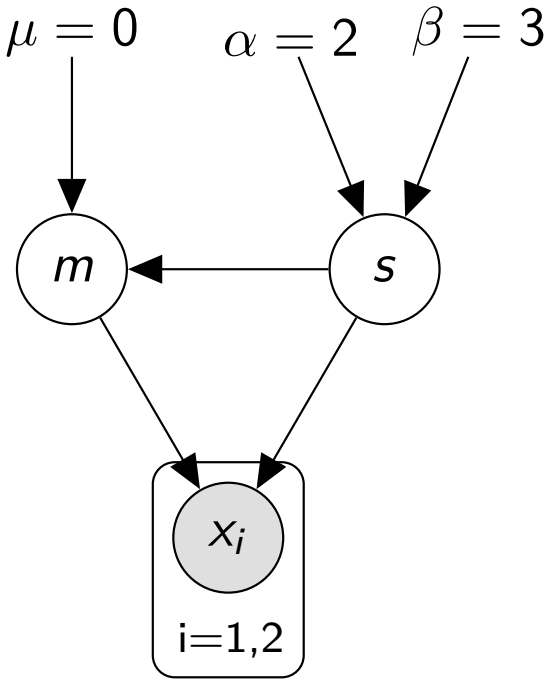
Approximations in Bayesian Inference

Václav Šmídl

March 3, 2020

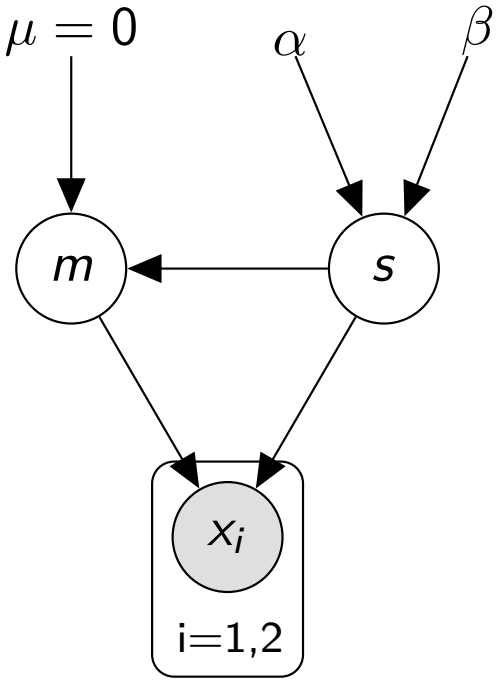
Previous models

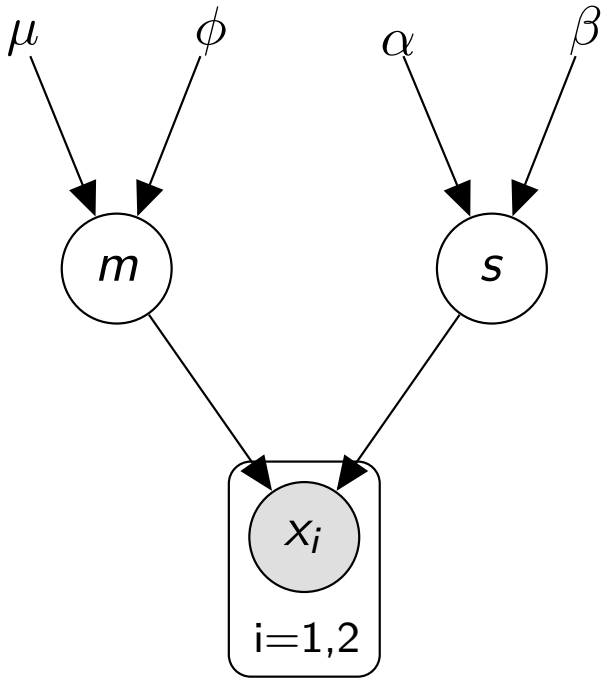


$$
\begin{aligned}
p(s) &= iG(\alpha_0, \beta_0) \\
p(m|s) &= \mathcal{N}(\mu, s) \\
p(x_i|m, s) &= \mathcal{N}(m, s)
\end{aligned}
$$

- Observations x_i are sampled from Gaussian with unknown mean and variance.
- We have some prior information about the mean and variance

Previous models





Inference:

$$p(m, s|x_1, x_2, \mu, \alpha, \beta, \phi) \propto \prod_{i=1}^{2} p(x_i, m, s)p(m|\mu, \phi)p(s|\alpha, \beta),$$

where $p(d)$ is a normalization constant.

