Non-Gaussian likelihoods for Gaussian Processes

Alan Saul

PROWLER.io
the decision company

Outline

Motivation

Non-Gaussian posteriors

Approximate methods

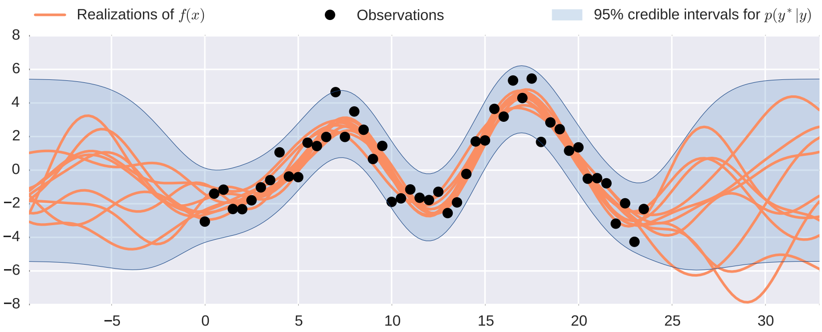
- Laplace approximation
- Variational bayes
- Expectation propagation
- Comparisons

GP regression - recap so far

Model the observations as a distorted version of the process $\mathbf{f}_i = f(\mathbf{x}_i)$:

$$\mathbf{y}_i \sim \mathcal{N}\left(f(\mathbf{x}_i), \sigma^2\right)$$

f is a non-linear function, in our case we assume it is latent, and is assigned a Gaussian process prior.



GP regression setting

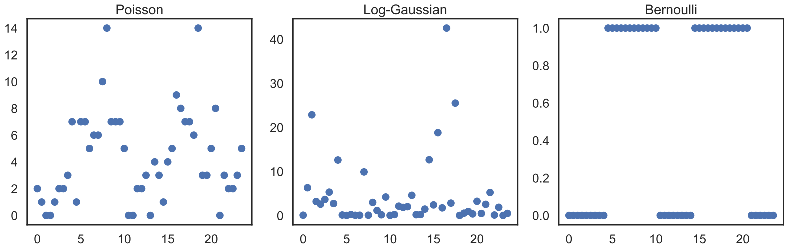
So far we have assumed that the latent values, $\mathbf{f}$, have been corrupted by Gaussian noise. Everything remains analytically tractable.

$$
\begin{aligned}
\text{Gaussian Prior:} \quad & \mathbf{f} \sim \mathcal{GP}(\mathbf{0}, \mathbf{K}_{\mathbf{ff}}) = p(\mathbf{f}) \\
\text{Gaussian likelihood:} \quad & \mathbf{y} \sim \mathcal{N}\left(\mathbf{f}, \sigma^2\mathbf{I}\right) = \prod_{i=1}^{n} p(\mathbf{y}_i|\mathbf{f}_i) \\
\text{Gaussian posterior:} \quad & p(\mathbf{f}|\mathbf{y}) \propto \mathcal{N}\left(\mathbf{y}|\mathbf{f}, \sigma^2\mathbf{I}\right)\mathcal{N}\left(\mathbf{f}|\mathbf{0}, \mathbf{K}_{\mathbf{ff}}\right)
\end{aligned}
$$

Outline

Motivation

- You have been given some data you wish to model.
- You believe that the observations are connected through some underlying unknown function.
- You know from your understanding of the data generation process, that the observations are not Gaussian.
- You still want to learn, as best as possible, what is the unknown function being used, and make predictions.



Likelihood

- $p(\mathbf{y}|\mathbf{f})$ is the probability that we would see some random variables, $\mathbf{y}$, if we knew the latent function values $\mathbf{f}$, which act as parameters.

Likelihood

- $p(\mathbf{y}|\mathbf{f})$ is the probability that we would see some random variables, $\mathbf{y}$, if we knew the latent function values $\mathbf{f}$, which act as parameters.
- Given the observed values for $\mathbf{y}$ are fixed, it can also be seen as the likelihood that some latent function values, $\mathbf{f}$, would give rise to the observed values of $\mathbf{y}$. Note this is a *function* of $\mathbf{f}$, and doesn't integrate to 1 in $\mathbf{f}$.

Likelihood

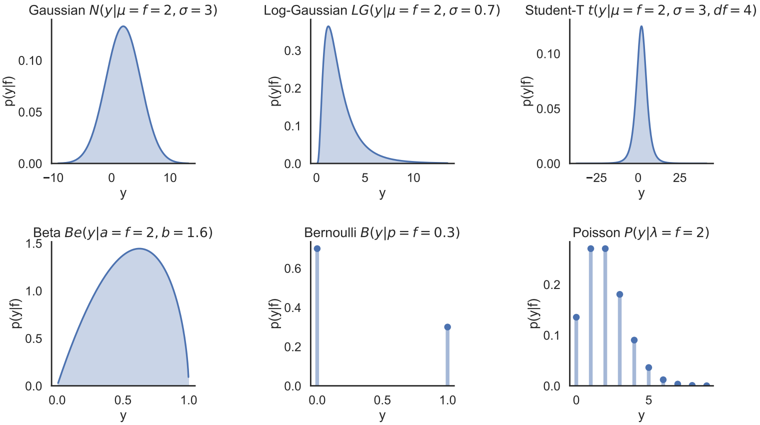
- $p(\mathbf{y}|\mathbf{f})$ is the probability that we would see some random variables, $\mathbf{y}$, if we knew the latent function values $\mathbf{f}$, which act as parameters.
- Given the observed values for $\mathbf{y}$ are fixed, it can also be seen as the likelihood that some latent function values, $\mathbf{f}$, would give rise to the observed values of $\mathbf{y}$. Note this is a *function* of $\mathbf{f}$, and doesn't integrate to 1 in $\mathbf{f}$.
- Often observations aren't observed by simple Gaussian corruptions of the underlying latent function, $\mathbf{f}$.

Likelihood

- $p(\mathbf{y}|\mathbf{f})$ is the probability that we would see some random variables, $\mathbf{y}$, if we knew the latent function values $\mathbf{f}$, which act as parameters.
- Given the observed values for $\mathbf{y}$ are fixed, it can also be seen as the likelihood that some latent function values, $\mathbf{f}$, would give rise to the observed values of $\mathbf{y}$. Note this is a *function* of $\mathbf{f}$, and doesn't integrate to 1 in $\mathbf{f}$.
- Often observations aren't observed by simple Gaussian corruptions of the underlying latent function, $\mathbf{f}$.
- In the case of count data, binary data, etc, we need to choose a different likelihood function.

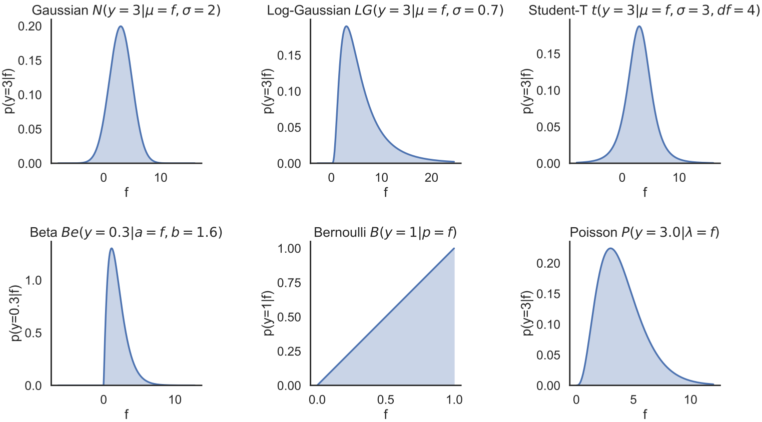
Likelihood

$p(y|f)$ as a function of y, with fixed f



Likelihood

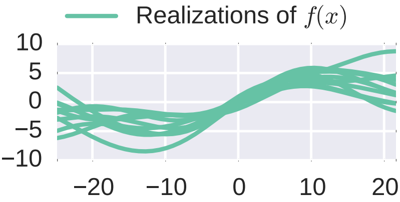
$p(y|f)$ as a function of f, with fixed y

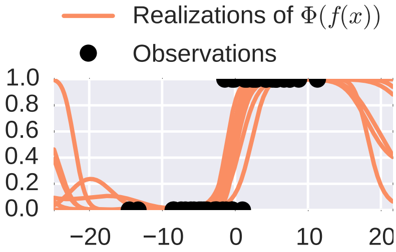


Binary example

- Binary outcomes for $\mathbf{y}_i$, $\mathbf{y}_i \in [0, 1]$.
- Model the probability of $\mathbf{y}_i = 1$ with transformation of GP, with Bernoulli likelihood.
- Probability of 1 must be between 0 and 1, thus use squashing transformation, $\lambda(\mathbf{f}_i) = \Phi(\mathbf{f}_i)$.

$$p(\mathbf{y}_i|\lambda(\mathbf{f}_i)) = \begin{cases} \lambda(\mathbf{f}_i), & \text{if } \mathbf{y}_i = 1 \\ 1 - \lambda(\mathbf{f}_i), & \text{if } \mathbf{y}_i = 0 \end{cases}$$

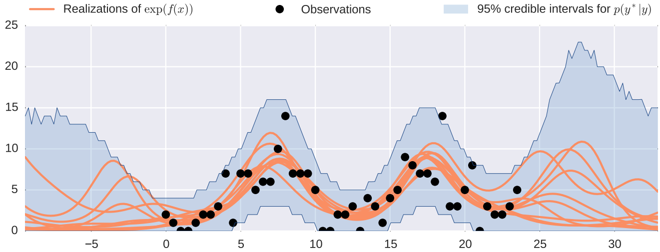




Count data example

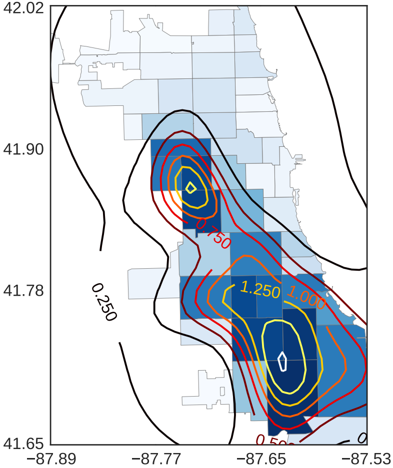
- Non-negative and discrete values only for $\mathbf{y}_i$, $\mathbf{y}_i \in \mathbb{N}$.
- Model the *rate* or *intensity*, λ, of events with a transformation of a Gaussian process.
- Rate parameter must remain positive, use transformation to maintain positiveness $\lambda(\mathbf{f}_i) = \exp(\mathbf{f}_i)$ or $\lambda(\mathbf{f}_i) = \mathbf{f}_i^2$

$$\mathbf{y}_i \sim \text{Poisson}(\mathbf{y}_i|\lambda_i = \lambda(\mathbf{f}_i)) \qquad \text{Poisson}(\mathbf{y}_i|\lambda_i) = \frac{\lambda_i^{\mathbf{y}_i}}{!\mathbf{y}_i} e^{-\lambda_i}$$



Application example

- Chicago crime counts.
- Same Poisson likelihood.
- 2D-input to kernel.



