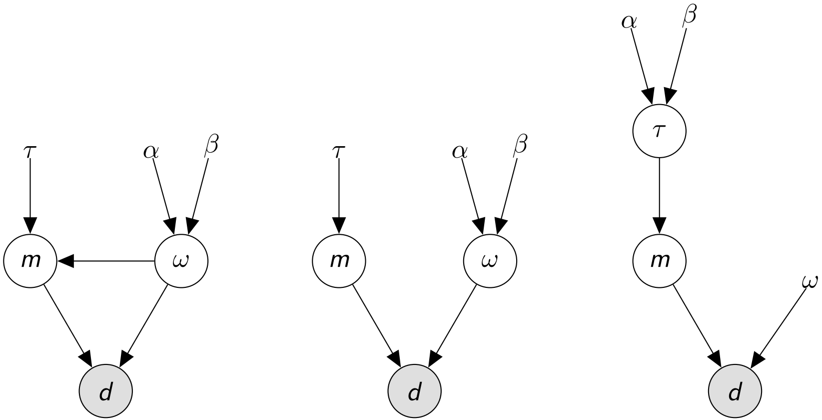
Approximations in Bayesian Inference

Václav Šmídl

February 27, 2018

Previous models



Inference:

$$p(m, \omega | d) = \frac{p(d, m, \omega)}{p(d)},$$

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

Laplace approximation

Consider (intractable) distribution $p(x)$. From which we are able to compute an extreme

$$\hat{x} = \arg\max p(x),$$

Using Taylor expansion of $\log p(x)$ at the extreme:

$$\begin{aligned}\log p(x) &\approx \log p(\hat{x}) + [\nabla \log p(\hat{x})]^T (x - \hat{x}) - \frac{1}{2}(x - \hat{x}) H (x - \hat{x}) \\ H &= -\nabla\nabla \log p(\hat{x})\end{aligned}$$

Yielding:

$$\begin{aligned}p(x) &\approx \mathcal{N}(\hat{x}, \Sigma), \\ \hat{x} &= \arg\max p(x), \\ \Sigma &= (-\nabla\nabla \log p(\hat{x}))^{-1}\end{aligned}$$

Toy problem

Noisy observation:

$$d = m + e,$$

$$p(m, \omega | d) \propto \sqrt{\omega} \exp\left(-\frac{1}{2}\omega(d-m)^2\right)$$
$$\sqrt{\omega\tau} \exp\left(-\frac{1}{2}\tau\omega m^2\right) \omega^{\alpha-1} \exp(-\beta\omega)$$

```
g = gradient(lpmom_d,[om,m]);
H = hessian(lpmom_d,[om,m]);
hat = solve(g,[m,om])
```

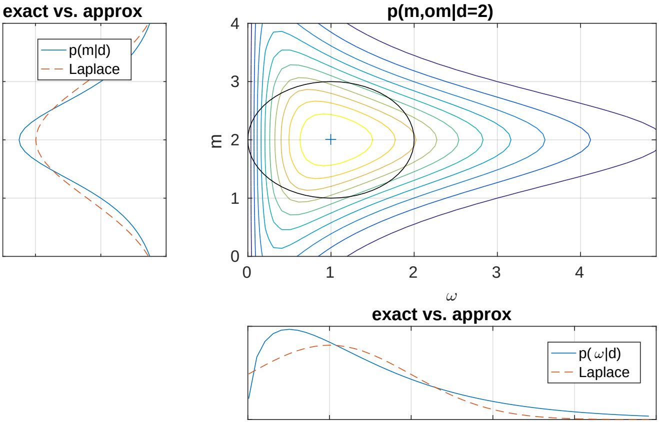
Extreme:

$$\hat{m} = \frac{d}{\tau + 1}$$
$$\hat{\omega} = \frac{2\alpha + 2\alpha\tau}{\tau d^2 + 2\beta + 2\beta\tau}$$

