Priors and Approximations in Bayesian Inference

Václav Šmídl

February 28, 2022

Recapitulation

1. independent observations with probability $p(y_1)$ and $p(y_2)$. What is the joint?
2. What is marginal distribution $p(x_1)$ of multivariate Gaussian

$$p(x_1, x_2) = \mathcal{N}\left(\begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix}, \begin{bmatrix}\sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22}\end{bmatrix}\right),$$

3. What $\propto$ means?

Homework: Joint likelihood

$$p(m, s|y_1, y_2, \mu, \alpha, \beta, \tau)$$

$$\propto \frac{1}{s}\frac{1}{s^{\alpha_0+1}} \exp\left(-\frac{1}{2}\frac{(m-y_1)^2}{s} - \frac{1}{2}\frac{(m-y_2)^2}{s} - \frac{1}{2}\frac{(m-\mu)^2}{\tau} - \frac{\beta_0}{s}\right)$$

$$p(m|s, y_1, y_2, \mu, \alpha, \beta, \tau)$$

$$\propto \exp\left(-\frac{1}{2}\frac{(m-y_1)^2}{s} - \frac{1}{2}\frac{(m-y_2)^2}{s} - \frac{1}{2}\frac{(m-\mu)^2}{\tau}\right)$$

$$\propto \exp\left(-\frac{1}{2}\left[m^2(\frac{1}{\tau}+\frac{2}{s}) - 2m(\frac{\mu}{\tau}+\frac{y_1+y_2}{s})\right]\right)$$

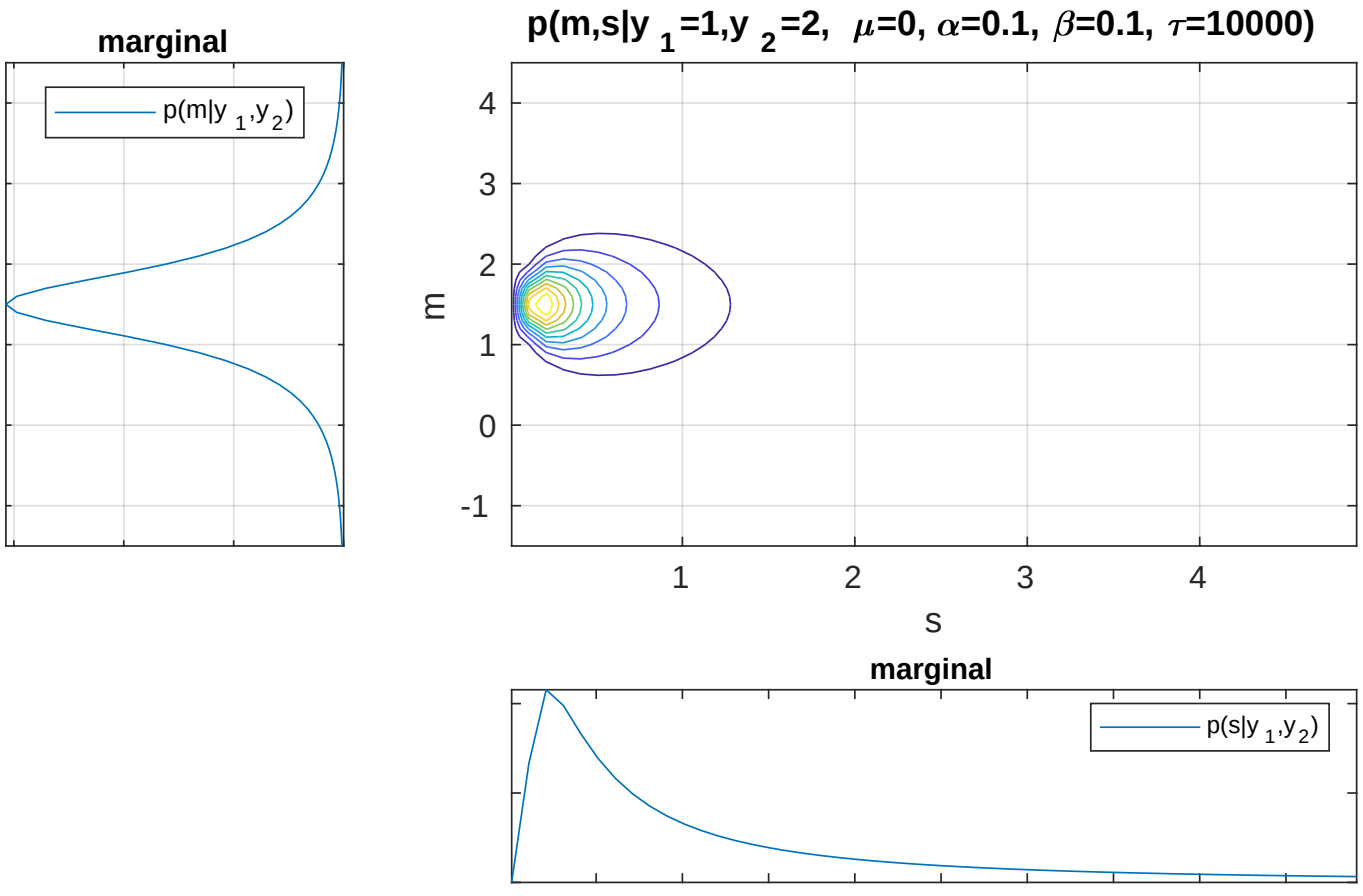
$$= \mathcal{N}\left(m; \left(\frac{1}{\tau}+\frac{2}{s}\right)^{-1}\left(\frac{\mu}{\tau}+\frac{y_1+y_2}{s}\right), \left(\frac{1}{\tau}+\frac{2}{s}\right)^{-1}\right)$$