Outline

Motivation

Non-Gaussian posteriors

Approximate methods

- Laplace approximation
- Variational bayes
- Expectation propagation
- Comparisons

Non-Gaussian posteriors

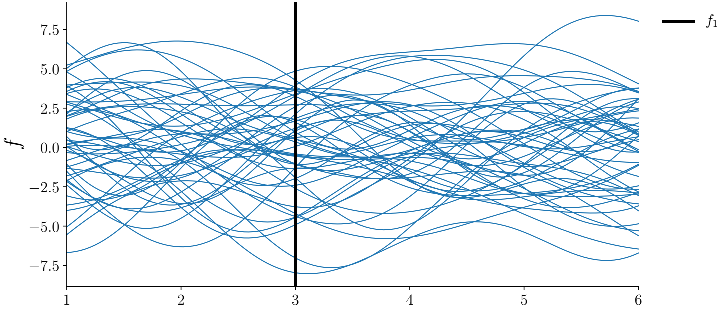
- Exact computation of posterior is no longer analytically tractable due to non-conjugate Gaussian process prior to non-Gaussian likelihood, $p(\mathbf{y}|\mathbf{f})$.

$$p(\mathbf{f}|\mathbf{y}) = \frac{p(\mathbf{f}) \prod_{i=1}^{n} p(\mathbf{y}_i|\mathbf{f}_i)}{\int p(\mathbf{f}) \prod_{i=1}^{n} p(\mathbf{y}_i|\mathbf{f}_i)\, d\mathbf{f}}$$

Why is it so difficult?

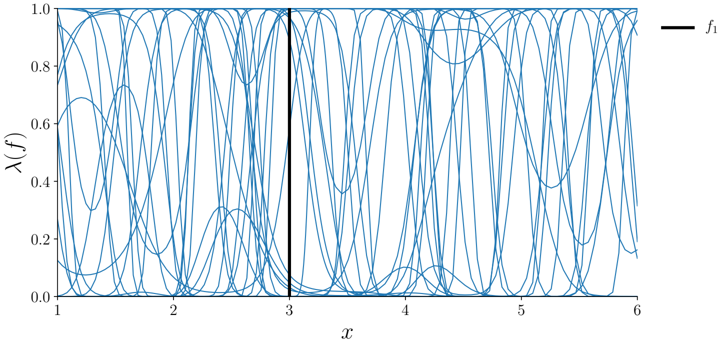
Non-Gaussian posteriors illustrated

- Consider one observation, $y_1 = 1$, at input x_1.
- Can normalise easily with numerical integration, $\int p(y_1 = 1|\lambda(f_1))p(f_1)df_1$.



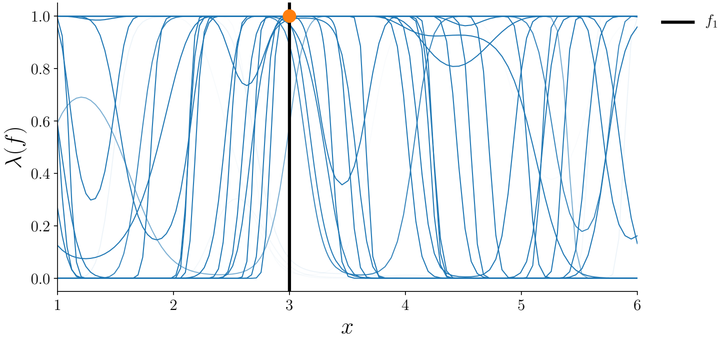
Non-Gaussian posteriors illustrated

- Consider one observation, $y_1 = 1$, at input x_1.
- Can normalise easily with numerical integration, $\int p(y_1 = 1|\lambda(f_1))p(f_1)df_1$.

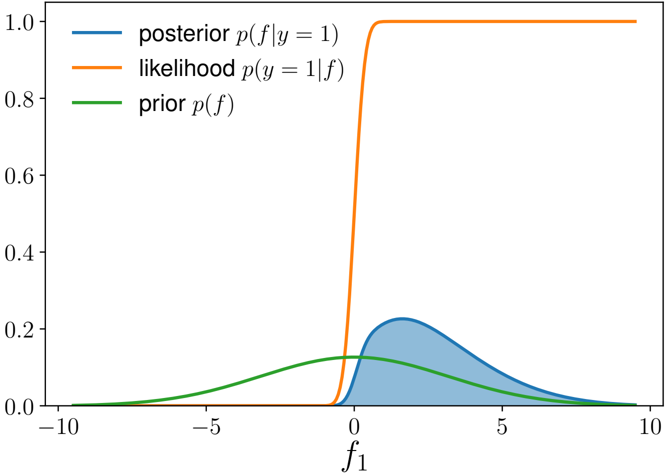


Non-Gaussian posteriors illustrated

- Consider one observation, $y_1 = 1$, at input x_1.
- Can normalise easily with numerical integration, $\int p(y_1 = 1|\lambda(f_1))p(f_1)df_1$.

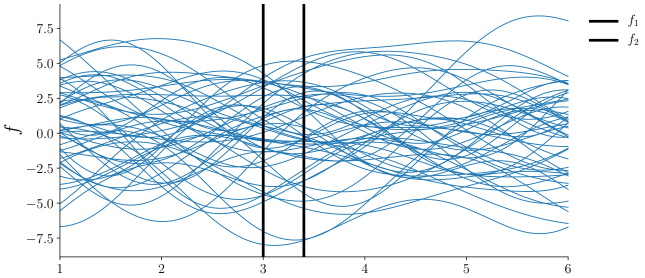


Non-Gaussian posteriors illustrated



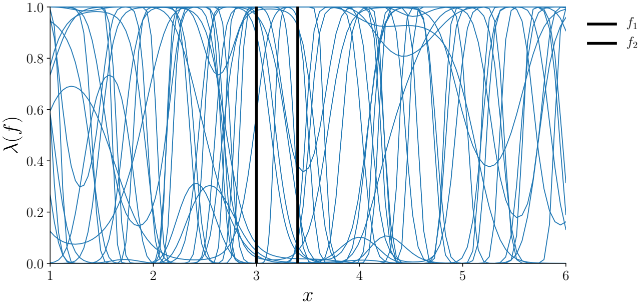
Non-Gaussian posteriors illustrated

- Now two observations, $y_1 = 1$ and $y_2 = 1$ at x_1 and x_2
- Need to calculate the joint posterior, $p(\mathbf{f}|\mathbf{y}) = p(f_1, f_2|y_1 = 1, y_2 = 1)$.
- Requires 2D integral $p(y_1 = 1, y_2 = 1|\lambda(f_1), \lambda(f_2))p(f_1, f_2)df_1df_2$.



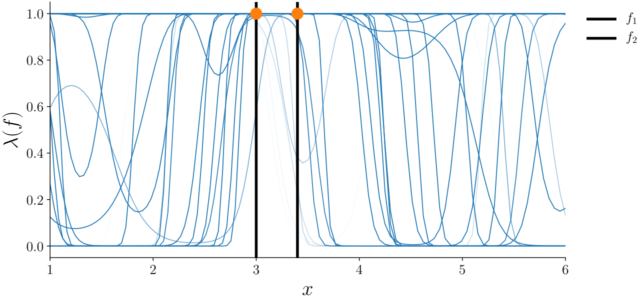
Non-Gaussian posteriors illustrated

- Now two observations, $y_1 = 1$ and $y_2 = 1$ at x_1 and x_2
- Need to calculate the joint posterior, $p(\mathbf{f}|\mathbf{y}) = p(f_1, f_2|y_1 = 1, y_2 = 1)$.
- Requires 2D integral $p(y_1 = 1, y_2 = 1|\lambda(f_1), \lambda(f_2))p(f_1, f_2)df_1df_2$.



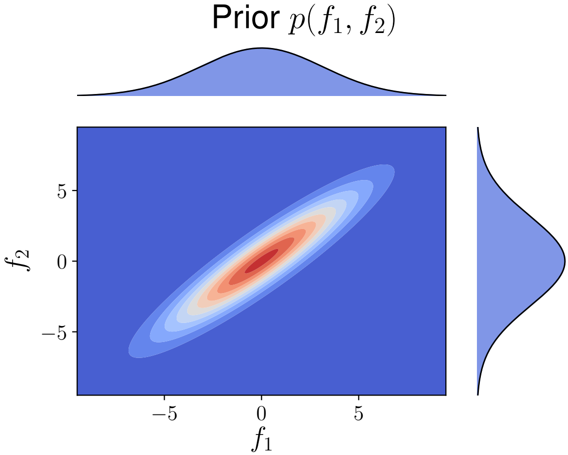
Non-Gaussian posteriors illustrated

- Now two observations, $y_1 = 1$ and $y_2 = 1$ at x_1 and x_2
- Need to calculate the joint posterior, $p(\mathbf{f}|\mathbf{y}) = p(f_1, f_2|y_1 = 1, y_2 = 1)$.
- Requires 2D integral $p(y_1 = 1, y_2 = 1|\lambda(f_1), \lambda(f_2))p(f_1, f_2)df_1 df_2$.



Non-Gaussian posteriors illustrated

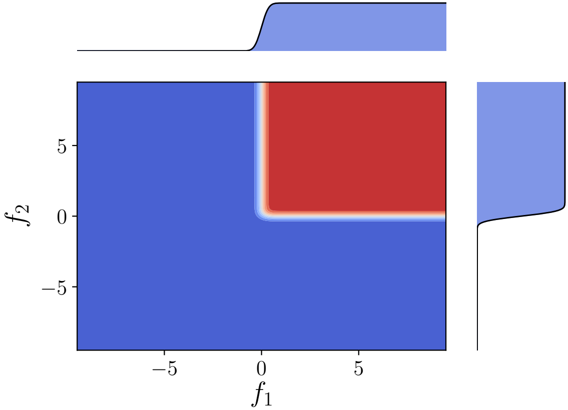
- To find the true posterior values, we need to perform a two dimensional integral.
- Still possible, but things are getting more difficult quickly.