Joint likelihood

$$p(m, s|x_1, x_2, \mu, \alpha, \beta, \phi)$$

$$\propto \frac{1}{s}\frac{1}{s^{\alpha_0+1}} \exp\left(-\frac{1}{2}\frac{(m-x_1)^2}{s} - \frac{1}{2}\frac{(m-x_2)^2}{s} - \frac{1}{2}\frac{(m-\mu)^2}{\phi} - \frac{\beta_0}{s}\right)$$

$$p(m|s, x_1, x_2, \mu, \alpha, \beta, \phi)$$

$$\propto \exp\left(-\frac{1}{2}\frac{(m-x_1)^2}{s} - \frac{1}{2}\frac{(m-x_2)^2}{s} - \frac{1}{2}\frac{(m-\mu)^2}{\phi}\right)$$

$$\propto \exp\left(-\frac{1}{2}\left[m^2(\frac{1}{\phi}+\frac{2}{s}) - 2m(\frac{\mu}{\phi}+\frac{x_1+x_2}{s})\right]\right)$$

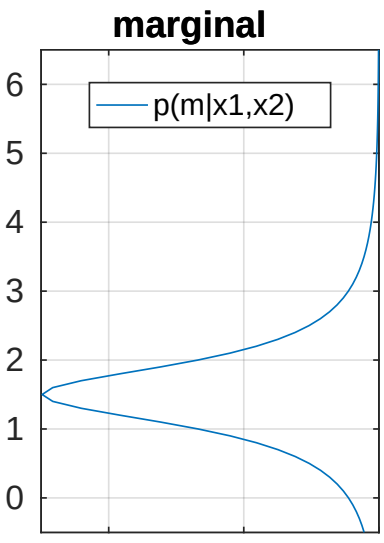
$$= \mathcal{N}\left(m; \left(\frac{1}{\phi}+\frac{2}{s}\right)^{-1}\left(\frac{\mu}{\phi}+\frac{x_1+x_2}{s}\right), \left(\frac{1}{\phi}+\frac{2}{s}\right)^{-1}\right)$$