$$p(s|m, y_1, y_2, \mu, \alpha, \beta, \tau)$$

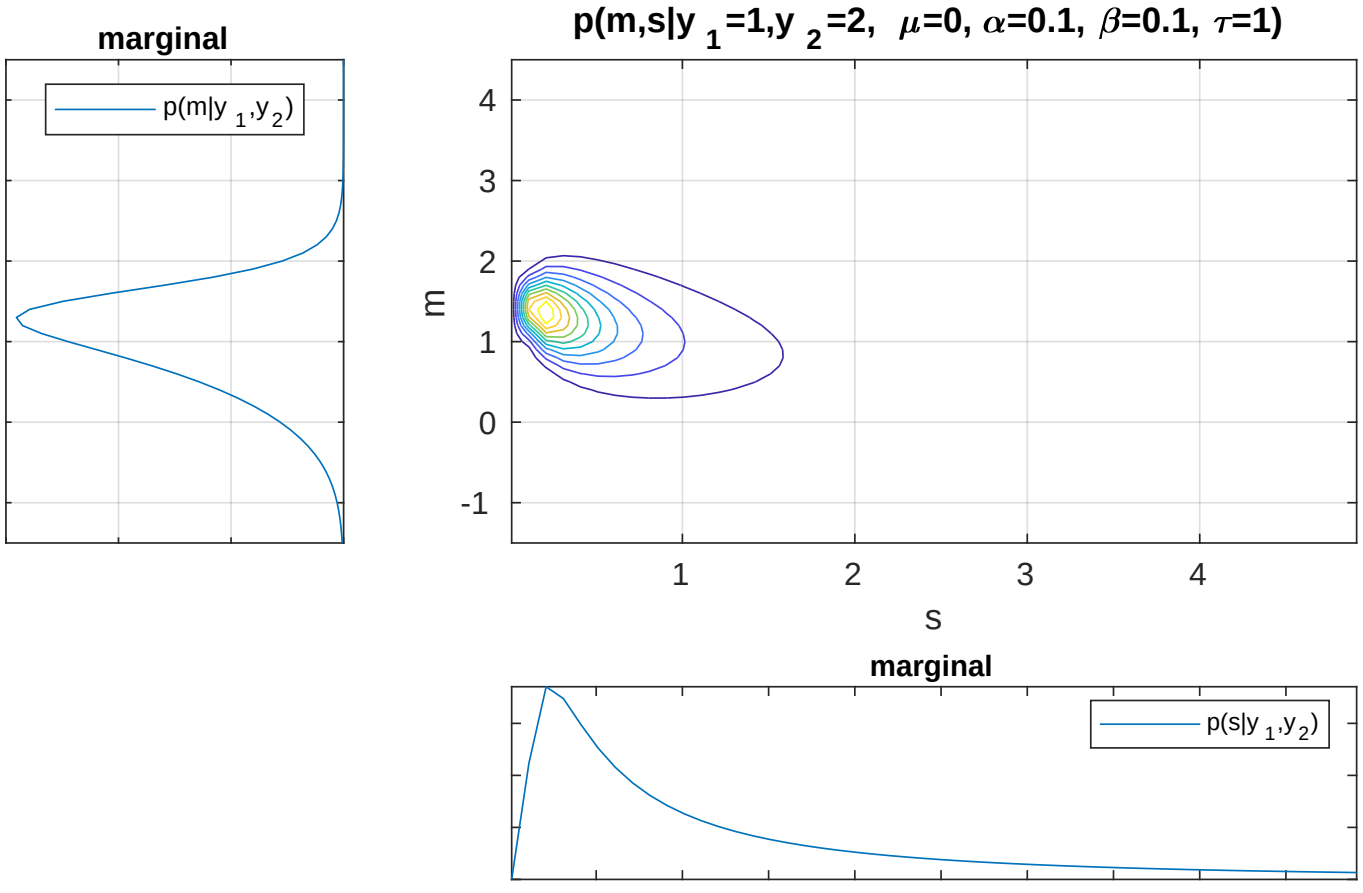
$$\propto \frac{1}{s^{\alpha_0+2}} \exp\left(-\frac{1}{2}\frac{(m-y_1)^2}{s} - \frac{1}{2}\frac{(m-y_2)^2}{s} - \frac{\beta_0}{s}\right)$$

$$= i\mathcal{G}(\alpha_0 + 1, 0.5(m-y_1)^2 + 0.5(m-y_2)^2 + \beta_0)$$

Numerical solution:



Numerical solution:



Prior meaning

$$\begin{aligned} p(s) &= iG(\alpha_0, \beta_0) \\ p(m|s) &= \mathcal{N}(0, s) \\ p(y_i|m, s) &= \mathcal{N}(m, s)\, i = 1, 2 \end{aligned}$$

$$\begin{aligned} p(s) &= iG(\alpha_0, \beta_0) \\ p(m|\tau) &= \mathcal{N}(0, \tau) \\ p(y_i|m, s) &= \mathcal{N}(m, s),\ i = 1, 2 \end{aligned}$$

Prior meaning

$$\begin{aligned} p(s) &= iG(\alpha_0, \beta_0) \\ p(m|s) &= \mathcal{N}(0, s) \\ p(y_i|m, s) &= \mathcal{N}(m, s)\, i = 1, 2 \end{aligned} \qquad \begin{aligned} p(s) &= iG(\alpha_0, \beta_0) \\ p(m|\tau) &= \mathcal{N}(0, \tau) \\ p(y_i|m, s) &= \mathcal{N}(m, s),\ i = 1, 2 \end{aligned}$$

Prior knowledge:

- zero mean m
- $m = 0$ has the same variance as observations
- **analytically solvable**

Define prior that is: i) analytically solvable, ii) can control tightness arbitrarily:

