Non-Gaussian likelihoods for Gaussian Processes

Alan Saul

University of Sheffield

Outline

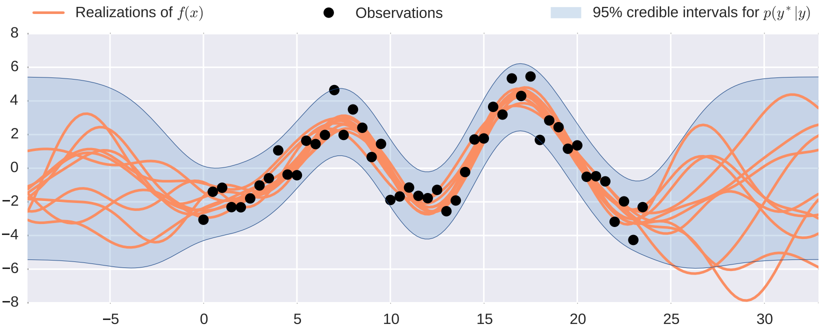
- Motivation
- Laplace approximation
- KL method
- Expectation Propagation
- Comparing approximations

GP regression

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

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

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.

Gaussian Prior: $\mathbf{f} \sim \mathcal{GP}(\mathbf{0}, \mathbf{K}_{\mathbf{ff}}) = p(\mathbf{f})$

Gaussian likelihood: $\mathbf{y} \sim \mathcal{N}\left(\mathbf{f}, \sigma^2\mathbf{I}\right) = \prod_{i=1}^{n} p(\mathbf{y}_i|\mathbf{f}_i)$

Gaussian posterior: $p(\mathbf{f}|\mathbf{y}) \propto \mathcal{N}\left(\mathbf{y}|\mathbf{f}, \sigma^2\mathbf{I}\right)\mathcal{N}\left(\mathbf{f}|\mathbf{0}, \mathbf{K}_{\mathbf{ff}}\right)$

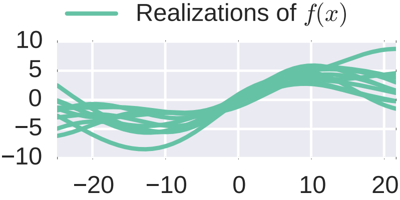
Likelihood

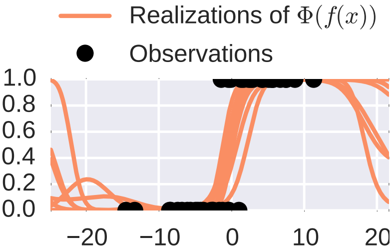
- $p(\mathbf{y}|\mathbf{f})$ is the probability that we would see some random variables, $\mathbf{y}$, if we knew the latent function values $\mathbf{f}$, which act as parameters.
- Given the observed values for $\mathbf{y}$ are fixed, it can also be seen as the likelihood that some latent function values, $\mathbf{f}$, 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 $\mathbf{f}$.
- So far assumed that the distortion of the underlying latent function, $\mathbf{f}$, that gives rise to the observed data, $\mathbf{y}$, is independent and normally distributed.
- This is often not the case, count data, binary data, etc.

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, $\lambda(\mathbf{f}_i) = \Phi(\mathbf{f}_i)$.

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

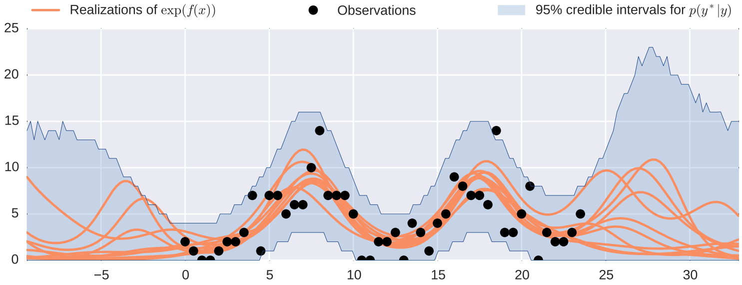




Count data example

- 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 $\lambda(\mathbf{f}_i) = \mathbf{f}_i^2$

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



Non-Gaussian posteriors

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

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

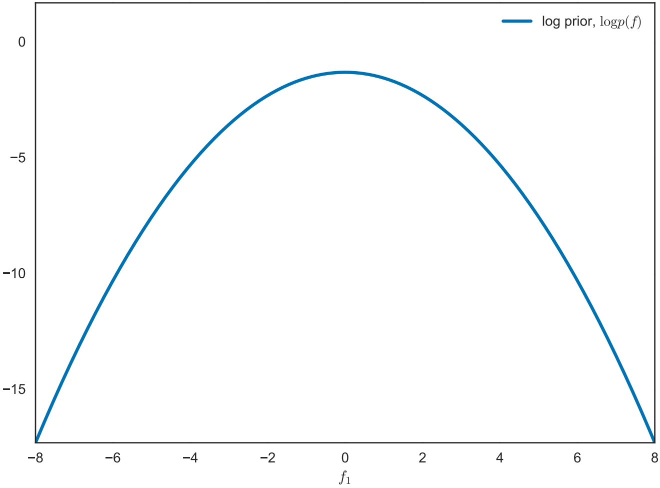
- Various methods to make a Gaussian approximation, $q(\mathbf{f}) \approx p(\mathbf{f}|\mathbf{y})$.
- Allows simple predictions

$$\begin{aligned} p(\mathbf{f}^*|\mathbf{y}) &= \int p(\mathbf{f}^*|\mathbf{f}) p(\mathbf{f}|\mathbf{y}) d\mathbf{f} \\ &\approx \int p(\mathbf{f}^*|\mathbf{f}) q(\mathbf{f}) d\mathbf{f} \end{aligned}$$