Non-Gaussian posteriors illustrated

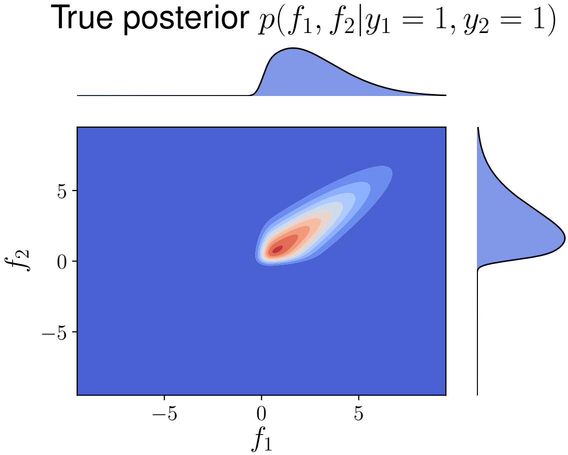
- To find the true posterior values, we need to perform a two dimensional integral.
- Still possible, but things are getting more difficult quickly.

Likelihood $p(y_1 = 1, y_2 = 1 | f_1, f_2)$



Non-Gaussian posteriors illustrated

- To find the true posterior values, we need to perform a two dimensional integral.
- Still possible, but things are getting more difficult quickly.

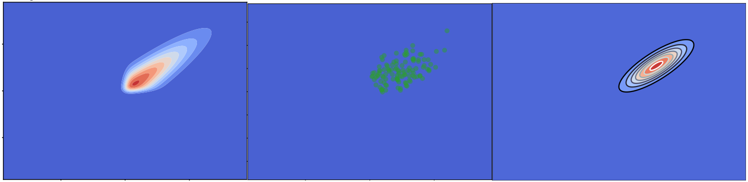


Approaches to handling non-Gaussian posteriors

Generally fall into two areas:

- Sampling methods that obtain samples of the posterior.
- Approximation of the posterior with something of known form.

Today we will focus on the latter.



Non-Gaussian posterior approximation

- Various methods to make a Gaussian approximation, $p(\mathbf{f}|\mathbf{y}) \approx q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mu =?, C =?)$.
- Only need to obtain an approximate posterior at the training locations.
- At test locations, the data only effects their probabily via the posterior at these locations.

$$
\begin{aligned}
p(\mathbf{f}, \mathbf{f}^*|\mathbf{x}^*, \mathbf{x}, \mathbf{y}) &= p(\mathbf{f}^*|\mathbf{f}, \mathbf{x}^*)p(\mathbf{f}|\mathbf{x}, \mathbf{y}) \\
&\propto p(\mathbf{f}^*|\mathbf{f}, \mathbf{x}^*)p(\mathbf{y}|\mathbf{f}, \mathbf{x})p(\mathbf{f}|\mathbf{x})
\end{aligned}
$$

Why do we want an the posterior anyway?

True posterior, posterior approximation, or samples are needed to make predictions at new locations, $\mathbf{x}^*$.

$$p(\mathbf{f}^*|\mathbf{x}^*, \mathbf{x}, \mathbf{y}) = \int p(\mathbf{f}^*|\mathbf{f}, \mathbf{x}^*) p(\mathbf{f}|\mathbf{y}, \mathbf{x}) d\mathbf{f}$$

$$q(\mathbf{f}^*|\mathbf{x}^*, \mathbf{x}, \mathbf{y}) = \int p(\mathbf{f}^*|\mathbf{f}, \mathbf{x}^*) q(\mathbf{f}|\mathbf{x}) d\mathbf{f}$$

Outline

Motivation

Non-Gaussian posteriors

Approximate methods

- Laplace approximation
- Variational bayes
- Expectation propagation
- Comparisons

Methods overview

Given choice of Gaussian approximation of posterior. How do we choose the parameter values μ and $\boldsymbol{C}$?

There a number of different methods in which to choose how to set the parameters of our Gaussian approximation.

Parameters effect - mean

Parameters effect - variance

How to choose the parameters?

Two approaches that we might take:

- Is to match the mean and variance at some point, for example the mode.
- Attempt to minimise some divergence measure between the approximate distribution and the true distribution.

- Laplace takes the former
- Variational bayes takes the latter
- EP kind of takes the latter

Outline

Laplace approximation

Task: for some generic random variable, z, and data, y, find a good approximation to difficult to compute posterior distribution, $p(z|y)$.

Laplace approach: fit a Gaussian by matching the curvature at the modal point of the posterior.

- Use a second-order taylor expansion around the mode of the log-posterior.
- Use the expansion to find an equivalent Gaussian in the probability space.

Laplace approximation

- Log of a Gaussian distribution, $q(\mathbf{f}) = \mathcal{N}\left(\mathbf{f}|\hat{\mu}, \hat{C}\right)$, is a quadratic function of $\mathbf{f}$.
- A second-order taylor expansion is an approximation of a function using only quadratic terms.
- Laplace approximation expands the un-normalised posterior, and then uses it to set the linear and quadratic terms of the log $q(\mathbf{f})$.
- The first and second derivatives of the form of the log-posterior, at the mode, will match the derivatives of the approximate Gaussian at this same point.

Second-order taylor expansion

$$p(\mathbf{f}|\mathbf{y}) = \frac{1}{Z}h(\mathbf{f})$$

Second-order taylor expansion

$$p(\mathbf{f}|\mathbf{y}) = \frac{1}{Z}h(\mathbf{f})$$

$$\log p(\mathbf{f}|\mathbf{y}) = \log\frac{1}{Z} + \log h(\mathbf{f})$$

Second-order taylor expansion

$$
\begin{aligned}
p(\mathbf{f}|\mathbf{y}) &= \frac{1}{Z} h(\mathbf{f}) \\
\log p(\mathbf{f}|\mathbf{y}) &= \log \frac{1}{Z} + \log h(\mathbf{f}) \\
&\approx \log \frac{1}{Z} + \log h(\mathbf{a}) + \frac{d \log h(\mathbf{a})}{d\mathbf{a}}(\mathbf{f} - \mathbf{a}) \\
&+ \frac{1}{2}(\mathbf{f} - \mathbf{a})^\top \frac{d^2 \log h(\mathbf{a})}{d\mathbf{a}^2}(\mathbf{f} - \mathbf{a})
\end{aligned}
$$

Second-order taylor expansion

$$
\begin{aligned}
p(\mathbf{f}|\mathbf{y}) &= \frac{1}{Z} h(\mathbf{f}) \\
\log p(\mathbf{f}|\mathbf{y}) &= \log \frac{1}{Z} + \log h(\mathbf{f}) \\
&\approx \log \frac{1}{Z} + \log h(\mathbf{a}) + \frac{d \log h(\mathbf{a})}{d\mathbf{a}}(\mathbf{f} - \mathbf{a}) \\
&+ \frac{1}{2}(\mathbf{f} - \mathbf{a})^\top \frac{d^2 \log h(\mathbf{a})}{d\mathbf{a}^2}(\mathbf{f} - \mathbf{a})
\end{aligned}
$$

In our case we want to make our expansion around the mode, $\hat{\mathbf{f}}$:

$$
\left. \frac{d \log h(\mathbf{a})}{d\mathbf{a}} \right|_{\mathbf{a}=\hat{\mathbf{f}}} = \mathbf{0}
$$

Second-order taylor expansion

$$\log p(\mathbf{f}|\mathbf{y}) \approx \log \frac{1}{\color{red}{Z}} + \log h(\hat{\mathbf{f}}) + \frac{d \log h(\hat{\mathbf{f}})}{d\hat{\mathbf{f}}}(\mathbf{f} - \hat{\mathbf{f}})$$

$$+ \frac{1}{2}(\mathbf{f} - \hat{\mathbf{f}})^\top \frac{d^2 \log h(\hat{\mathbf{f}})}{d\hat{\mathbf{f}}^2}(\mathbf{f} - \hat{\mathbf{f}})$$

Second-order taylor expansion

$$\log p(\mathbf{f}|\mathbf{y}) \approx \log \frac{1}{{\color{red}Z}} + \log h(\hat{\mathbf{f}}) + \frac{d \log h(\hat{\mathbf{f}})}{d\hat{\mathbf{f}}}(\mathbf{f} - \hat{\mathbf{f}})$$
$$+ \frac{1}{2}(\mathbf{f} - \hat{\mathbf{f}})^\top \frac{d^2 \log h(\hat{\mathbf{f}})}{d\hat{\mathbf{f}}^2}(\mathbf{f} - \hat{\mathbf{f}})$$
$$p(\mathbf{f}|\mathbf{y}) = \frac{1}{{\color{red}Z}} h(\hat{\mathbf{f}}) \exp\left\{-\frac{1}{2}(\mathbf{f} - \hat{\mathbf{f}})^\top \left(-\frac{d^2 \log h(\hat{\mathbf{f}})}{d\mathbf{f}^2}\right)(\mathbf{f} - \hat{\mathbf{f}})\right\}$$

Second-order taylor expansion

$$\log p(\mathbf{f}|\mathbf{y}) \approx \log \frac{1}{Z} + \log h(\hat{\mathbf{f}}) + \frac{d \log h(\hat{\mathbf{f}})}{d\hat{\mathbf{f}}}(\mathbf{f} - \hat{\mathbf{f}})$$

$$+ \frac{1}{2}(\mathbf{f} - \hat{\mathbf{f}})^\top \frac{d^2 \log h(\hat{\mathbf{f}})}{d\hat{\mathbf{f}}^2}(\mathbf{f} - \hat{\mathbf{f}})$$

$$p(\mathbf{f}|\mathbf{y}) = \frac{1}{Z} h(\hat{\mathbf{f}}) \exp\left\{-\frac{1}{2}(\mathbf{f} - \hat{\mathbf{f}})^\top \left(-\frac{d^2 \log h(\hat{\mathbf{f}})}{d\mathbf{f}^2}\right)(\mathbf{f} - \hat{\mathbf{f}})\right\}$$

$$= \mathcal{N}\left(\mathbf{f}|\hat{\mathbf{f}}, \left(-\frac{d^2 \log h(\hat{\mathbf{f}})}{d\mathbf{f}^2}\right)^{-1}\right)$$