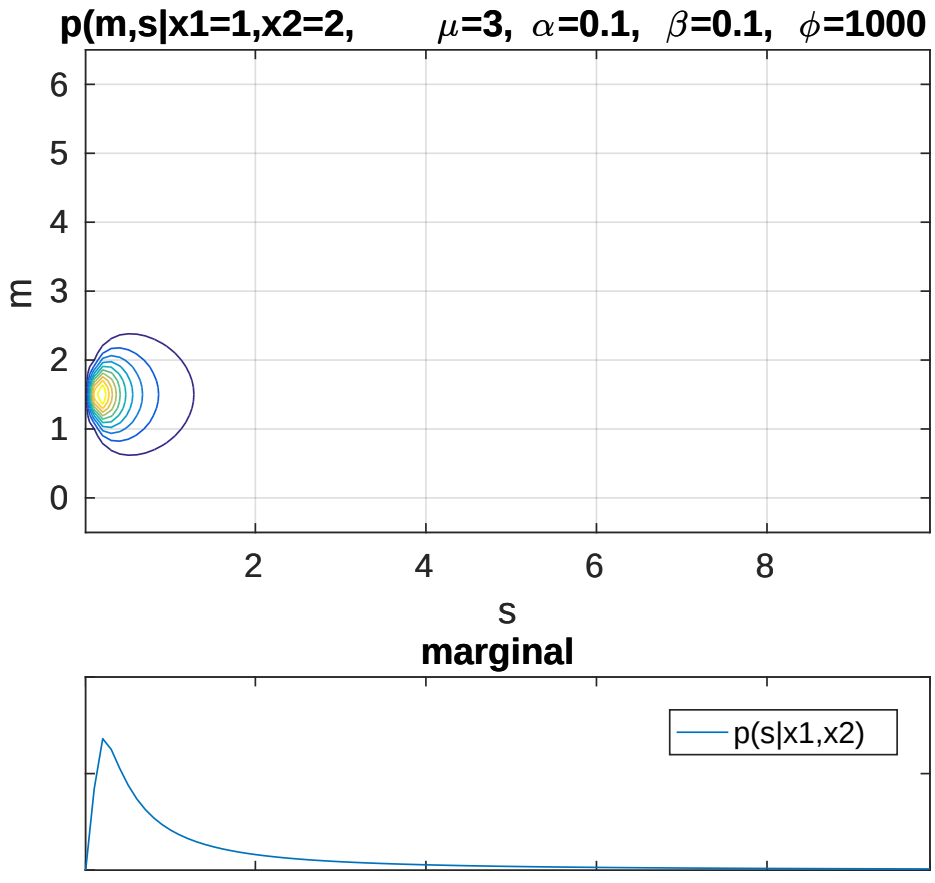
$$p(s|m, x_1, x_2, \mu, \alpha, \beta, \phi)$$

$$\propto \frac{1}{s^{\alpha_0+2}} \exp\left(-\frac{1}{2}\frac{(m-x_1)^2}{s} - \frac{1}{2}\frac{(m-x_2)^2}{s} - \frac{\beta_0}{s}\right)$$

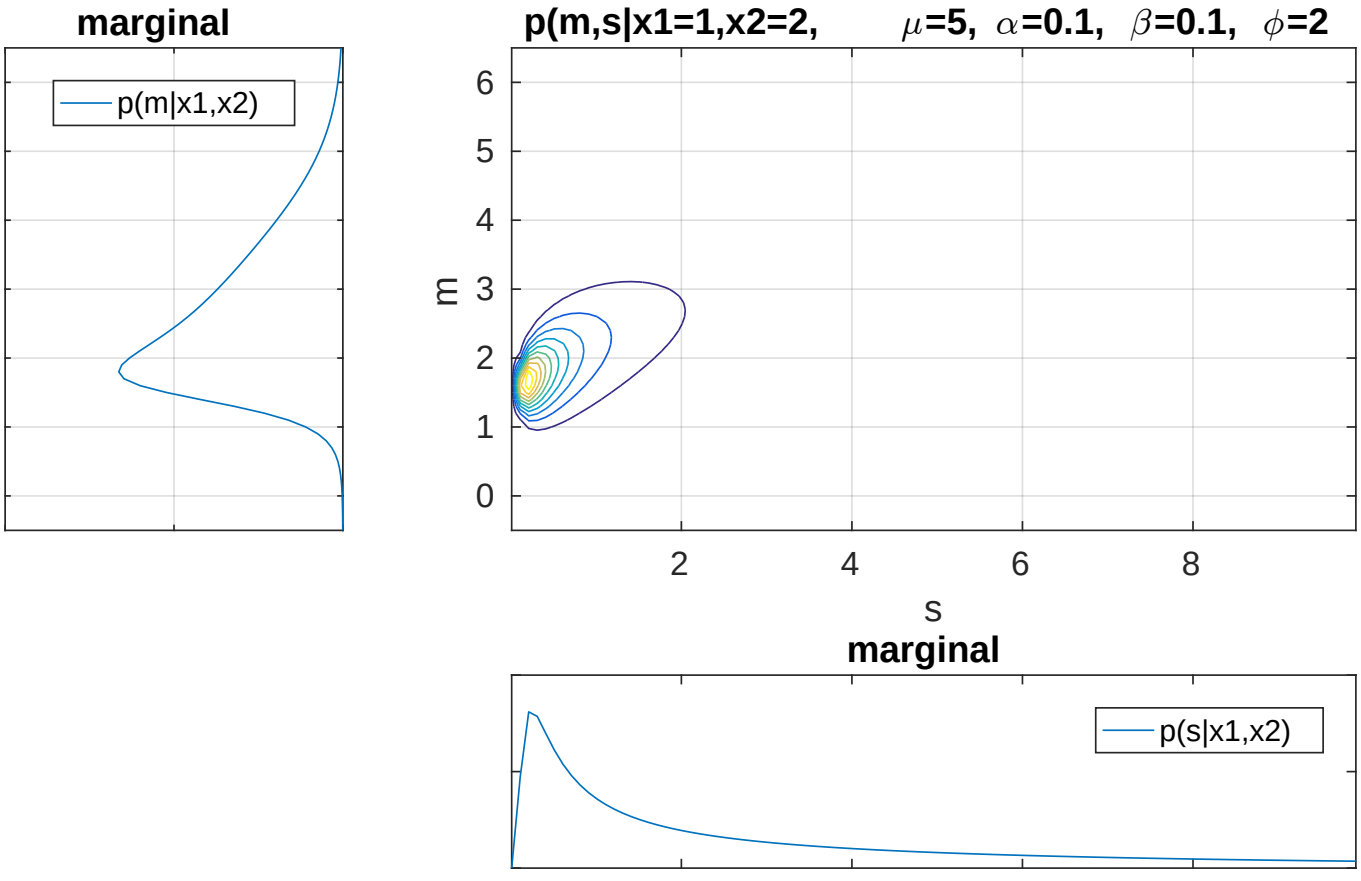
$$= i\mathcal{G}(\alpha_0+1, 0.5(m-x_1)^2 + 0.5(m-x_2)^2 + \beta_0)$$