Prior meaning

$$\begin{aligned} p(s) &= iG(\alpha_0, \beta_0) \\ p(m|s) &= \mathcal{N}(0, s) \\ p(y_i|m, s) &= \mathcal{N}(m, s)\, i = 1, 2 \end{aligned}$$

$$\begin{aligned} p(s) &= iG(\alpha_0, \beta_0) \\ p(m|\tau) &= \mathcal{N}(0, \tau) \\ p(y_i|m, s) &= \mathcal{N}(m, s),\ i = 1, 2 \end{aligned}$$

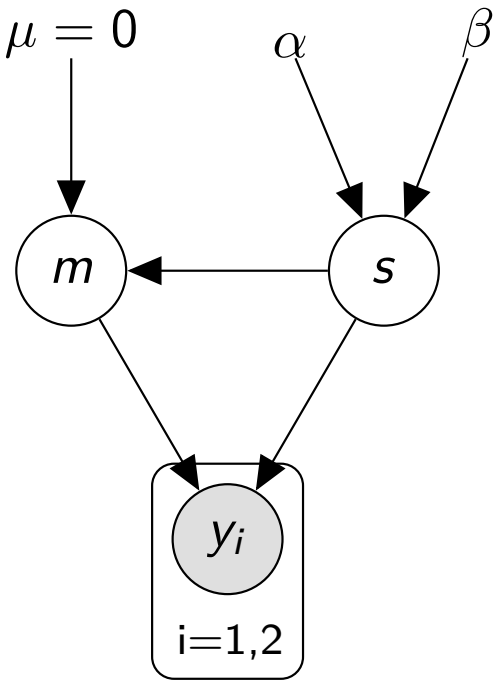
Prior knowledge:

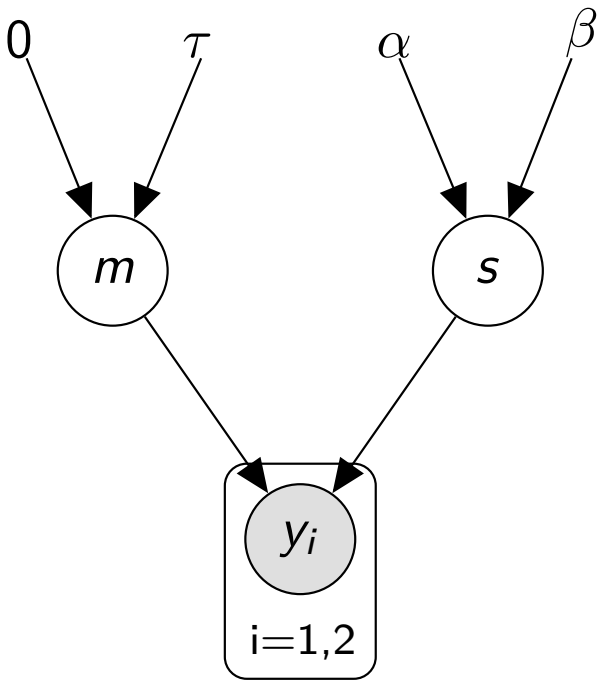
- zero mean m
- variance is independent of m
- **tightness to zero controlled by τ**

Define prior that is: i) analytically solvable, ii) can control tightness arbitrarily:

$$p(m|s, \tau) = \mathcal{N}(0, s\tau)$$

Previous models





Circle = probability distribution,
Gray = observation,
arrow = conditional dependence.

What is the role of prior?

1. Uninformative
   - needed to make a consistent answer
   - Jeffrey's

What is the role of prior?

1. Uninformative
   - needed to make a consistent answer
   - Jeffrey's
2. Non-committal
   - make only minor adjustment (numerical stability)

What is the role of prior?