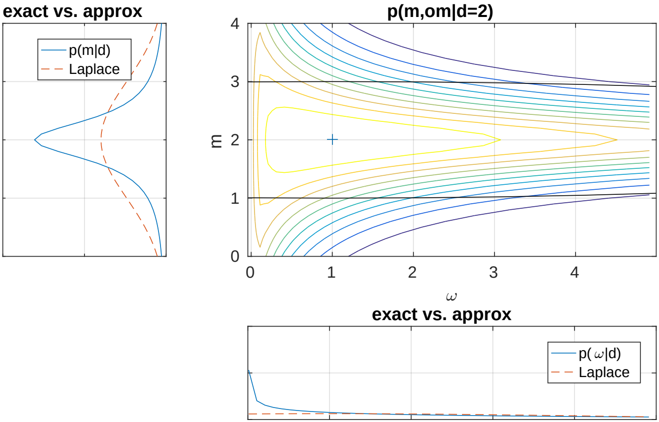
Hamiltonian

$$H = -\begin{bmatrix} \frac{\alpha}{\hat{\omega}^2} & 0 \\ 0 & \hat{\omega}(1+\tau) \end{bmatrix}$$

Toy: $\alpha = 1, \beta = 1, \tau = 0$



Toy: $\alpha = 0.1, \beta = 0.1, \tau = 0$



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.

How to choose: i) $q(x, \theta)$, 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.

How to choose: i) $q(x, \theta)$, and ii) D.

Theory: Choose θ^* that minimizes expected risk (Bernardo, 1979)

$$\theta^* = \arg\min_\theta E_p \left(\log \frac{p}{q} \right)$$

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.

How to choose: i) $q(x, \theta)$, and ii) D.

Theory: Choose θ^* that minimizes expected risk (Bernardo, 1979)

$$\theta^* = \arg\min_\theta E_p \left(\log \frac{p}{q} \right) = \arg\min_\theta KL(p||q).$$

Results to moment-matching.

Variational Bayes

Is a divergence minimization technique with

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

$$q(x_1, x_2) = q(x_1|\theta_1)q(x_2|\theta_2).$$

which allows free-form optimization.

Variational Bayes

Is a divergence minimization technique with

$$\theta^* = \arg\min_\theta KL(q||p) = \arg\min_\theta E_q\left(\log\frac{q}{p}\right)$$
$$q(x_1, x_2) = q(x_1|\theta_1)q(x_2|\theta_2).$$

which allows free-form optimization.
Result:

$$q(x_1|\theta_1) \propto \exp\left(E_{q(x_2)}\left[\log p(x_1, x_2)\right]\right)$$
$$q(x_2|\theta_2) \propto \exp\left(E_{q(x_1)}\left[\log p(x_1, x_2)\right]\right)$$

which is a set of implicit functions.

- Proportionality allows to use $p(x_1, x_2, d)$ in place of $p(x_1, x_2|d)$

Toy: Variational Bayes

General rule: $q(x_1|\theta_1) \propto \exp\left(E_{qx(1)}\left[\log p(x_1, x_2)\right]\right)$
Toy:

$$\log p(m, \omega, d) \propto \frac{1}{2}\log\omega - \frac{1}{2}\omega(d-m)^2$$
$$\frac{1}{2}\log\omega - \frac{1}{2}\tau\omega m^2 + (\alpha-1)\log\omega - \beta\omega$$

Toy: Variational Bayes

General rule: $q(x_1|\theta_1) \propto \exp\left(E_{qx(1)}\left[\log p(x_1, x_2)\right]\right)$
Toy:

$$\log p(m, \omega, d) \propto \frac{1}{2}\log\omega - \frac{1}{2}\omega(d-m)^2$$
$$\frac{1}{2}\log\omega - \frac{1}{2}\tau\omega m^2 + (\alpha - 1)\log\omega - \beta\omega$$

Factor $q(m|d)$:

$$\log q(m|d) \propto E_{q(\omega)}\left[-\frac{1}{2}\omega(d-m)^2 - \frac{1}{2}\tau\omega m^2\right]$$
$$q(m|d) \propto \exp\left(-\frac{1}{2}\langle\omega\rangle(d-m)^2 - \frac{1}{2}\tau\langle\omega\rangle m^2\right)$$