Laplace appoximation for Gaussian processes

In our case, $h(\mathbf{f}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f})$, so we need to evaluate

$$\begin{aligned}
-\frac{d^2 \log h(\hat{\mathbf{f}})}{d\mathbf{f}^2} &= -\frac{d^2(\log p(\mathbf{y}|\hat{\mathbf{f}}) + \log p(\hat{\mathbf{f}}))}{d\mathbf{f}^2} \\
&= -\frac{d^2 \log p(\mathbf{y}|\hat{\mathbf{f}})}{d\mathbf{f}^2} + \mathbf{K}^{-1} \\
&\triangleq \boldsymbol{W} + \mathbf{K}^{-1}
\end{aligned}$$

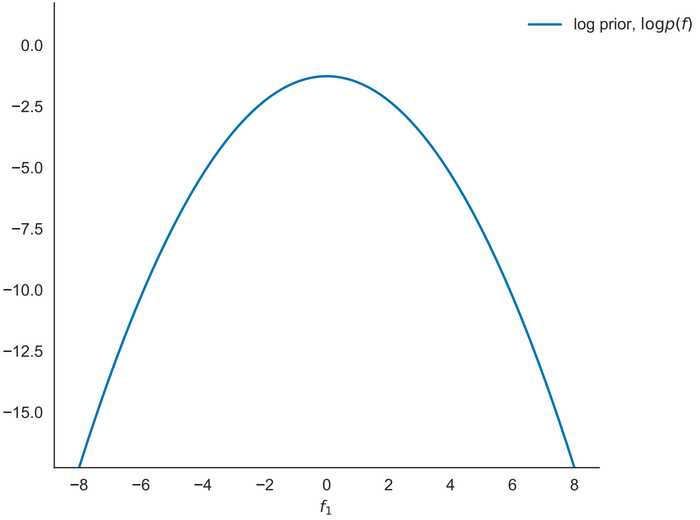
giving a posterior approximation:

$$p(\mathbf{f}|\mathbf{y}) \approx q(\mathbf{f}) = \mathcal{N}\left(\mathbf{f}|\hat{\mathbf{f}}, \left(\boldsymbol{W} + \mathbf{K}^{-1}\right)^{-1}\right)$$

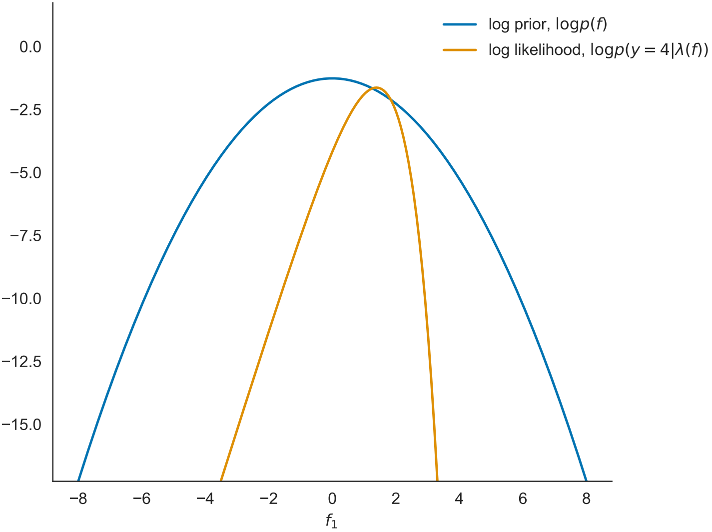
Laplace approximation - algorithm overview

- Find the mode, $\hat{\mathbf{f}}$ of the true log posterior, via Newton's method.
- Use second-order Taylor expansion around this modal value.
- Form Gaussian approximation setting the mean equal to the posterior mode, $\hat{\mathbf{f}}$, and matching the curvature.
- $p(\mathbf{f}|\mathbf{y}) \approx q(\mathbf{f}|\boldsymbol{\mu}, \boldsymbol{C}) = \mathcal{N}\left(\mathbf{f}|\hat{\mathbf{f}}, (\mathbf{K}^{-1} + \boldsymbol{W})^{-1}\right)$
- $\boldsymbol{W} \triangleq -\frac{d^2 \log p(\mathbf{y}|\hat{\mathbf{f}})}{d\hat{\mathbf{f}}^2}$.
- For factorizing likelihoods (most), $\boldsymbol{W}$ is diagonal.

