Binary probit regression with I-priors

Haziq Jamil
Supervisors: Dr. Wicher Bergsma & Prof. Irini Moustaki

Social Statistics (Year 3)
London School of Economics & Political Science

8-9 May 2017

PhD Presentation Event

http://phd3.haziqj.ml

Outline

Variational inference introduction

- Name derived from calculus of variations which deals with maximising or minimising functionals.

$$\begin{array}{lll} \text{Functions} & p : \theta \mapsto \mathbb{R} & \text{(standard calculus)} \\ \text{Functionals} & \mathcal{H} : p \mapsto \mathbb{R} & \text{(variational calculus)} \end{array}$$

- Using standard calculus, we can solve

$$\arg\max_\theta p(\theta) =: \hat\theta$$

e.g. p is a likelihood function, and $\hat\theta$ is the ML estimate.

- Using variational calculus, we can solve

$$\arg\max_p \mathcal{H}(p) =: \tilde p$$

e.g. $\mathcal{H}$ is the entropy $\mathcal{H} = -\int p(x) \log p(x) \, \mathrm{d}x$, and $\tilde p$ is the entropy maximising distribution.

C. M. Bishop (2006). *Pattern Recognition and Machine Learning*. Springer

Variational inference introduction (cont.)

- Consider a statistical model where we have observations $(y_1, \dots, y_n)$ and also some latent variables $(z_1, \dots, z_n)$.
- The z_i could be random effects or some auxiliary latent variables.
- In a Bayesian setting, this could also include the parameters to be estimated.
- **GOAL**: Find approximations for
  - The posterior distribution $p(\mathbf{z}|\mathbf{y})$; and
  - The marginal likelihood (or model evidence) $p(\mathbf{y})$.
- Variational inference is a deterministic approach, unlike MCMC.

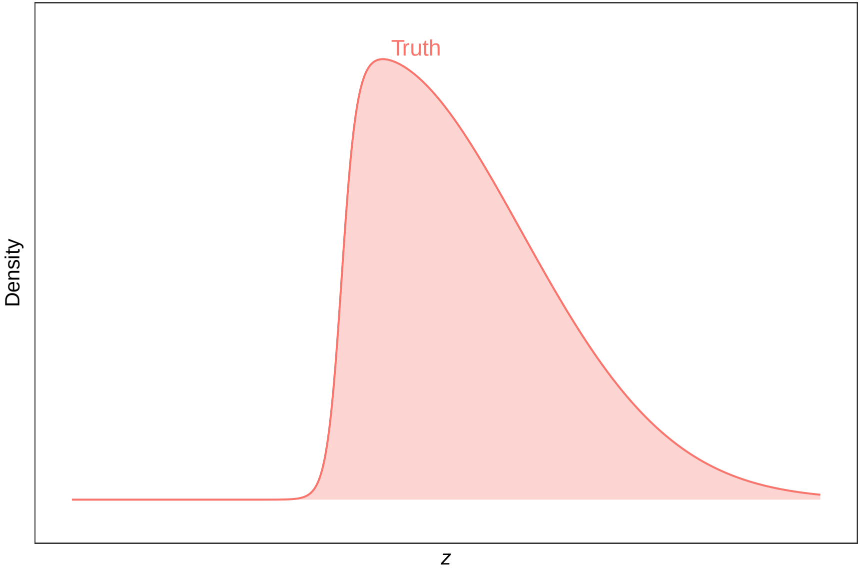
Decomposition of the log marginal

- Let $q(\mathbf{z})$ be some density function to approximate $p(\mathbf{z}|\mathbf{y})$. Then the log-marginal density can be decomposed into