Numerical solution:





Numerical solution:



What is the role of prior?

1. Uninformative
   - needed to make a consistent answer
   - Jeffrey's
2. Non-committal
   - make minor adjustment
3. Informative
   - incorporate important information
   - 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

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

- For normal likelihood

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

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

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

- For normal likelihood

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

- Uniform prior on scale is informative...

Conjugate prior

For normal likelihood with parameters m, s

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

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

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

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:

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

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

Use:

$$\begin{aligned} p(x_1|\theta) &= h(x_1) \exp\left(\eta(\theta)^\top T(x_1) - A(\theta)\right), \\ p(x_2|\theta) &= h(x_2) \exp\left(\eta(\theta)^\top T(x_2) - A(\theta)\right), \end{aligned}$$

Exponential family:

Likelihood of the data is in form:

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

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

Use:

$$\begin{aligned}
p(x_1|\theta) &= h(x_1) \exp\left(\eta(\theta)^\top T(x_1) - A(\theta)\right), \\
p(x_2|\theta) &= h(x_2) \exp\left(\eta(\theta)^\top T(x_2) - A(\theta)\right), \\
p(\theta) &= h_0 \exp\left(\eta(\theta)^\top T_0 - \nu_0 A(\theta)\right) \\
p(x_1, \ldots x_n|\theta) &= \prod_{i=1}^{n} h(x_i) \exp\left(\eta(\theta)^\top \sum_{i=1}^{n} T(x_i) - nA(\theta)\right),
\end{aligned}$$