1. Uninformative
   - needed to make a consistent answer
   - Jeffrey's
2. Non-committal
   - make only minor adjustment (numerical stability)
3. Informative
   - incorporate important information (range)
   - do we know exact shape of the distribution?
4. Structural, weakly informative

Non-informative (Jeffreys)

- The answer should be invariant to the change of coordinates of parameter θ
- Solution

$$p(\theta) \propto \sqrt{\det \mathcal{F}(\theta)},$$

where $\mathcal{F}$ is the Fisher information matrix

$$\mathcal{F}(\theta) = -\mathrm{E}\left[\frac{\partial^2}{\partial \theta^2} \log p(Y;\theta) \middle| \theta \right]$$

- For normal likelihood

$$\begin{aligned} \log p(y|m,s) &= -\frac{1}{2}\log s - \frac{1}{2}\frac{(m-y)^2}{s} + c \\ &= \dots \\ p(m,s) &\propto \frac{1}{s} \end{aligned}$$

Non-informative (Jeffreys)

- The answer should be invariant to the change of coordinates of parameter θ
- Solution

$$p(\theta) \propto \sqrt{\det \mathcal{F}(\theta)},$$

where $\mathcal{F}$ is the Fisher information matrix

$$\mathcal{F}(\theta) = -\mathrm{E}\left[\frac{\partial^2}{\partial \theta^2} \log p(Y;\theta) \middle| \theta \right]$$

- For normal likelihood

$$\begin{aligned} \log p(y|m,s) &= -\frac{1}{2}\log s - \frac{1}{2}\frac{(m-y)^2}{s} + c \\ &= \dots \\ p(m,s) &\propto \frac{1}{s} = i\mathcal{G}(0,0)\mathcal{N}(0,\infty) \end{aligned}$$

- Uniform prior on scale is informative...

Conjugate prior

For normal likelihood with parameters m, s

$$\begin{aligned}
\log p(y|m,s) &= -\frac{1}{2}\log s - \frac{1}{2}\frac{(m-y)^2}{s} + c \\
&= -\frac{1}{2}\log s - \frac{1}{2}\frac{m^2 - 2my + y^2}{s} + c \\
&= [-\frac{1}{2}, -\frac{1}{2}, y, -\frac{1}{2}y^2][\log s, \frac{m^2}{s}, \frac{m}{s}, \frac{1}{s}]
\end{aligned}$$

is a composition of bases functions $[\log s, \frac{m^2}{s}, \frac{m}{s}, \frac{1}{s}]$.

Conjugate prior

For normal likelihood with parameters m, s

$$\begin{aligned}
\log p(y|m,s) &= -\frac{1}{2}\log s - \frac{1}{2}\frac{(m-y)^2}{s} + c \\
&= -\frac{1}{2}\log s - \frac{1}{2}\frac{m^2 - 2my + y^2}{s} + c \\
&= [-\frac{1}{2}, -\frac{1}{2}, y, -\frac{1}{2}y^2][\log s, \frac{m^2}{s}, \frac{m}{s}, \frac{1}{s}]
\end{aligned}$$

is a composition of bases functions $[\log s, \frac{m^2}{s}, \frac{m}{s}, \frac{1}{s}]$. It is advantageous to choose prior with the same basis functions.

$$\begin{aligned}
\log p(m,s) &= -\frac{1}{2}\log s - \frac{1}{2}\frac{(m-\mu)^2}{s} + c - (\alpha_0 + 1)\log s - \frac{\beta_0}{s} \\
&=
\end{aligned}$$

Conjugate prior

For normal likelihood with parameters m, s

$$\begin{aligned}
\log p(y|m,s) &= -\frac{1}{2}\log s - \frac{1}{2}\frac{(m-y)^2}{s} + c \\
&= -\frac{1}{2}\log s - \frac{1}{2}\frac{m^2 - 2my + y^2}{s} + c \\
&= [-\frac{1}{2}, -\frac{1}{2}, y, -\frac{1}{2}y^2][\log s, \frac{m^2}{s}, \frac{m}{s}, \frac{1}{s}]
\end{aligned}$$

is a composition of bases functions $[\log s, \frac{m^2}{s}, \frac{m}{s}, \frac{1}{s}]$. It is advantageous to choose prior with the same basis functions.