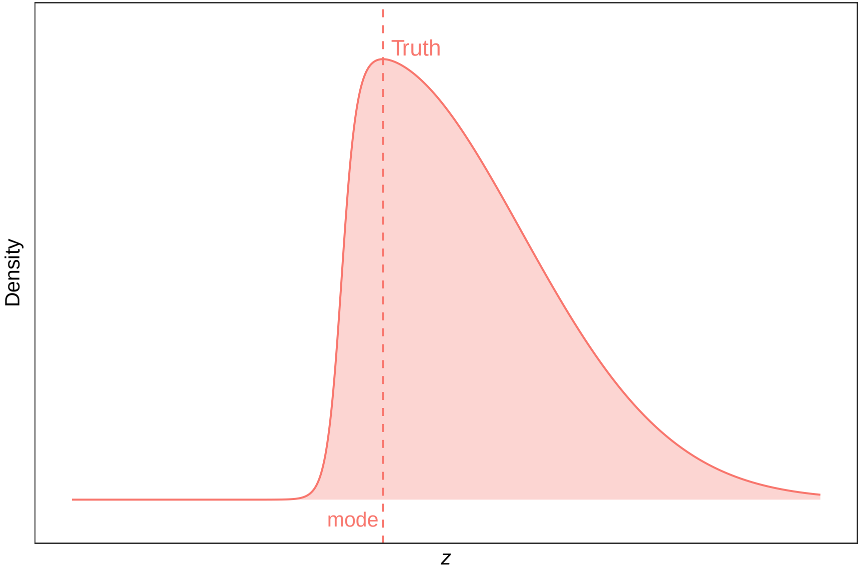
$$
\begin{aligned}
\log p(\mathbf{y}) &= \log p(\mathbf{y}, \mathbf{z}) - \log p(\mathbf{z}|\mathbf{y}) \\
&= \int \left\{ \log \frac{p(\mathbf{y}, \mathbf{z})}{q(\mathbf{z})} - \log \frac{p(\mathbf{z}|\mathbf{y})}{q(\mathbf{z})} \right\} q(\mathbf{z}) \, d\mathbf{z} \\
&= \mathcal{L}(q) + \mathrm{KL}(q \| p) \\
&\geq \mathcal{L}(q)
\end{aligned}
$$

- $\mathcal{L}$ is referred to as the "lower-bound", and it serves as a surrogate function to the marginal.
- Maximising the $\mathcal{L}(q)$ is equivalent to minimising $\mathrm{KL}(q\|p)$.
- Although $\mathrm{KL}(q\|p)$ is minimised at $q(\mathbf{z}) \equiv p(\mathbf{z}|\mathbf{y})$ (c.f. EM algorithm), we are unable to work with $p(\mathbf{z}|\mathbf{y})$.

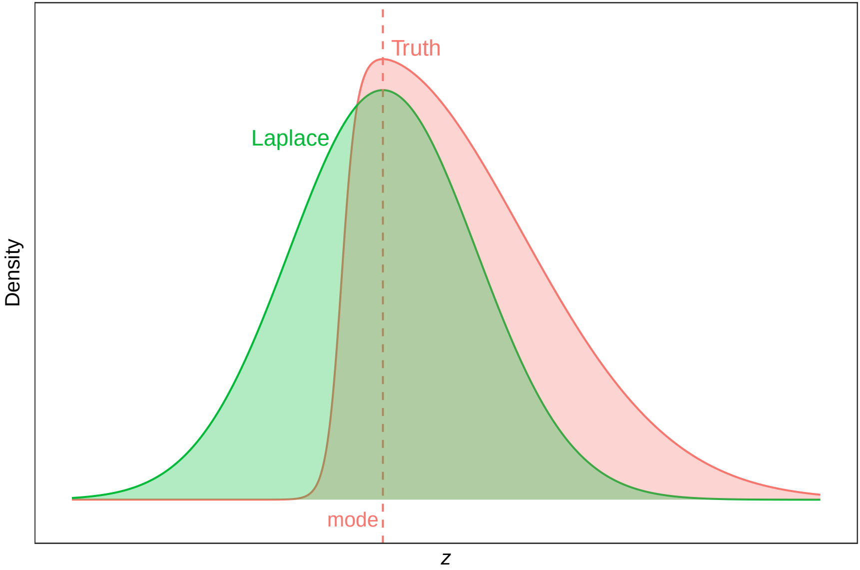
Comparison of approximations (density)