Exponential family of normal distribution

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

Or

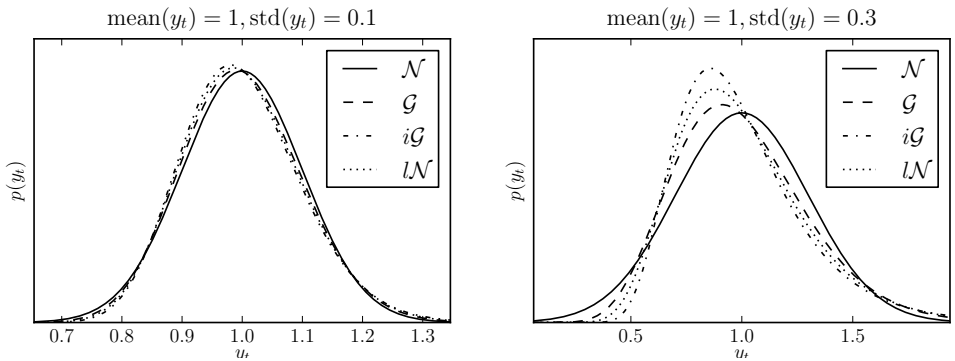
$$
p(x|m,s) = \underbrace{\frac{1}{\sqrt{2\pi}}}_{h(x)} \exp\left(\underbrace{\left[\frac{m}{s}, -\frac{1}{2s}, -\frac{m^2}{2s}\right]}_{\eta} \underbrace{\left[x, x^2, 1\right]^\top}_{T(x)} \underbrace{-\frac{1}{2}\log s}_{-A(\theta)}\right)
$$

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



Working out the non-conjugate prior example

$$p(m, s|x_1, x_2, \mu, \alpha, \beta, \phi)$$
$$\propto \frac{1}{s}\frac{1}{s^{\alpha_0+1}} \exp\left(-\frac{1}{2}\frac{(m-x_1)^2}{s} - \frac{1}{2}\frac{(m-x_2)^2}{s} - \frac{1}{2}\frac{(m-\mu)^2}{\phi} - \frac{\beta_0}{s}\right)$$

Conditional distributions are available!

EM algorithm

- provides maximum of one marginal distribution

Variational Bayes approximation

- provides conditionally independent posterior

Gibbs sampler

- provides samples from posterior (efficient)

EM algorithm classical form

Splits unknown vector into two parts:

1. quantity of interest θ (e.g. mean m)
2. missing data z (e.g. variance s)

General EM algorithm [Dempster, Laird, Rubin, 1977]:

$$\hat{\theta} = \arg\max_{\theta} \int p(x|\theta, z)p(z)dz,$$

can be (approximately) found by alternating:

E-step: $q(\theta|\theta^{(i)}) = \int \log p(x|\theta, z)p(z|\theta^{(i)})dz$

M-step: $\theta^{(i+1)} = \arg\max_{\theta} q(\theta|\theta^{(i)})$

E-step: $\theta = m$, $z = s$

Proxy distribution

$$Q(m|m^{(i)}) = \int \log p(x_1, x_2, m, s) p(s|m^{(i)}) ds$$
$$= \mathrm{E}_{p(s|m^{(i)})}(\log p(x_1, x_2, m, s))$$

using

$$p(x_1, x_2|m, s) \propto \frac{1}{s} \exp\left(-\frac{1}{2}\frac{(m-x_1)^2}{s} - \frac{1}{2}\frac{(m-x_2)^2}{s} - \frac{1}{2}\frac{(m-\mu)^2}{\phi}\right)$$

$$\log p(x_1, x_2|m, s) = c + \log s - \frac{1}{2}\frac{(m-x_1)^2}{s} - \frac{1}{2}\frac{(m-x_2)^2}{s} - \frac{1}{2}\frac{(m-\mu)^2}{\phi},$$

$$Q(m|m^{(i)}) = c - \frac{1}{2}\mathrm{E}(\frac{1}{s})\left[(m-x_1)^2 + (m-x_2)^2\right] - \frac{1}{2}\frac{(m-\mu)^2}{\phi}$$

$$\mathrm{E}(\frac{1}{s}) = \frac{\alpha}{\beta} \qquad \hat{s} = \frac{\beta}{\alpha}$$

$$Q(m|m^{(i)}) = c - \frac{1}{2}\left[\frac{(m-x_1)^2}{\hat{s}} + \frac{(m-x_2)^2}{\hat{s}}\right] - \frac{1}{2}\frac{(m-\mu)^2}{\phi}$$