Toy: Variational Bayes

General rule: $q(x_1|\theta_1) \propto \exp\left(E_{qx(1)}\left[\log p(x_1, x_2)\right]\right)$
Toy:

$$\log p(m, \omega, d) \propto \frac{1}{2}\log\omega - \frac{1}{2}\omega(d-m)^2$$
$$\frac{1}{2}\log\omega - \frac{1}{2}\tau\omega m^2 + (\alpha - 1)\log\omega - \beta\omega$$

Factor $q(m|d)$:

$$\log q(m|d) \propto E_{q(\omega)}\left[-\frac{1}{2}\omega(d-m)^2 - \frac{1}{2}\tau\omega m^2\right]$$
$$q(m|d) \propto \exp\left(-\frac{1}{2}\langle\omega\rangle(d-m)^2 - \frac{1}{2}\tau\langle\omega\rangle m^2\right)$$

Factor $q(\omega|d)$:

$$\log q(\omega|d) \propto E_{q(m)}\left[\alpha\log\omega - \frac{1}{2}\omega\left((d-m)^2 + \tau m^2 + \beta\right)\right]$$
$$q(\omega|d) \propto \omega^{\alpha}\exp\left[-\frac{1}{2}\omega\left\langle(d-m)^2 + \tau m^2 + 2\beta\right\rangle\right]$$

Toy: Variational Bayes

Factors:

$$\begin{aligned} q(m|d) &= N(\hat{m}, \sigma_m), \\ q(\omega|d) &= G(\alpha_\omega, \beta_\omega), \end{aligned}$$

with

$$\begin{aligned} \hat{m} &= \frac{d}{1+\tau}, & \hat{\omega} &= \frac{\alpha_\omega}{\beta_\omega}, \\ \sigma_m &= \frac{1}{(1+\tau)\langle\omega\rangle}, & \langle m\rangle &= \hat{m}, \\ \alpha_\omega &= \alpha + 1, & \left\langle m^2\right\rangle &= \hat{m}^2 + \sigma_m \\ \beta_\omega &= \beta + \frac{1}{2}\left(d^2 - 2d\langle m\rangle + (1+\tau)\left\langle m^2\right\rangle\right), \end{aligned}$$

which needs to be solved.

Toy: Variational Bayes Iterations

Factors: with

$$\hat{m} = \frac{d}{1+\tau}, \qquad \hat{\omega} = \frac{\alpha_\omega}{\beta_\omega},$$

$$\sigma_m = \frac{1}{(1+\tau)\langle\omega\rangle}, \qquad \langle m \rangle = \hat{m},$$

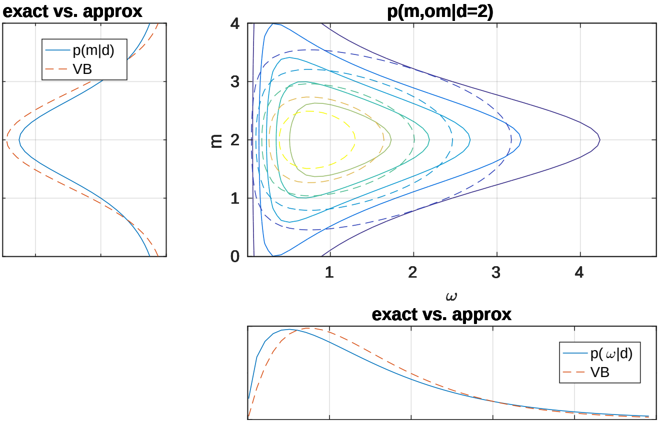
$$\alpha_\omega = \alpha + 1, \qquad \left\langle m^2 \right\rangle = \hat{m}^2 + \sigma_m$$

$$\beta_\omega = \beta + \frac{1}{2}\left(d^2 - 2d\langle m\rangle + (1+\tau)\left\langle m^2\right\rangle\right),$$