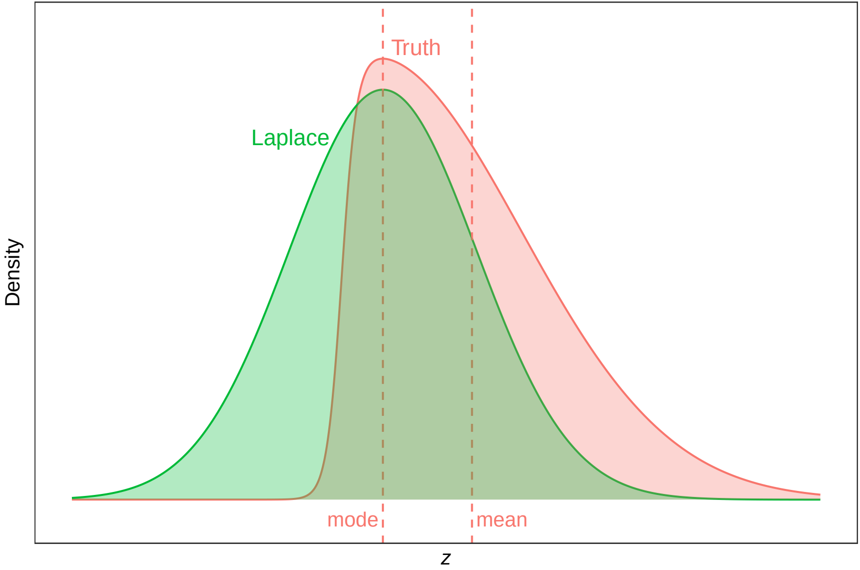
Comparison of approximations (density)



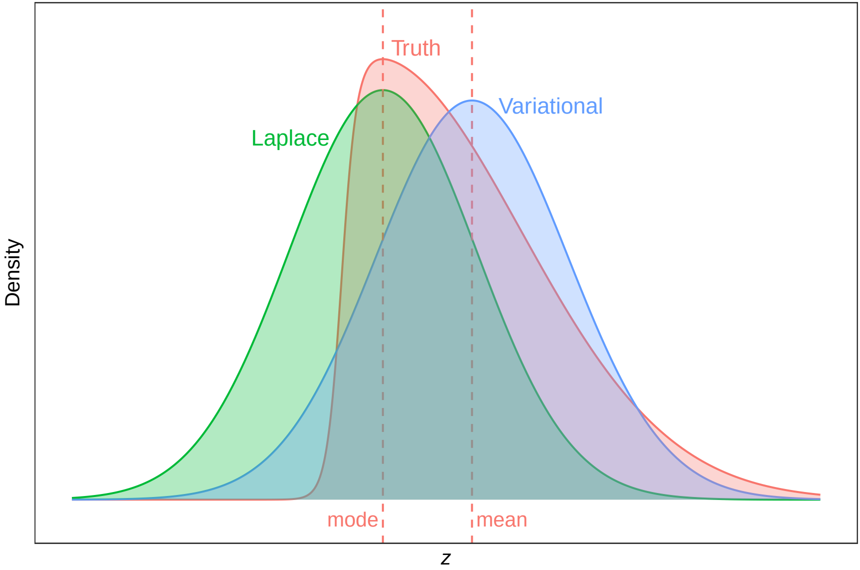
Comparison of approximations (density)



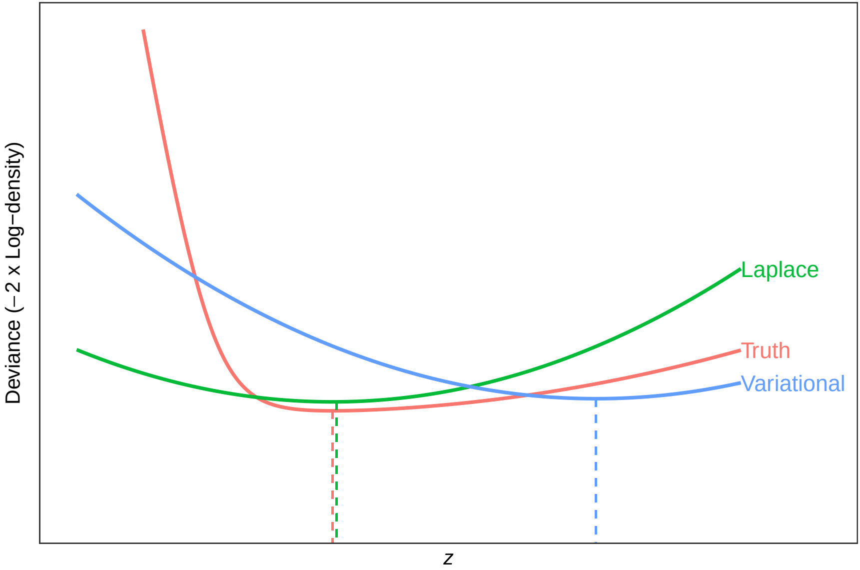
Comparison of approximations (density)