M-step

New point estimate $m^{(i+1)} = \arg\max_m q(m|m^{(i)})$

1. set first derivative equal to 0
2. completion of squares

$$m^{(i+1)} = \left(\frac{1}{\phi} + \frac{2}{\hat{s}}\right)^{-1} \left(\frac{\mu}{\phi} + \frac{x_1 + x_2}{\hat{s}}\right),$$

based on $\hat{s}$ estimate of conditional

$$\begin{aligned} p(s|m^{(i)}) &= i\mathcal{G}(\alpha, \beta) \\ \alpha &= \alpha_0 + 1, \\ \beta &= 0.5(m^{(i)} - x_1)^2 + 0.5(m^{(i)} - x_2)^2 + \beta_0 \end{aligned}$$

repeat for $p(s|m^{(i+1)})$.

Note: meaning of variables can be swapped. E-step over m and M-step on s.

Divergence minimization

We seek best approximation of intractable distribution $p(x)$ in the chosen class of parametric functions, $q(x|\theta)$, such that

$$\theta^* = \arg\min_{\theta} D(p, q),$$

where $D(p, q)$ is a statistical divergence.

Divergence minimization

We seek best approximation of intractable distribution $p(x)$ in the chosen class of parametric functions, $q(x|\theta)$, such that

$$\theta^* = \arg\min_{\theta} D(p, q),$$

where $D(p, q)$ is a statistical divergence.

Different results for different choices of: i) $q(\theta)$, and ii) D.

Divergence minimization

We seek best approximation of intractable distribution $p(x)$ in the chosen class of parametric functions, $q(x|\theta)$, such that

$$\theta^* = \arg\min_{\theta} D(p, q),$$

where $D(p, q)$ is a statistical divergence.

Different results for different choices of: i) $q(\theta)$, and ii) D.

Variational Bayes: i) $q(\theta_1, \theta_2) = q(\theta_1)q(\theta_2)$, and ii) Kullback-Leibler.

Variational Bayes

Is a divergence minimization technique with

$$q^* = \arg\min_q KL(q||p) = \arg\min_q E_q\left(\log\frac{q}{p}\right)$$

$$q(m,s) = q(m|x_1,x_2)q(s|x_1,x_2).$$

which allows free-form optimization.

Variational Bayes

Is a divergence minimization technique with

$$q^* = \arg\min_q KL(q||p) = \arg\min_q E_q\left(\log\frac{q}{p}\right)$$
$$q(m,s) = q(m|x_1,x_2)q(s|x_1,x_2).$$

which allows free-form optimization.
Result:

$$q(m|x_1,x_2) \propto \exp\left(E_{q(s)}\left[\log p(x_1,x_2,m,s)\right]\right)$$
$$q(s|x_1,x_2) \propto \exp\left(E_{q(m)}\left[\log p(x_1,x_2,m,s)\right]\right)$$

which is a set of implicit functions.

- Proportionality above allows to use $p(x_1,x_2,m,s)$ in place of $p(m,s|x_1,x_2)$
- Variational EM algorithm (E-E algorithm).

E-step: m

Proxy distribution

$$q(m) = \mathsf{E}_{q(s)}(\log p(x_1, x_2, m, s))$$

using

$$p(x_1, x_2, m, s) \propto \frac{1}{s}\frac{1}{s^{\alpha_0+1}} \exp\left(-\frac{1}{2}\frac{(m-x_1)^2}{s} - \frac{1}{2}\frac{(m-x_2)^2}{s} - \frac{1}{2}\frac{(m-\mu)^2}{\phi} - \frac{\beta_0}{s}\right)$$

$$\log p(x_1, x_2, m, s) \propto (\alpha_0 + 2)\log s - \frac{1}{2}\frac{(m-x_1)^2}{s} - \frac{1}{2}\frac{(m-x_2)^2}{s} - \frac{1}{2}\frac{(m-\mu)^2}{\phi} - \frac{\beta_0}{s},$$

$$\propto -\frac{1}{2}\mathsf{E}\left(\frac{1}{s}\right)\left[(m-x_1)^2 + (m-x_2)^2\right] - \frac{1}{2}\frac{(m-\mu)^2}{\phi}$$