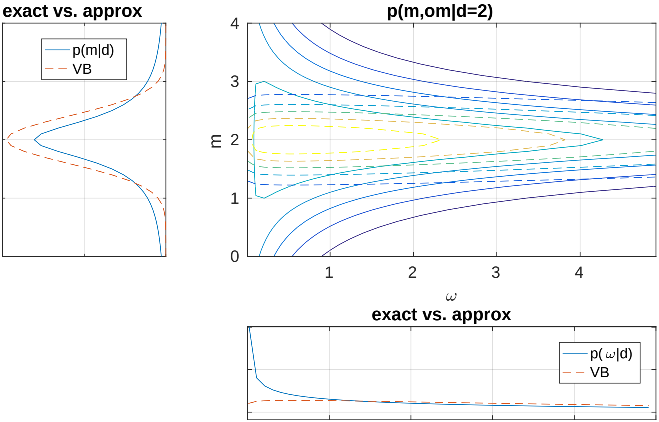
Iterations:

1. evaluate shaping parameters of $q(m)$: $\hat{m}, \sigma_m$
2. evaluate moments $\langle m\rangle, \left\langle m^2\right\rangle$
3. evaluate shaping parameters of $q(\omega)$: $\alpha_\omega, \beta_\omega$
4. evaluate moment $\langle\omega\rangle$

Toy: Variational Bayes Iterations



Toy: Variational Bayes Iterations



Monte Carlo methods

Approximation of a distribution by "Dirac train"

$$p(x) \approx \frac{1}{N} \sum_{i=1}^{N} \delta(x - x^{(i)}).$$

Approximation of moments, cumulative density.

1. Importance sampling,
   1.1 Adaptive importance sapling
   1.2 Population Monte Carlo
2. Monte Carlo Markov Chain
   2.1 Metropolis-Hastings (Gibbs sampler)
   2.2 Hybrid MC (Hamiltonian Monte Carlo)

Convergence assured under mild conditions, different convergence rate.

Importance Sampling

To represent

$$p(x) \approx \frac{1}{N} \sum_{i=1}^{N} \delta(x - x^{(i)}). \tag{1}$$

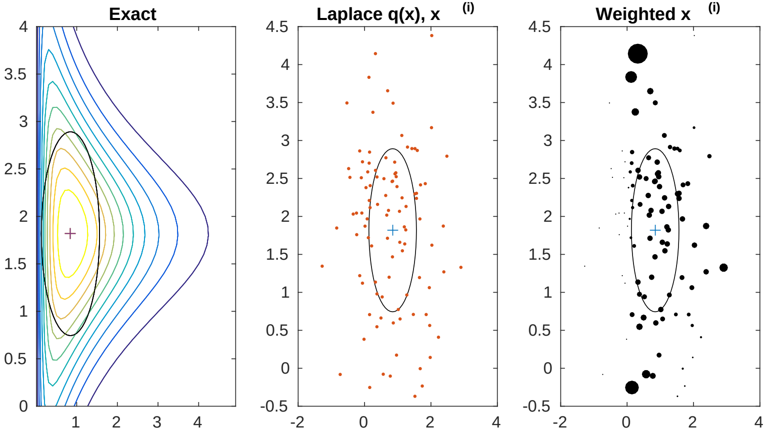
an ideal sampler should sample $x^{(i)} \sim p(x)$, which is not available.
Using

$$p(x) = p(x) \frac{q(x)}{q(x)},$$

we can approximate $q(x)$ by (1) by sampling $x^{(i)} \sim q(x)$.

$$\begin{aligned} p(x) &\propto \frac{p(x)}{q(x)} \frac{1}{N} \sum_{i=1}^{N} \delta(x - x^{(i)}), \\ &\propto \sum_{i=1}^{N} \tilde{w}_i \delta(x - x^{(i)}), & \tilde{w}_i &= \frac{p(x^{(i)})}{q(x^{(i)})} \\ &= \sum_{i=1}^{N} w_i \delta(x - x^{(i)}) & w_i &= \frac{\tilde{w}_i}{\sum_{i=1}^{N} \tilde{w}_i} \end{aligned}$$

Toy: Importance sampling



Adaptive Importance sampling

What if $q(x)$ is too far from $p(x)$?

Adaptive Importance sampling

What if $q(x)$ is too far from $p(x)$? Move it.
Choose parametric form $q(x|\theta)$ and adapt parameter.