Comparison of approximations (density)



Comparison of approximations (deviance)



Factorised distributions (Mean-field theory)

- Maximising $\mathcal{L}$ over all possible q not feasible. Need some restrictions, but only to achieve tractability.
- Suppose we partition elements of $\mathbf{z}$ into m disjoint groups $\mathbf{z} = (\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(m)})$, and assume

$$q(\mathbf{z}) = \prod_{j=1}^{m} q_j(\mathbf{z}^{(j)})$$

- Under this restriction, the solution to $\arg\max_q \mathcal{L}(q)$ is

$$\tilde{q}_j(\mathbf{z}^{(j)}) \propto \exp\left(\mathrm{E}_{-j}[\log p(\mathbf{y}, \mathbf{z})]\right) \tag{1}$$

  for $j \in \{1, \dots, m\}$.
- In practice, these unnormalised densities are of recognisable form (especially if conjugate priors are used).

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe (2016). "Variational Inference: A Review for Statisticians". arXiv: 1601.00670

Coordinate ascent mean-field variational inference (CAVI)

- The optimal distributions are coupled with another, i.e. each $\tilde{q}_j(\mathbf{z}^{(j)})$ depends on the optimal moments of $\mathbf{z}^{(k)}$, $k \in \{1, \dots, m : k \neq j\}$.
- One way around this to employ an iterative procedure.
- Assess convergence by monitoring the lower bound

$$\mathcal{L}(q) = \mathrm{E}_q[\log p(\mathbf{y}, \mathbf{z})] - \mathrm{E}_q[\log q(\mathbf{z})].$$

Algorithm 1 CAVI

1: **initialise** Variational factors $q_j(\mathbf{z}^{(j)})$
2: **while** $\mathcal{L}(q)$ not converged **do**
3:     **for** $j = 1, \dots, m$ **do**
4:         $\log q_j(\mathbf{z}^{(j)}) \leftarrow \mathrm{E}_{-j}[\log p(\mathbf{y}, \mathbf{z})] + \text{const.}$
5:     **end for**
6:     $\mathcal{L}(q) \leftarrow \mathrm{E}_q[\log p(\mathbf{y}, \mathbf{z})] - \mathrm{E}_q[\log q(\mathbf{z})]$
7: **end while**
8: **return** $\tilde{q}(\mathbf{z}) = \prod_{j=1}^m \tilde{q}_j(\mathbf{z}^{(j)})$

Estimation of a 1-dim Gaussian mean and variance

- GOAL: Bayesian inference of mean μ and variance ψ^{-1}

$$
\begin{aligned}
y_i &\overset{\text{iid}}{\sim} \mathrm{N}(\mu, \psi^{-1}) && \text{Data} \\
\mu | \psi &\sim \mathrm{N}\big(a, (b\psi)^{-1}\big) && \text{Priors} \\
\psi &\sim \Gamma(c, d) \\
i &= 1, \dots, n
\end{aligned}
$$

Estimation of a 1-dim Gaussian mean and variance

- GOAL: Bayesian inference of mean μ and variance ψ^{-1}

$$
\begin{aligned}
y_i &\overset{\text{iid}}{\sim} \mathrm{N}(\mu, \psi^{-1}) && \text{Data} \\
\mu | \psi &\sim \mathrm{N}\left(a, (b\psi)^{-1}\right) && \multirow{2}{*}{Priors} \\
\psi &\sim \Gamma(c, d) \\
i &= 1, \ldots, n
\end{aligned}
$$