$$\begin{aligned}
\log p(m,s) &= -\frac{1}{2}\log s - \frac{1}{2}\frac{(m-\mu)^2}{s} + c - (\alpha_0 + 1)\log s - \frac{\beta_0}{s} \\
&= \\
&= [-\frac{1}{2} - \alpha_0 - 1, -\frac{1}{2}, \mu, -\frac{1}{2}\mu^2 - \beta_0][\log s, \frac{m^2}{s}, \frac{m}{s}, \frac{1}{s}]
\end{aligned}$$

Posterior is of the same form as prior.

Exponential family:

Likelihood of the data is in form:

$$p(y|\theta) = h(y) \exp\left(\eta(\theta)^\top T(y) - A(\theta)\right)$$

where
$\eta(\theta)$ is natural parameter, sometimes used $\eta = [\eta_1, \eta_2, \ldots]$
$T(y)$ is sufficient statistics,

Use:

$$\begin{aligned} p(y_1|\theta) &= h(y_1) \exp\left(\eta(\theta)^\top T(y_1) - A(\theta)\right), \\ p(y_2|\theta) &= h(y_2) \exp\left(\eta(\theta)^\top T(y_2) - A(\theta)\right), \end{aligned}$$

Exponential family:

Likelihood of the data is in form:

$$p(y|\theta) = h(y) \exp\left(\eta(\theta)^\top T(y) - A(\theta)\right)$$

where
$\eta(\theta)$ is natural parameter, sometimes used $\eta = [\eta_1, \eta_2, \ldots]$
$T(y)$ is sufficient statistics,

Use:

$$\begin{aligned}
p(y_1|\theta) &= h(y_1) \exp\left(\eta(\theta)^\top T(y_1) - A(\theta)\right), \\
p(y_2|\theta) &= h(y_2) \exp\left(\eta(\theta)^\top T(y_2) - A(\theta)\right), \\
p(\theta) &= h_0 \exp\left(\eta(\theta)^\top T_0 - \nu_0 A(\theta)\right) \\
p(y_1, \ldots y_n|\theta) &= \prod_{i=1}^{n} h(y_i) \exp\left(\eta(\theta)^\top \sum_{i=1}^{n} T(y_i) - nA(\theta)\right),
\end{aligned}$$

Exponential family of normal distribution

$$
\begin{aligned}
p(y|m,s) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\log s - \frac{1}{2}\frac{(m-y)^2}{s}\right) \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\log s - \frac{1}{2}\left[\frac{m^2}{s} - 2\frac{m}{s}y + \frac{1}{s}y^2\right]\right) \\
&= \underbrace{\frac{1}{\sqrt{2\pi}}}_{h(y)} \exp\left(\underbrace{\left[\frac{m}{s}, -\frac{1}{2s}\right]}_{\eta} \underbrace{\left[y, y^2\right]^\top}_{T(y)} \underbrace{-\frac{1}{2}\log s - \frac{m^2}{2s}}_{-A(\theta)}\right)
\end{aligned}
$$

Or

$$
p(y|m,s) = \underbrace{\frac{1}{\sqrt{2\pi}}}_{h(y)} \exp\left(\underbrace{\left[\frac{m}{s}, -\frac{1}{2s}, -\frac{m^2}{2s}\right]}_{\eta} \underbrace{\left[y, y^2, 1\right]^\top}_{T(y)} \underbrace{-\frac{1}{2}\log s}_{-A(\theta)}\right)
$$

Non-commital: any prior conjugate statistics with minimum impact of sufficient statistics.

Prior

Prior is typically in the hands of the modeller:

- for sufficient number of data, use Jeffrey's
- Conjugate prior beneficial for analytical tractability
- Care needed for tail behaviour

Different choices of distributions on positive support with mean at 1 and different standard deviation.