$$\propto -\frac{1}{2}\left[\frac{(m-x_1)^2}{\hat{s}} + \frac{(m-x_2)^2}{\hat{s}}\right] - \frac{1}{2}\frac{(m-\mu)^2}{\phi}$$

$$q(m) = \mathcal{N}\left(m; \left(\frac{1}{\phi} + \frac{2}{\hat{s}}\right)^{-1}\left(\frac{\mu}{\phi} + \frac{x_1 + x_2}{\hat{s}}\right), \left(\frac{1}{\phi} + \frac{2}{\hat{s}}\right)^{-1}\right)$$

E-step: s

Proxy distribution

$$q(s) = \mathrm{E}_{q(m)}(\log p(x_1, x_2, m, s))$$

using

$$p(x_1, x_2, m, s) \propto \frac{1}{s}\frac{1}{s^{\alpha_0+1}} \exp\left(-\frac{1}{2}\frac{(m-x_1)^2}{s} - \frac{1}{2}\frac{(m-x_2)^2}{s} - \frac{1}{2}\frac{(m-\mu)^2}{\phi} - \frac{\beta_0}{s}\right)$$

$$\log p(x_1, x_2, m, s) = -(\alpha_0+2)\log s - \frac{1}{2}\frac{(m-x_1)^2}{s} - \frac{1}{2}\frac{(m-x_2)^2}{s} - \frac{1}{2}\frac{(m-\mu)^2}{\phi} - \frac{\beta_0}{s},$$

$$\begin{aligned}
\mathrm{E}_{q(m)}(\log p(x_1, \cdot)) &= -(\alpha_0+2)\log s - \frac{1}{2s}\mathrm{E}_{q(m)}\left[(m-x_1)^2 + (m-x_2)^2\right] - \frac{\beta_0}{s} \\
&= -(\alpha_0+2)\log s - \frac{1}{2s}\mathrm{E}_{q(m)}\left[(m^2 - 2mx_1 + x_2^2 + m^2 - 2x_2 m + x_2^2\right] \\
q(s) &= i\mathcal{G}(\alpha, \beta), \\
\alpha &= \alpha_0 + 1, \\
\beta &= 0.5\mathrm{E}(m^2) - \mathrm{E}(m)(x_1 + x_2) + 0.5(x_1^2 + x_2^2) + \beta_0
\end{aligned}$$

Toy: Variational Bayes

Factors:

$$q(m) = \mathcal{N}\left(m; \left(\frac{1}{\phi} + \frac{2}{\hat{s}}\right)^{-1}\left(\frac{\mu}{\phi} + \frac{x_1 + x_2}{\hat{s}}\right), \left(\frac{1}{\phi} + \frac{2}{\hat{s}}\right)^{-1}\right)$$

$$q(s) = i\mathcal{G}\left(\alpha_0 + 1, \mathrm{E}(m^2) - \mathrm{E}(m)(x_1 + x_2) + 0.5(x_1^2 + x_2^2) + \beta_0\right)$$