- Substitute $p(\mu, \psi | \mathbf{y})$ with the mean-field approximation

$$q(\mu, \psi) = q_\mu(\mu) q_\psi(\psi)$$

Estimation of a 1-dim Gaussian mean and variance

- GOAL: Bayesian inference of mean μ and variance ψ^{-1}

$$
\begin{aligned}
y_i &\overset{\text{iid}}{\sim} \mathrm{N}(\mu, \psi^{-1}) && \text{Data} \\
\mu|\psi &\sim \mathrm{N}\left(a, (b\psi)^{-1}\right) && \text{Priors} \\
\psi &\sim \Gamma(c, d) \\
i &= 1, \dots, n
\end{aligned}
$$

- Substitute $p(\mu, \psi|\mathbf{y})$ with the mean-field approximation

$$q(\mu, \psi) = q_\mu(\mu) q_\psi(\psi)$$

- From (1), we can work out the solutions

Estimation of a 1-dim Gaussian mean and variance

- GOAL: Bayesian inference of mean μ and variance ψ^{-1}

- Under the mean-field restriction, the solution to $\arg\max_q \mathcal{L}(q)$ is

$$\tilde{q}_j(\mathbf{z}^{(j)}) \propto \exp\left(\mathrm{E}_{-j}[\log p(\mathbf{y}, \mathbf{z})]\right)$$

for $j \in \{1, \ldots, m\}$.

$$q(\mu, \psi) = q_\mu(\mu) q_\psi(\psi)$$

- From (1), we can work out the solutions

Estimation of a 1-dim Gaussian mean and variance

- GOAL: Bayesian inference of mean μ and variance ψ^{-1}

> - Under the mean-field restriction, the solution to $\arg\max_q \mathcal{L}(q)$ is
>
> $$\tilde{q}_j(\mathbf{z}^{(j)}) \propto \exp\left(\mathrm{E}_{-j}[\log p(\mathbf{y}, \mathbf{z})]\right)$$
>
> for $j \in \{1, \dots, m\}$.

$$q(\mu, \psi) = q_\mu(\mu) q_\psi(\psi)$$

- From (1), we can work out the solutions

$$\tilde{q}_\mu(\mu) \equiv \mathrm{N}\left(\frac{ab + n\bar{y}}{b+n}, \frac{1}{(b+n)\,\mathrm{E}_q[\psi]}\right)$$

Estimation of a 1-dim Gaussian mean and variance

- GOAL: Bayesian inference of mean μ and variance ψ^{-1}

> - Under the mean-field restriction, the solution to $\arg\max_q \mathcal{L}(q)$ is
>
> $$\tilde{q}_j(\mathbf{z}^{(j)}) \propto \exp\left(\mathrm{E}_{-j}[\log p(\mathbf{y}, \mathbf{z})]\right)$$
>
> for $j \in \{1, \ldots, m\}$.

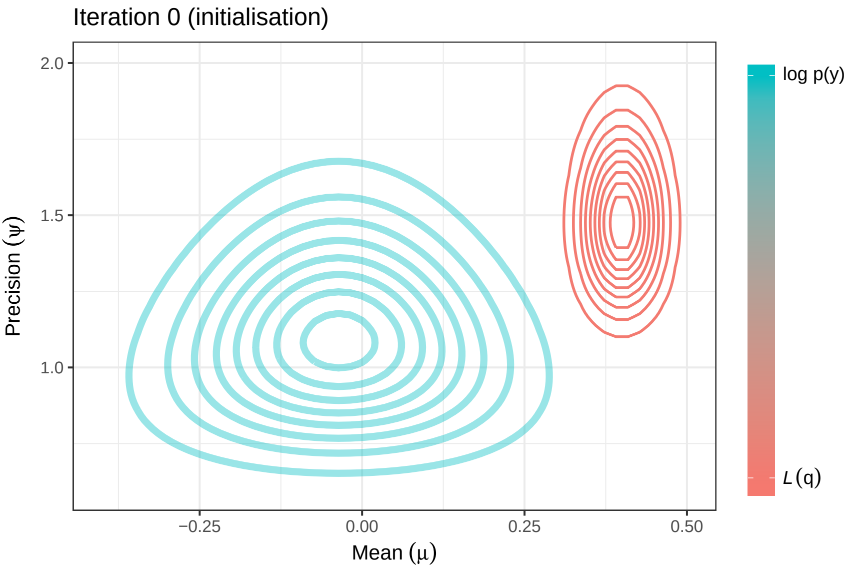
$$q(\mu, \psi) = q_\mu(\mu) q_\psi(\psi)$$

