org/10.4108/eai.25-10-2016.2266879

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer. http://www.nrbook.com/devroye/