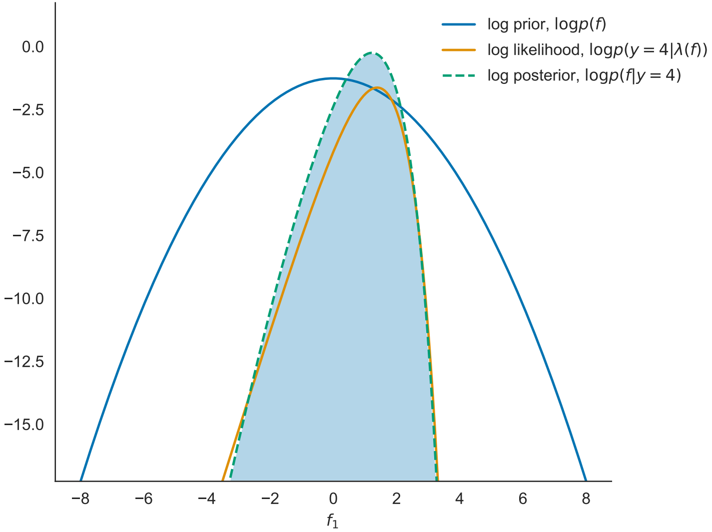
Visualization of Laplace



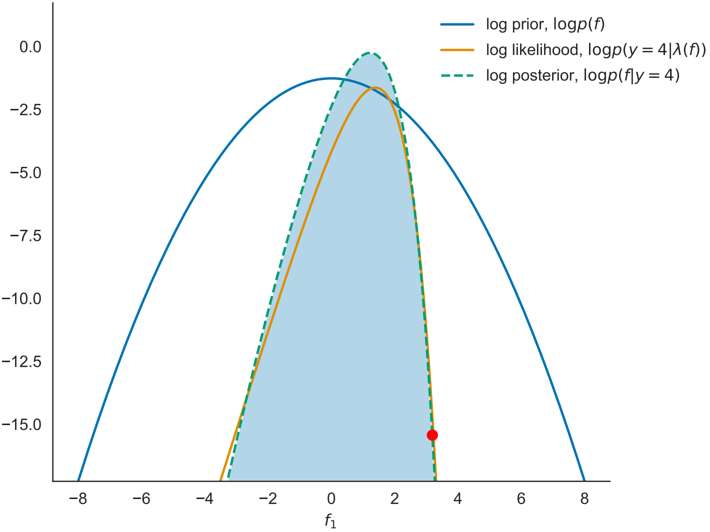
Visualization of Laplace



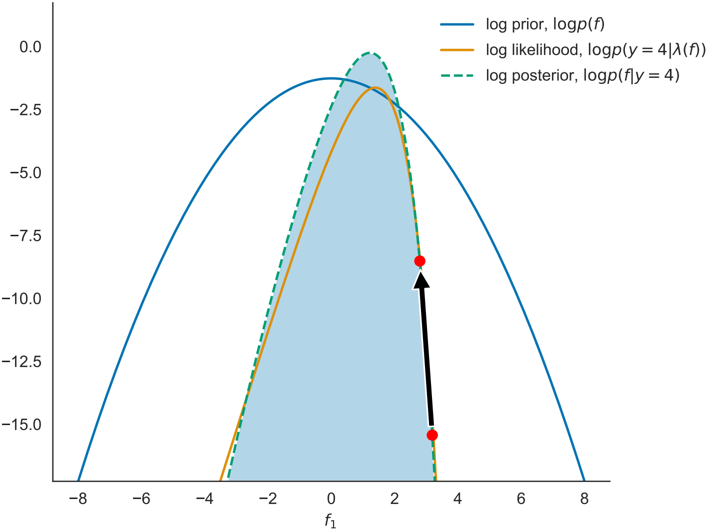
Visualization of Laplace



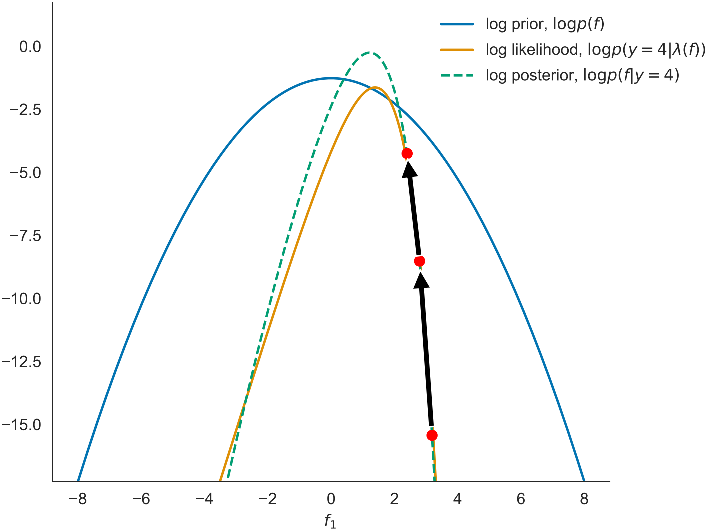
Visualization of Laplace



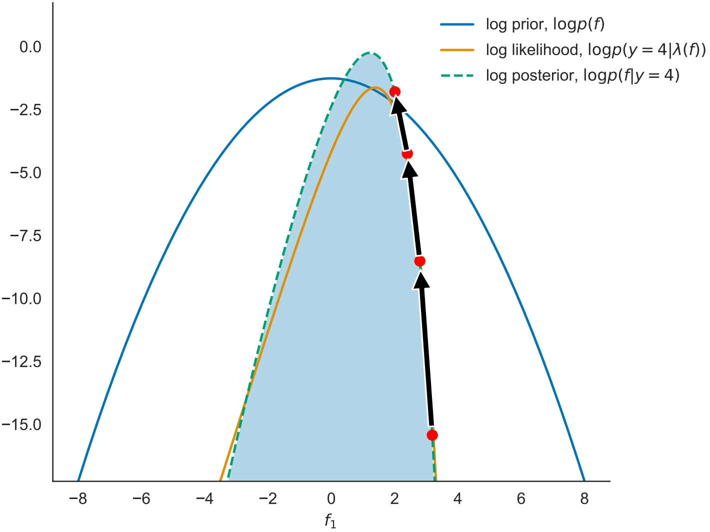
Visualization of Laplace



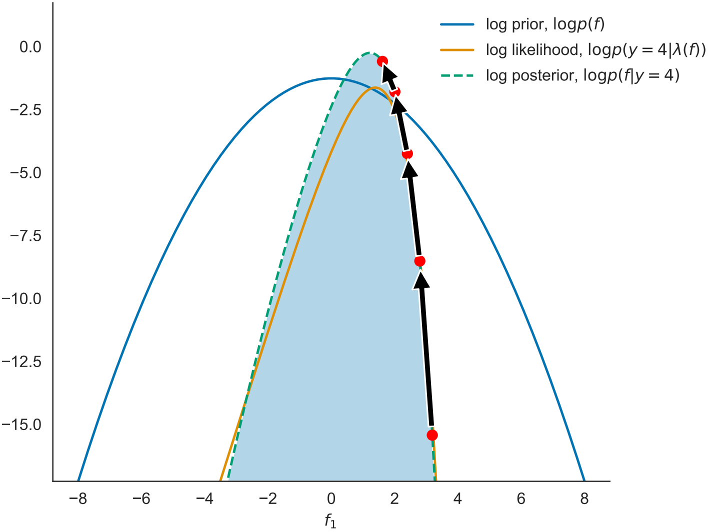
Visualization of Laplace



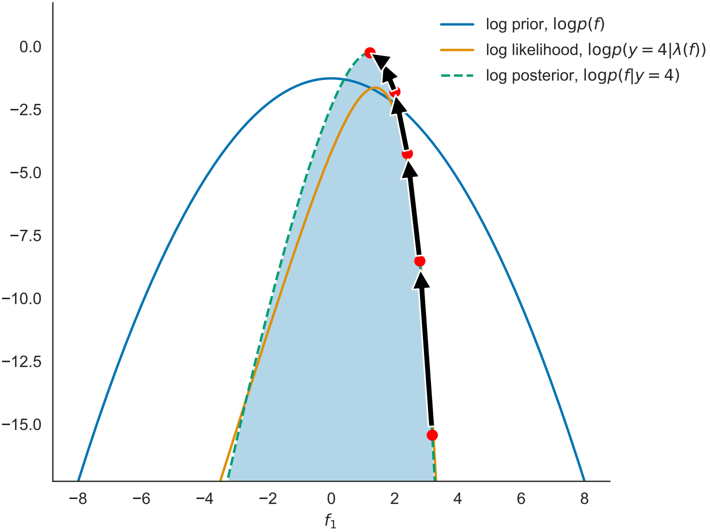
Visualization of Laplace



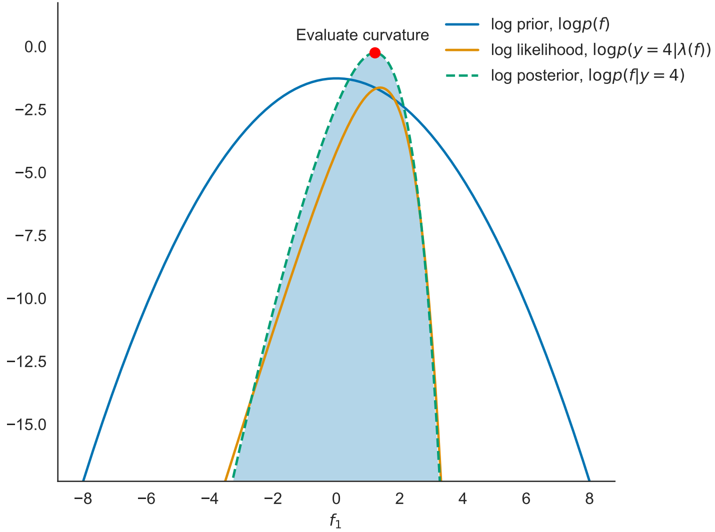
Visualization of Laplace



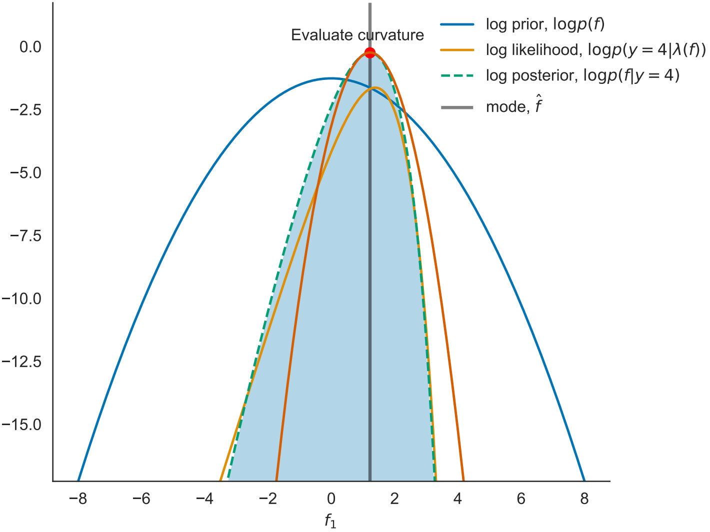
Visualization of Laplace



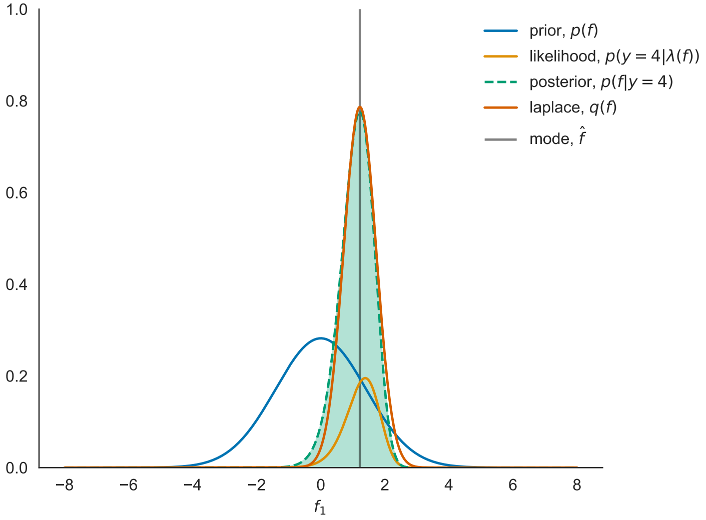
Visualization of Laplace



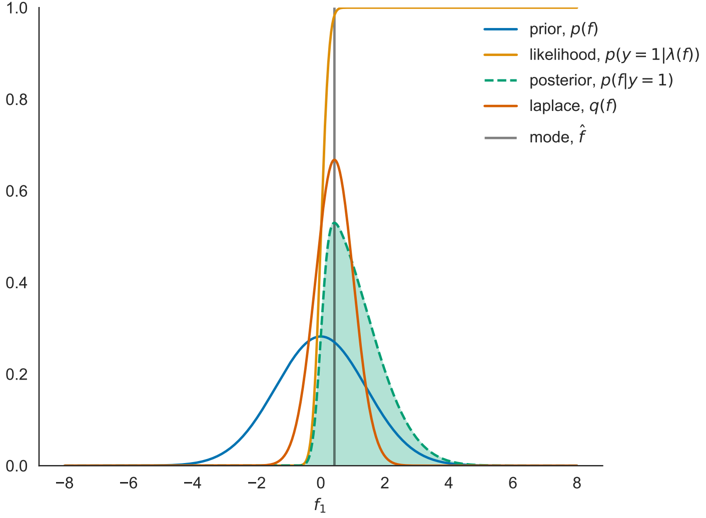
Visualization of Laplace



Visualization of Laplace



Visualise of Laplace - Bernoulli



Outline

Motivation

Non-Gaussian posteriors

Approximate methods

- Laplace approximation
- Variational bayes
- Expectation propagation
- Comparisons

Variational Bayes (VB)

Task: for some generic random variable, z, and data, y, find a good approximation to difficult to compute posterior distribution, $p(z|y)$.

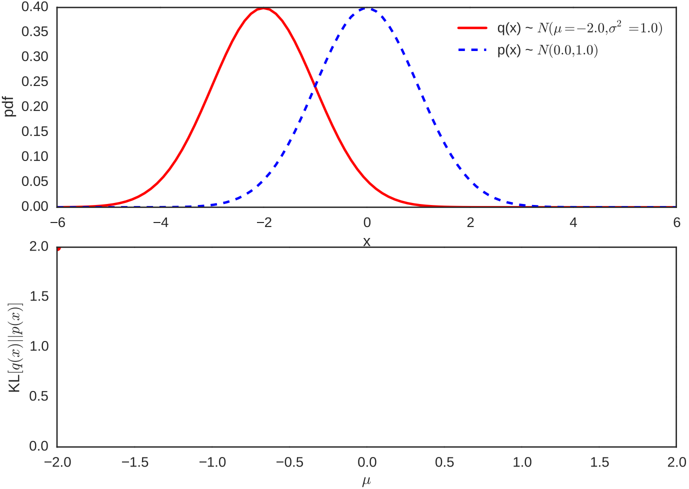
VB approach: minimise a divergence measure between an approximate posterior, $q(z)$ and true posterior, $p(z|y)$.

- KL divergence, $\mathrm{KL}\left(q(z) \,\|\, p(z|y)\right)$.
- Minimize this with respect to parameters of $q(z)$.

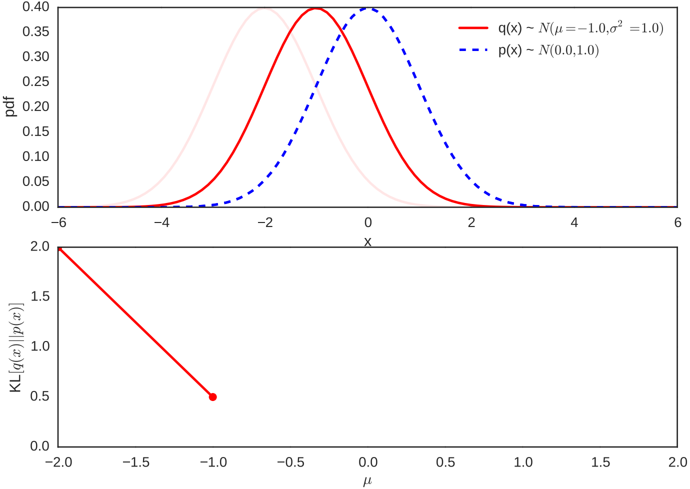
KL divergence

- General for any two distributions $q(\mathbf{x})$ and $p(\mathbf{x})$.
- $\mathrm{KL}\left(q(\mathbf{x})\,\|\,p(\mathbf{x})\right)$ is the average additional amount of information required to specify the values of $\mathbf{x}$ as a result of using an approximate distribution $q(\mathbf{x})$ instead of the true distribution, $p(\mathbf{x})$.
- $\mathrm{KL}\left(q(\mathbf{x})\,\|\,p(\mathbf{x})\right) = \left\langle \log \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\rangle_{q(\mathbf{x})}$
- Always 0 or positive, not symmetric.
- Lets look at how it changes with response to changes in the approximating distribution.