- From (1), we can work out the solutions

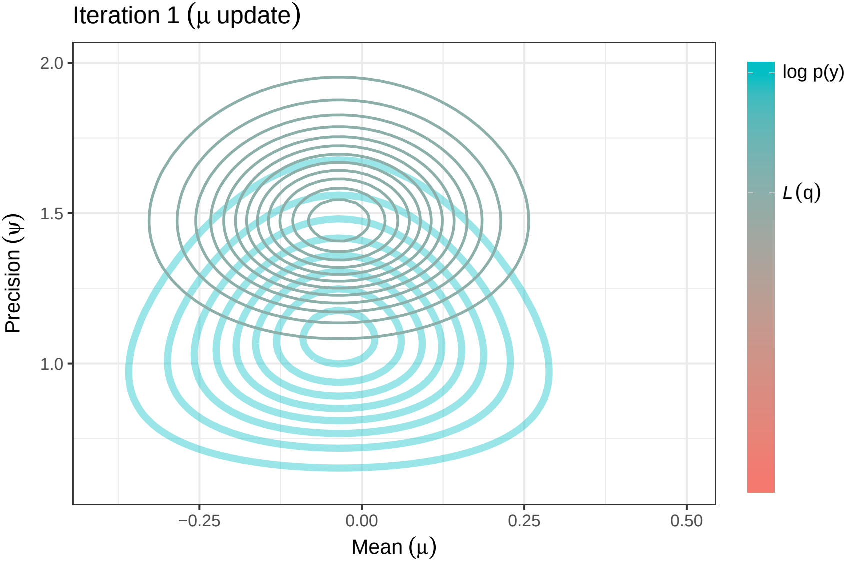
$$\tilde{q}_\mu(\mu) \equiv \mathrm{N}\left(\frac{ab + n\bar{y}}{b + n}, \frac{1}{(b+n)\,\mathrm{E}_q[\psi]}\right) \quad \text{and} \quad \tilde{q}_\psi(\psi) \equiv \Gamma(\tilde{c}, \tilde{d})$$

$$\tilde{c} = c + \frac{n}{2} \qquad \tilde{d} = d + \frac{1}{2}\,\mathrm{E}_q\left[\textstyle\sum_{i=1}^n (y_i - \mu)^2 + b(\mu - a)^2\right]$$

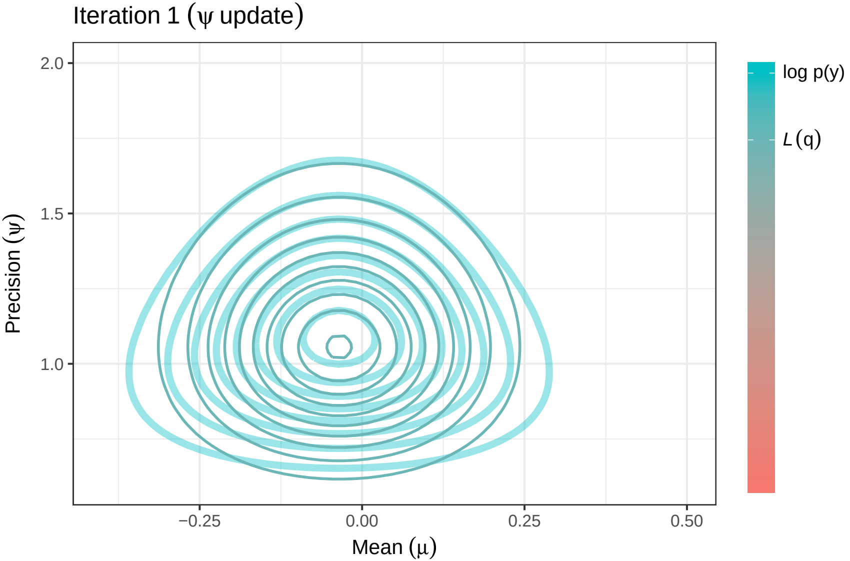
Estimation of a 1-dim Gaussian mean and variance (cont.)