Population MC:

- Sample one generation
- compute weights
- estimate parameter
- Sample next generation

Adaptive Importance sampling

What if $q(x)$ is too far from $p(x)$? Move it.
Choose parametric form $q(x|\theta)$ and adapt parameter.

Population MC:

- Sample one generation
- compute weights
- estimate parameter
- Sample next generation

AMIS:

- Consider each generation to be a component in deterministic mixture

$$q(x) = \sum_{g=1}^{G} q_g(x)$$

MCMC: Metropolis Hastings

Instead of fixed distribution, we define a Markov chain that converges to the true distribution.

1. choose transition kernel $q(x|x^{(i)})$,
2. generate sample $x^* \sim q(x|x^{(i)})$,
3. With probability
$$\min\left(1, \frac{p(x^*)q(x^{(i)}|x^*)}{p(x^{(i)})q(x^*|x^{(i)})}\right)$$
accept $(i = i + 1, x^{(i)} = x^*)$, else reject; goto 2.

MCMC: Metropolis Hastings

Instead of fixed distribution, we define a Markov chain that converges to the true distribution.

1. choose transition kernel $q(x|x^{(i)})$,
2. generate sample $x^* \sim q(x|x^{(i)})$,
3. With probability

$$\min\left(1, \frac{p(x^*)q(x^{(i)}|x^*)}{p(x^{(i)})q(x^*|x^{(i)})}\right)$$

   accept $(i = i + 1, x^{(i)} = x^*)$, else reject; goto 2.

How to choose the kernel:

- Random walk (Gaussian), with parameters θ
- Use known properties: conditionals

MCMC: Gibbs sampler

Special case of MH for mutidimensional distributions.

$$p(x_1, x_2, \ldots, x_k)$$

with MH probability of acceptance equal to one.

1. generate sample $x_1^{(i+1)} \sim p(x_1 | x_2^{(i)}, \ldots x_k^{(i)})$,
2. generate sample $x_2^{(i+1)} \sim p(x_2 | x_1^{(i+1)}, \ldots x_k^{(i)})$,

   $\vdots$

3. generate sample $x_k^{(i+1)} \sim p(x_2 | x_1^{(i+1)}, \ldots x_{k-1}^{(i+1)})$,

Suitable when these distributions are tractable.

MCMC: Gibbs sampler

Special case of MH for mutidimensional distributions.

$$p(x_1, x_2, \ldots, x_k)$$

with MH probability of acceptance equal to one.

1. generate sample $x_1^{(i+1)} \sim p(x_1 | x_2^{(i)}, \ldots x_k^{(i)})$,
2. generate sample $x_2^{(i+1)} \sim p(x_2 | x_1^{(i+1)}, \ldots x_k^{(i)})$,

   $\vdots$

3. generate sample $x_k^{(i+1)} \sim p(x_2 | x_1^{(i+1)}, \ldots x_{k-1}^{(i+1)})$,

Suitable when these distributions are tractable.

- not suitable for parallel computing

Toy: Gibbs sampler

Toy:

$$\begin{aligned}\log p(m, \omega, d) &\propto \frac{1}{2}\log \omega - \frac{1}{2}\omega(d-m)^2 \\ &\frac{1}{2}\log \omega - \frac{1}{2}\tau\omega m^2 + (\alpha - 1)\log \omega - \beta\omega\end{aligned}$$

Toy: Gibbs sampler

Toy:

$$
\begin{aligned}
\log p(m, \omega, d) \propto & \frac{1}{2}\log\omega - \frac{1}{2}\omega(d-m)^2 \\
& \frac{1}{2}\log\omega - \frac{1}{2}\tau\omega m^2 + (\alpha - 1)\log\omega - \beta\omega
\end{aligned}
$$

Conditional $p(m|d,\omega)$:

$$
\begin{aligned}
p(m|d,\omega) &\propto \exp\left(-\frac{1}{2}\omega(d-m)^2 - \frac{1}{2}\tau\omega m^2\right) \\
&= \mathcal{N}\left(\frac{d}{1+\tau}, ((1+\tau)\omega)^{-1}\right)
\end{aligned}
$$