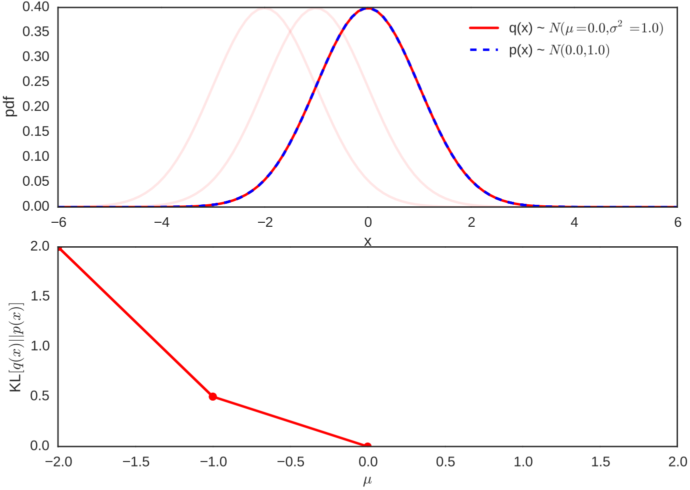
KL varying mean



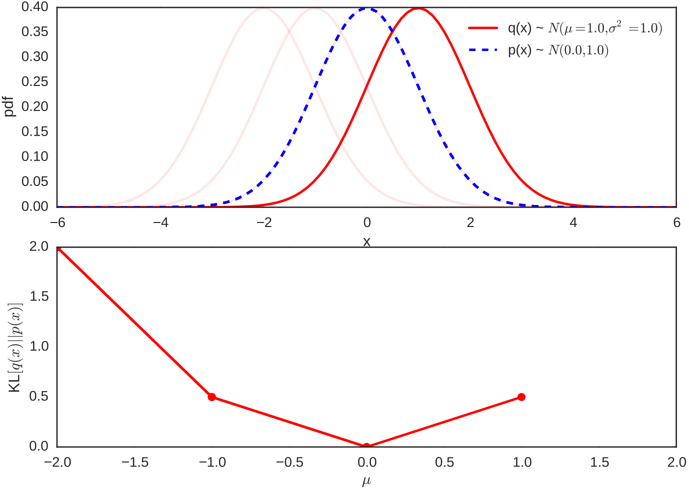
KL varying mean



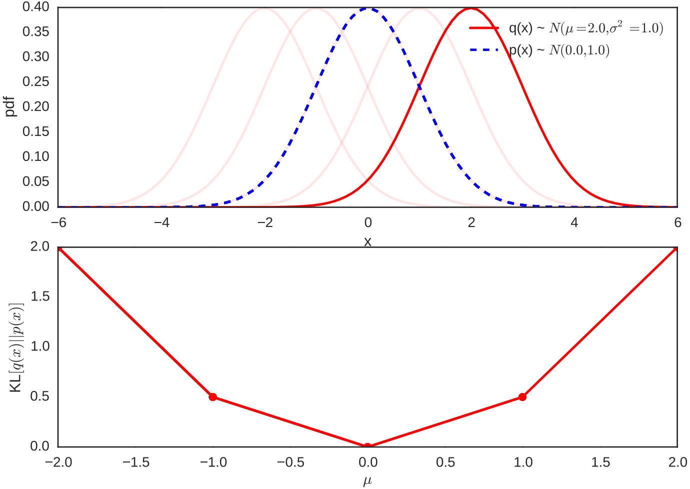
KL varying mean



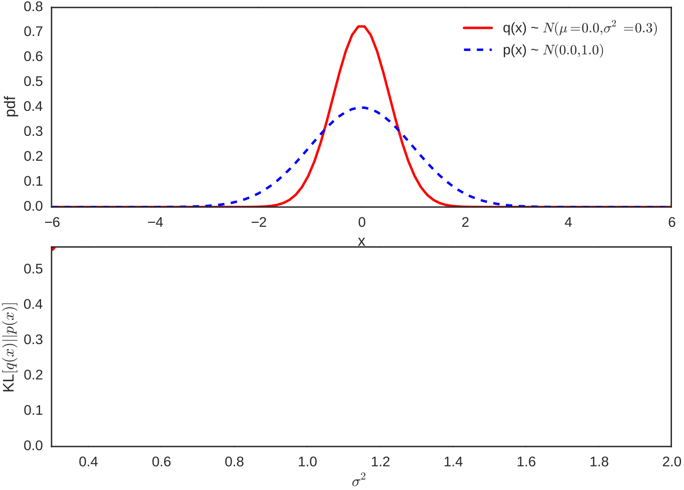
KL varying mean



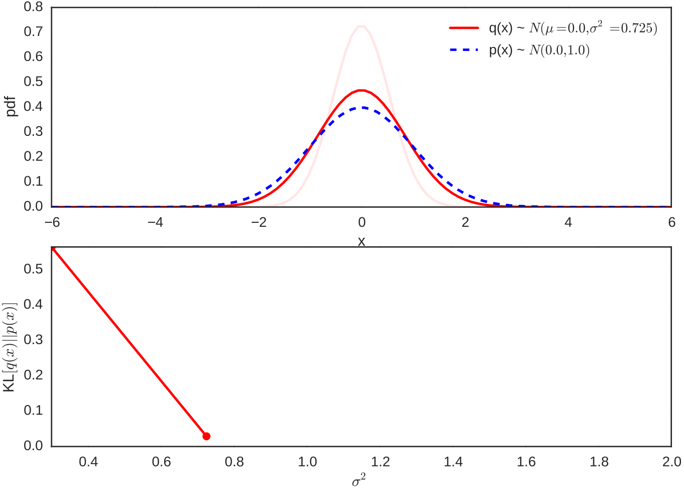
KL varying mean



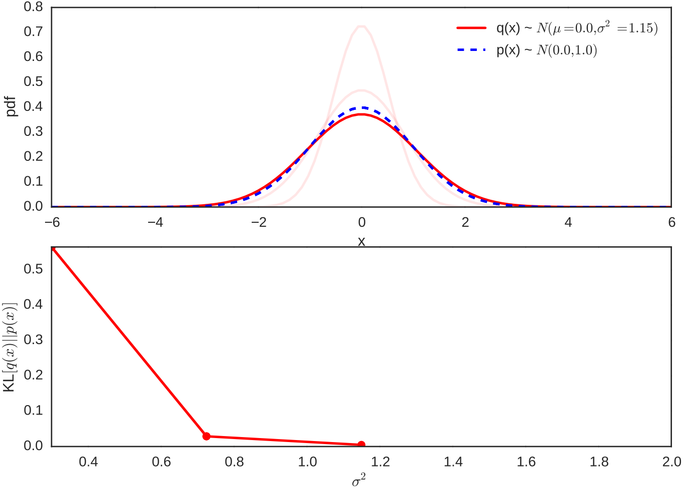
KL varying variance



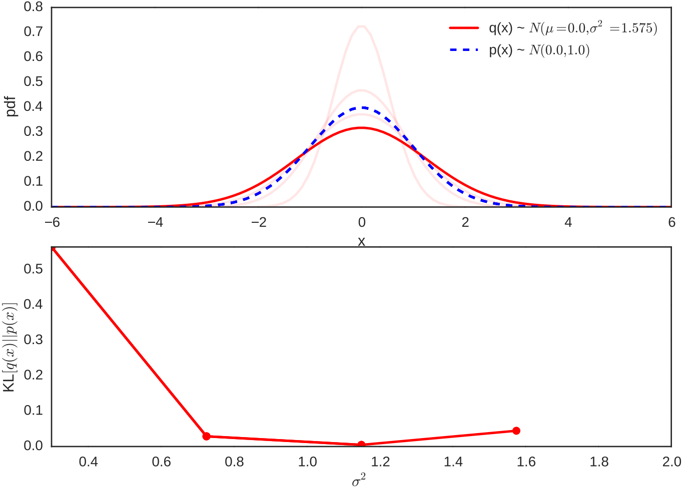
KL varying variance



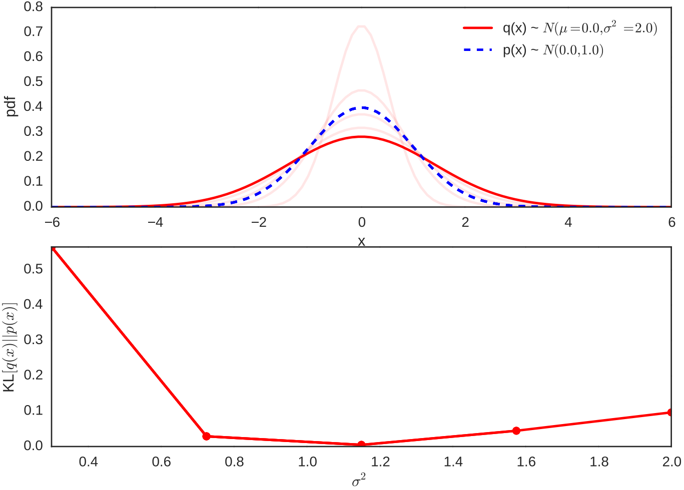
KL varying variance



KL varying variance



KL varying variance



Variational Bayes

Don't have access to or can't compute for computational reasons: $p(z|y)$ or $p(y)$, and hence $\mathrm{KL}\left(q(z) \,\|\, p(z|y)\right)$

How can we minimize something we can't compute?

- Can compute $q(z)$ and $p(y|z)$ for any z.
- $q(z)$ is parameterised by 'variational parameters'.
- True posterior using Bayes rule, $p(z|y) = \frac{p(y|z)p(z)}{p(y)}$.
- $p(y)$ doesn't change when variational parameters are changed.