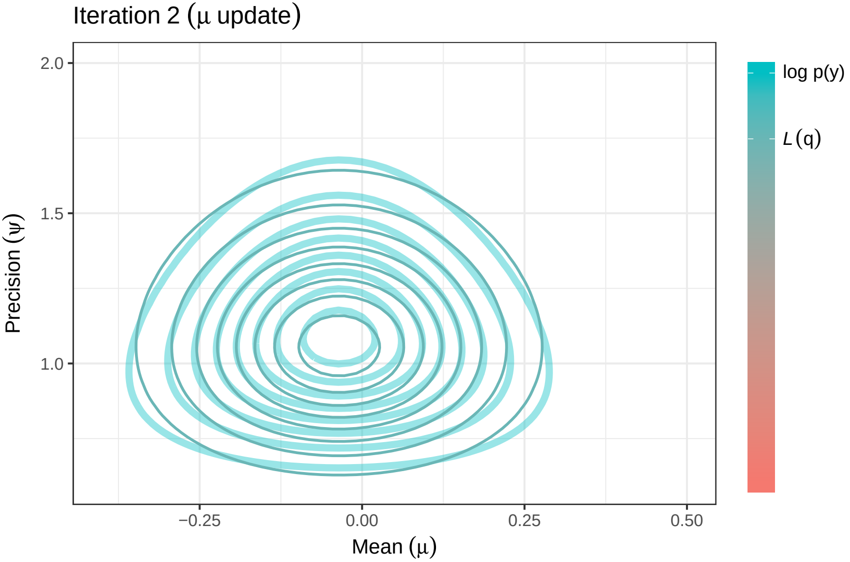
Estimation of a 1-dim Gaussian mean and variance (cont.)



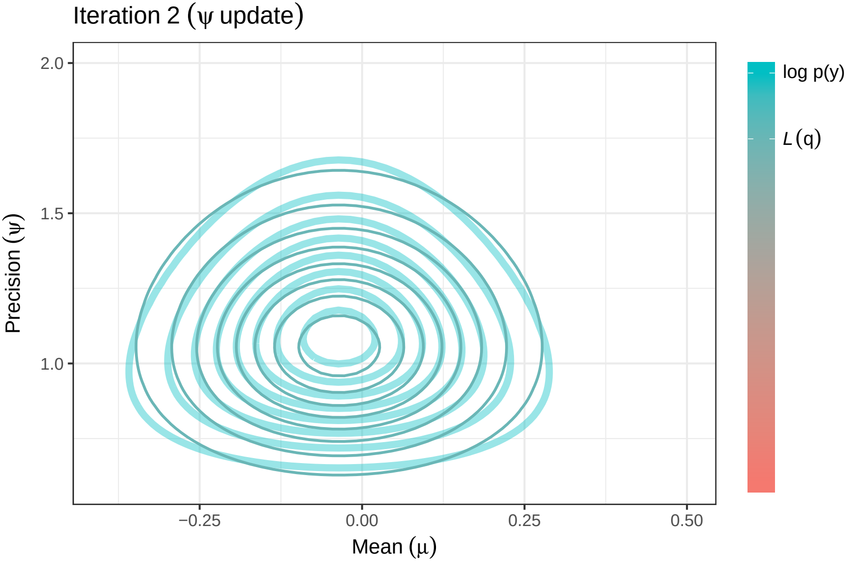
Estimation of a 1-dim Gaussian mean and variance (cont.)



Estimation of a 1-dim Gaussian mean and variance (cont.)



Estimation of a 1-dim Gaussian mean and variance (cont.)



End

Thank you!