with

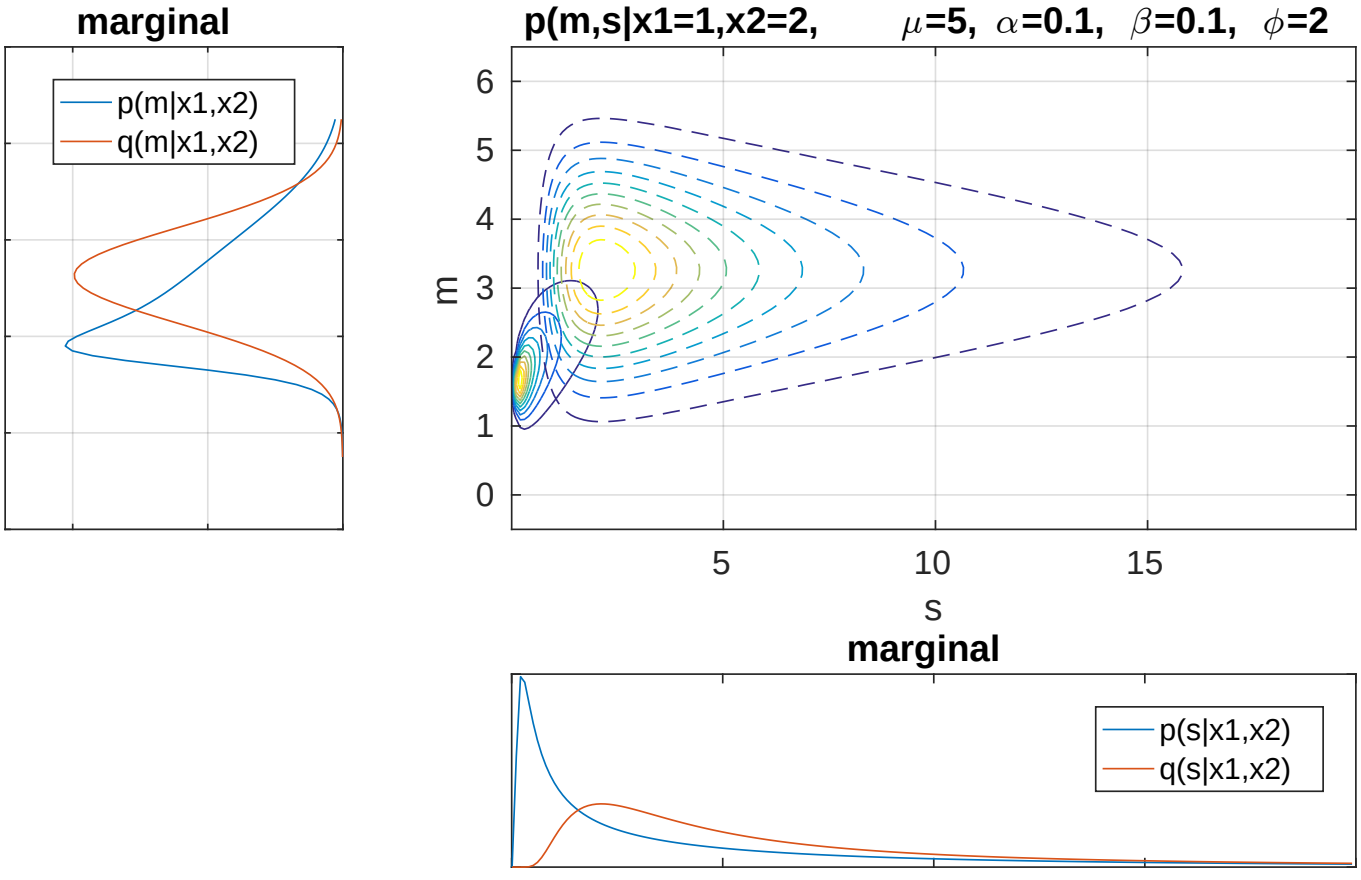
$$\hat{s} = \frac{\mathrm{E}(m^2) - \mathrm{E}(m)(x_1 + x_2) + 0.5(x_1^2 + x_2^2) + \beta_0}{\alpha_0 + 1},$$

$$\mathrm{E}(m) = \left(\frac{1}{\phi} + \frac{2}{\hat{s}}\right)^{-1}\left(\frac{\mu}{\phi} + \frac{x_1 + x_2}{\hat{s}}\right),$$

$$\mathrm{E}(m^2) = \mathrm{E}(m)^2 + \left(\frac{1}{\phi} + \frac{2}{\hat{s}}\right)^{-1},$$

which needs to be (Iterated).

Toy: Variational Bayes Iterations



Homework assignment

- Working code for EM estimation of toy problem:
  1. $\hat{m} = \arg\max p(m|x_1, x_2)$ (5 points)
  2. $\hat{s} = \arg\max p(s|x_1, x_2)$ (5 points)
- Working code for Variational Bayes estimation of the toy problem (10 points).