Toy: Gibbs sampler

Toy:

$$
\begin{aligned}
\log p(m, \omega, d) \propto \frac{1}{2}\log\omega - \frac{1}{2}\omega(d-m)^2 \\
\frac{1}{2}\log\omega - \frac{1}{2}\tau\omega m^2 + (\alpha - 1)\log\omega - \beta\omega
\end{aligned}
$$

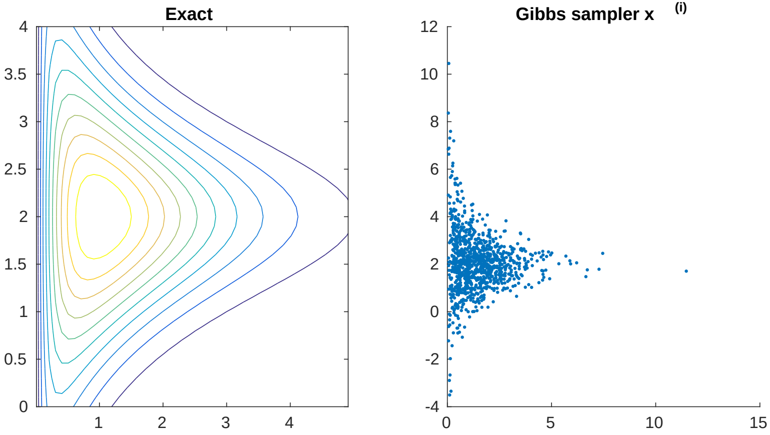
Conditional $p(m|d,\omega)$:

$$
\begin{aligned}
p(m|d,\omega) &\propto \exp\left(-\frac{1}{2}\omega(d-m)^2 - \frac{1}{2}\tau\omega m^2\right) \\
&= \mathcal{N}\left(\frac{d}{1+\tau}, ((1+\tau)\omega)^{-1}\right)
\end{aligned}
$$

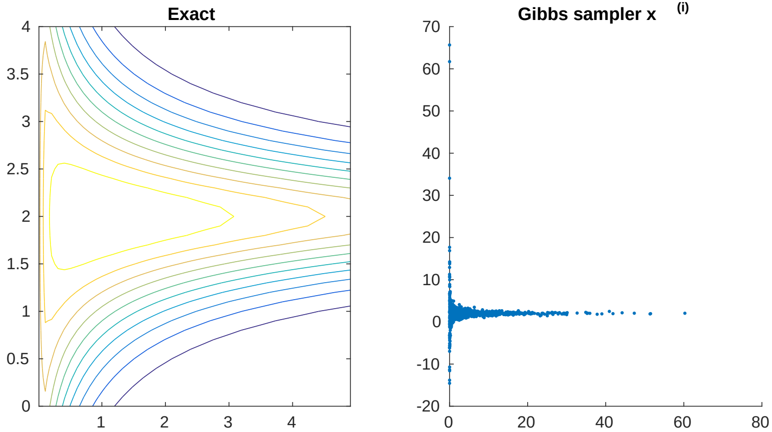
Conditional $p(\omega|d,m)$:

$$
\begin{aligned}
p(\omega|d,m) &\propto \omega^{\alpha}\exp\left[-\frac{1}{2}\omega\left((d-m)^2 + \tau m^2 + 2\beta\right)\right] \\
&= G\left(\alpha + 1, \beta + \frac{1}{2}\left((d-m)^2 + \tau m^2\right)\right)
\end{aligned}
$$

Toy: Gibbs sampler



Toy: Gibbs sampler



Assignments

- Classification scale: A: 50+, B: 40-50, C: 30-40
- Small assignment: one model, one or two methods

	$d = m + e$, $var(m) = \tau\omega$	$d = m + e$, $var(m) = \tau$	$d = m + e$, $var(m) = \tau, p(\tau)$
Laplace	5	5	5
Variational Bayes	5	8	8
Importance Sampling	5	8	8
Gibbs Sampling	5	8	8
2 methods	8	12	12