Variational Bayes - Derivation

$$\mathrm{KL}\left(q(z)\,||\,p(z|y)\right)$$

Variational Bayes - Derivation

$$\mathrm{KL}\left(q(z) \,\|\, p(z|y)\right)$$

$$= \int q(z) \left[\log \frac{q(z)}{p(z|y)} \right] dz$$

Variational Bayes - Derivation

$$\mathrm{KL}\left(q(z)\,\|\,p(z|y)\right)$$
$$= \int q(z)\left[\log \frac{q(z)}{p(z|y)}\right]dz$$
$$= \int q(z)\left[\log \frac{q(z)}{p(z)} - \log p(y|z) + \log p(y)\right]dz$$

Variational Bayes - Derivation

$$
\begin{aligned}
&\mathrm{KL}\left(q(z) \,\|\, p(z|y)\right) \\
&= \int q(z) \left[\log \frac{q(z)}{p(z|y)}\right] dz \\
&= \int q(z) \left[\log \frac{q(z)}{p(z)} - \log p(y|z) + \log p(y)\right] dz \\
&= \mathrm{KL}\left(q(z) \,\|\, p(z)\right) - \int q(z) \left[\log p(y|z)\right] dz + \log p(y)
\end{aligned}
$$

Variational Bayes - Derivation

$$
\begin{aligned}
&\mathrm{KL}\left(q(z) \,\|\, p(z|y)\right) \\
&= \int q(z) \left[\log \frac{q(z)}{p(z|y)}\right] dz \\
&= \int q(z) \left[\log \frac{q(z)}{p(z)} - \log p(y|z) + \log p(y)\right] dz \\
&= \mathrm{KL}\left(q(z) \,\|\, p(z)\right) - \int q(z) \left[\log p(y|z)\right] dz + \log p(y)
\end{aligned}
$$

$$
\log p(y) = \int q(z) \left[\log p(y|z)\right] dz - \mathrm{KL}\left(q(z) \,\|\, p(z)\right) + \mathrm{KL}\left(q(z) \,\|\, p(z|y)\right)
$$

Variational Bayes - Derivation

$$\begin{aligned}
\log p(y) &= \int q(z) \left[\log p(y|z)\right] dz - \mathrm{KL}\left(q(z) \,\|\, p(z)\right) + \mathrm{KL}\left(q(z) \,\|\, p(z|y)\right) \\
&\geq \int q(z) \left[\log p(y|z)\right] dz - \mathrm{KL}\left(q(z) \,\|\, p(z)\right)
\end{aligned}$$

- Tractable terms give lower bound on $\log p(y)$ as $\mathrm{KL}\left(q(z) \,\|\, p(z|y)\right)$ always positive.
- Adjust variational parameters of $q(z)$ to make tractable terms as large as possible, thus $\mathrm{KL}\left(q(z) \,\|\, p(z|y)\right)$ as small as possible.

VB optimisation illustration

Variational Bayes for Gaussian processes

- Make a Gaussian approximation, $q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, \boldsymbol{C})$, as similar possible to true posterior, $p(\mathbf{f}|\mathbf{y})$.
- Treat $\boldsymbol{\mu}$ and $\boldsymbol{C}$ as 'variational parameters', effecting quality of approximation.
- True posterior using Bayes rule, $p(\mathbf{f}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f})}{p(\mathbf{y})}$.
- Cannot compute the KL divergence as we cannot compute the true posterior, $p(\mathbf{f}|\mathbf{y})$.

$$\begin{aligned}
\mathrm{KL}\left(q(\mathbf{f})\,\|\,p(\mathbf{f}|\mathbf{y})\right) &= \left\langle \log \frac{q(\mathbf{f})}{p(\mathbf{f}|\mathbf{y})} \right\rangle_{q(\mathbf{f})} \\
&= \left\langle \log \frac{q(\mathbf{f})}{p(\mathbf{f})} - \log p(\mathbf{y}|\mathbf{f}) + \log p(\mathbf{y}) \right\rangle_{q(\mathbf{f})} \\
&= \mathrm{KL}\left(q(\mathbf{f})\,\|\,p(\mathbf{f})\right) - \left\langle \log p(\mathbf{y}|\mathbf{f}) \right\rangle_{q(\mathbf{f})} + \log p(\mathbf{y}) \\
\log p(\mathbf{y}) &= \left\langle \log p(\mathbf{y}|\mathbf{f}) \right\rangle_{q(\mathbf{f})} - \mathrm{KL}\left(q(\mathbf{f})\,\|\,p(\mathbf{f})\right) + \mathrm{KL}\left(q(\mathbf{f})\,\|\,p(\mathbf{f}|\mathbf{y})\right)
\end{aligned}$$

Variational Bayes for Gaussian processes - bound

$$
\begin{aligned}
\log p(\mathbf{y}) &= \left\langle \log p(\mathbf{y}|\mathbf{f}) \right\rangle_{q(\mathbf{f})} - \mathrm{KL}\left(q(\mathbf{f}) \,\|\, p(\mathbf{f})\right) + \mathrm{KL}\left(q(\mathbf{f}) \,\|\, p(\mathbf{f}|\mathbf{y})\right) \\
&\geq \left\langle \log p(\mathbf{y}|\mathbf{f}) \right\rangle_{q(\mathbf{f})} - \mathrm{KL}\left(q(\mathbf{f}) \,\|\, p(\mathbf{f})\right)
\end{aligned}
$$

- Adjust variational parameters $\boldsymbol{\mu}$ and $\boldsymbol{C}$ to make tractable terms as large as possible, thus $\mathrm{KL}\left(q(\mathbf{f}) \,\|\, p(\mathbf{f}|\mathbf{y})\right)$ as small as possible.
- $\left\langle \log p(\mathbf{y}|\mathbf{f}) \right\rangle_{q(\mathbf{f})}$ with factorizing likelihood can be done with a series of n 1 dimensional integrals.
- In practice, can reduce the number of variational parameters by reparameterizing $\boldsymbol{C} = (\mathbf{K}_{\mathbf{ff}} - 2\Lambda)^{-1}$ by noting that the bound is constant in off diagonal terms of $\boldsymbol{C}$.

VB optimisation illustration for Gaussian processes

Outline

Motivation

Non-Gaussian posteriors

Approximate methods

- Laplace approximation
- Variational bayes
- Expectation propagation
- Comparisons

Expectation propagation

$$p(\mathbf{f}|\mathbf{y}) \propto p(\mathbf{f}) \prod_{i=1}^{n} p(\mathbf{y}_i|\mathbf{f}_i)$$

$$q(\mathbf{f}|\mathbf{y}) \triangleq \frac{1}{Z_{ep}} p(\mathbf{f}) \prod_{i=1}^{n} t_i(\mathbf{f}_i|\tilde{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2) = \mathcal{N}\left(\mathbf{f}|\boldsymbol{\mu}, \Sigma\right)$$

$$t_i \triangleq \tilde{Z}_i \mathcal{N}\left(\mathbf{f}_i|\tilde{\boldsymbol{\mu}}_i, \tilde{\boldsymbol{\sigma}}_i^2\right)$$

- Individual likelihood terms, $p(\mathbf{y}_i|\mathbf{f}_i)$, replaced by independent local likelihood functions, t_i.
- Uses an iterative algorithm to update t_i's.

Expectation propagation

1. From the approximate current posterior, $q(\mathbf{f}|\mathbf{y})$, leave out one of the local likelihoods, t_i, and marginalise $\mathbf{f}_j$ where $j \neq i$, giving rise to the marginal *cavity distribution*, $q_{-i}(\mathbf{f}_i)$.
2. Combine resulting cavity distribution, $q_{-i}(\mathbf{f}_i)$, with exact likelihood contribution, $p(\mathbf{y}_i|\mathbf{f}_i)$, giving non-Gaussian un-normalized distribution, $\hat{q}(\mathbf{f}_i) \triangleq p(\mathbf{y}_i|\mathbf{f}_i)q_{-i}(\mathbf{f}_i)$.
3. Choose a un-normalized Gaussian approximation to this distribution, $\mathcal{N}\left(\mathbf{f}_i|\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\sigma}}_i^2\right)\hat{Z}_i$, by finding moments of $\hat{q}(\mathbf{f}_i)$.
4. Replace parameters of t_i with those that produce the same moments as this approximation.
5. Choose another i and start again. Repeat to convergence.

Expectation propagation - in math

Step 1. First choose a local likelihood contribution, i, to leave out, and find the marginal cavity distribution,

$$q(\mathbf{f}|\mathbf{y}) \propto p(\mathbf{f}) \prod_{j=1}^{n} t_j(\mathbf{f}_j) \rightarrow \frac{p(\mathbf{f}) \prod_{j=1}^{n} t_j(\mathbf{f}_j)}{t_i(\mathbf{f}_i)} \rightarrow p(\mathbf{f}) \prod_{j \neq i}^{n} t_j(\mathbf{f}_j)$$

$$\rightarrow \int p(\mathbf{f}) \prod_{j \neq i} t_j(\mathbf{f}_j)\, d\mathbf{f}_{j \neq i} \triangleq q_{-i}(\mathbf{f}_i)$$

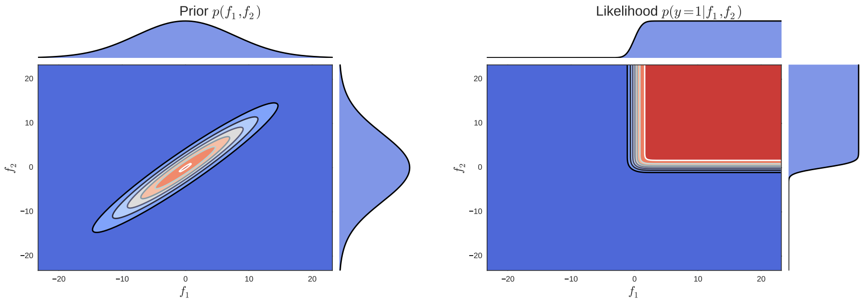
Step 2. $p(\mathbf{y}_i|\mathbf{f}_i) q_{-i}(\mathbf{f}_i) \triangleq \hat{q}(\mathbf{f}_i)$

Step 3. $\hat{q}(\mathbf{f}_i) \approx \mathcal{N}\left(\mathbf{f}_i | \hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\sigma}}_i^2\right) \hat{Z}_i$

Step 4: Compute parameters of $t_i(\mathbf{f}_i|\tilde{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2)$ making moments of $q(\mathbf{f}_i)$ match those of $\hat{Z}_i \mathcal{N}\left(\mathbf{f}_i | \hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\sigma}}_i^2\right)$.

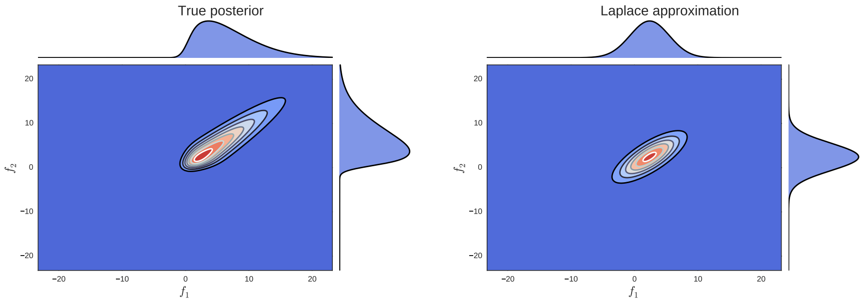
Outline

Comparing posterior approximations



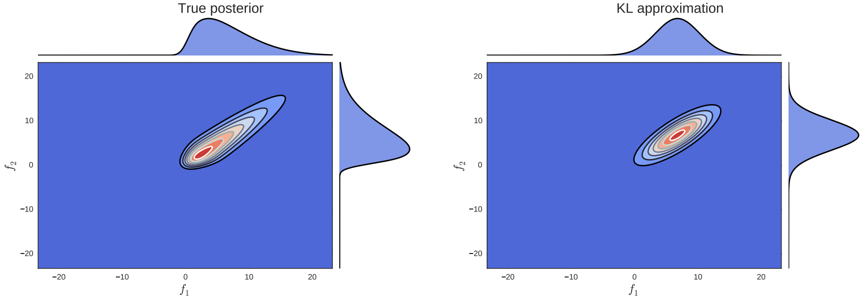
- Gaussian prior between two function values $\{\mathbf{f}_1, \mathbf{f}_2\}$, at $\{\mathbf{x}_1, \mathbf{x}_2\}$ respectively.
- Bernoulli likelihood, $\mathbf{y}_1 = 1$ and $\mathbf{y}_2 = 1$.

Comparing posterior approximations



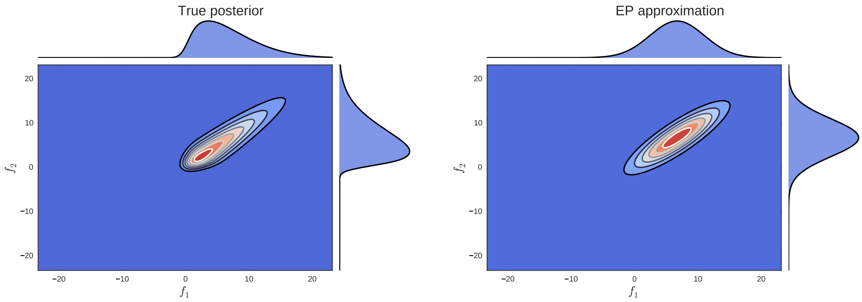
- $p(\mathbf{f}|\mathbf{y}) \propto \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f})}{p(\mathbf{y})}$
- True posterior is non-Gaussian.
- Laplace approximates with a Gaussian at the mode of the posterior.

Comparing posterior approximations



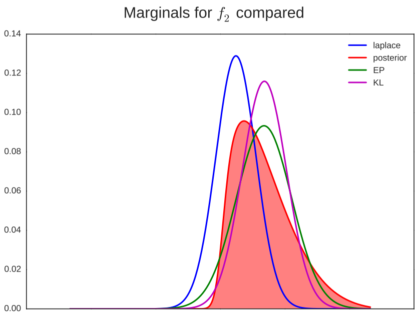
- True posterior is non-Gaussian.
- VB approximate with a Gaussian that has minimal KL divergence, $\mathrm{KL}\left(q(\mathbf{f})\,\|\,p(\mathbf{f}|\mathbf{y})\right)$.
- This leads to distributions that avoid regions in which $p(\mathbf{f}|\mathbf{y})$ is small.
- It has a large penality for assigning density where there is none.

Comparing posterior approximations



- True posterior is non-Gaussian.
- EP tends to try and put density where $p(\mathbf{f}|\mathbf{y})$ is large
- Cares less about assigning density density where there is none. Contrasts to VB method.

Comparing posterior marginal approximations



- Laplace: Poor approximation.
- VB: Avoids assigning density to areas where there is none, at the expense of areas where there is some (right tail).
- EP: Assigns density to areas with density, at the expense of areas where there is none (left tail).

Pros - Cons - When - Laplace

Laplace approximation

- Pros
  - Very fast.
- Cons
  - Poor approximation if the mode does not well describe the posterior, for example Bernoulli likelihood.
- When
  - When the posterior *is* well characterized by its mode, for example Poisson.

Pros - Cons - When - VB

Variational Bayes

- Pros
  - Principled in that it we are directly optimizing a measure of divergence between an approximation and true distribution.
  - Lends itself to sparse extensions.
- Cons
  - Requires factorizing likelihoods to avoid n dimensional integral.
  - As seen, can result in underestimating the variance, i.e. becomes overconfident.
- When
  - Applicable to a range of likelihoods, but is known in some cases to underestimate variance, might need to be careful if you wish to be conservative with predictive uncertainty.
  - In conjunction with sparse methods.

Pros - Cons - When - EP

EP method

- Pros
  - Very effective for certain likelihoods (classification).
  - Also lends itself to sparse approximations.
- Cons
  - Standard algorithm is slow though possible to extend to sparse case.
  - Convergence issues for some likelihoods.
  - Must be able to match moments.
- When
  - Binary data (Nickisch and Rasmussen, 2008; Kuß, 2006), perhaps with truncated likelihood (censored data) (Vanhatalo et al., 2015).
  - In conjunction with sparse methods.

Pros - Cons - When - MCMC

MCMC methods

- Pros
  - Theoretical limit gives true distribution
- Cons
  - Can be very slow
- When
  - If time is not an issue, but exact accuracy is.
  - If you are unsure whether a different approximation is appropriate, can be used as a "ground truth"

Conclusion

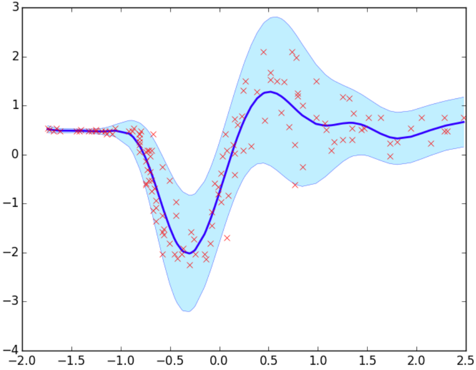
- Many real world tasks require non-Gaussian observation models.
- Non-Gaussian likelihoods cause complications in applying our framework.
- Several different ways to deal with the problem. Many are based on Gaussian approximations.
- Different methods have their own advantages and disadvantages.

Questions

Thanks for listening.

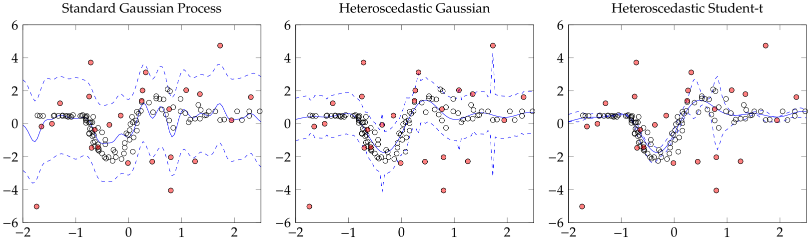
Any questions?

Bonus - Hetroscedastic likelihoods



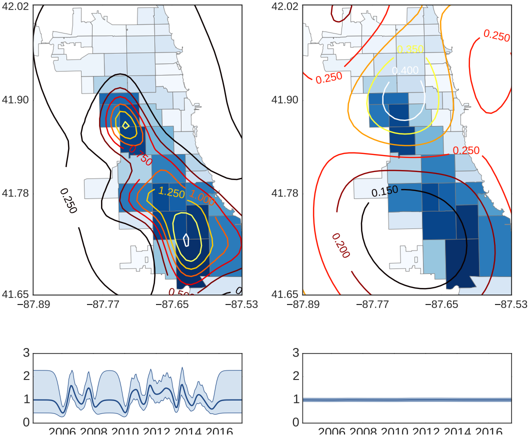
- Likelihood whos parameters are governed by two known functions, $\mathbf{f}$ and $\mathbf{g}$.
- $p(\mathbf{y}|\mathbf{f}, \mathbf{g}) = \mathcal{N}\left(\mathbf{y}|\mu = \mathbf{f}, \sigma^2 = \exp(\mathbf{g})\right)$

Bonus - non-Gaussian hetroscedastic likelihoods



- Likelihood whos parameters are governed by two known functions, $\mathbf{f}$ and $\mathbf{g}$.
- $p(\mathbf{y}|\mathbf{f}, \mathbf{g}) = t(\mathbf{y}|\mu = \mathbf{f}, \sigma^2 = \exp(\mathbf{g}), \nu = 3.0)$

Bonus - non-Gaussian hetroscedastic likelihoods



- $\Lambda(\mathbf{x}, \mathbf{t}) = \lambda_1(\mathbf{x})\mu_1(\mathbf{t}) + \lambda_2(\mathbf{x})\mu_2(\mathbf{t})$

References I

Hensman, J., Matthews, A. G. D. G., and Ghahramani, Z. (2015). Scalable variational gaussian process classification. In *In 18th International Conference on Artificial Intelligence and Statistics*, pages 1–9, San Diego, California, USA.

Kuß, M. (2006). *Gaussian Process Models for Robust Regression, Classification, and Reinforcement Learning*. PhD thesis, TU Darmstadt.

Nickisch, H. and Rasmussen, C. E. (2008). Approximations for Binary Gaussian Process Classification. *Journal of Machine Learning Research*, 9:2035–2078.

Vanhatalo, J., Riihimaki, J., Hartikainen, J., Jylanki, P., Tolvanen, V., and Vehtari, A. (2015). Gpstuff. `http://mloss.org/software/view/451/`.