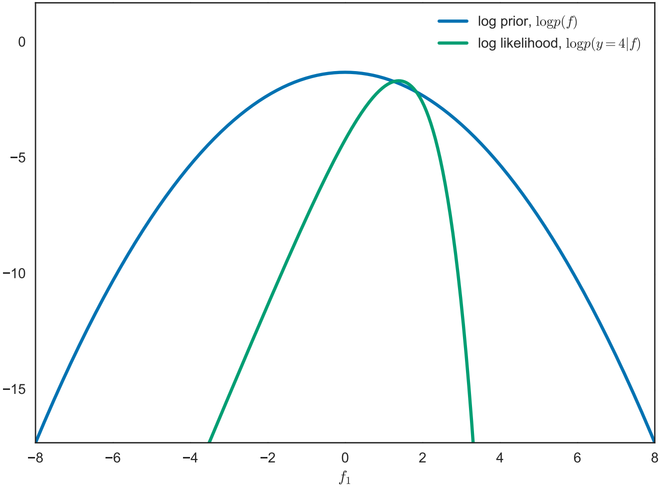
Laplace approximation

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

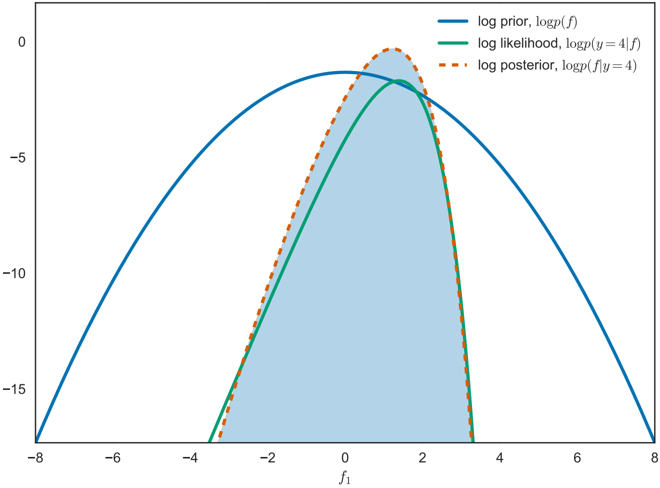
Visualization of Laplace



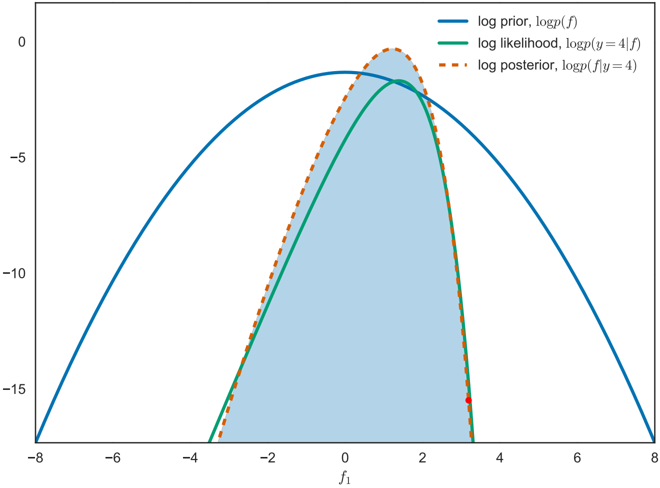
Visualization of Laplace



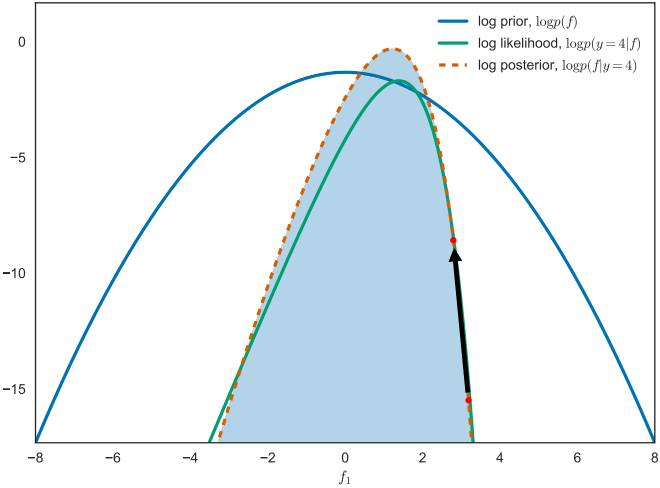
Visualization of Laplace



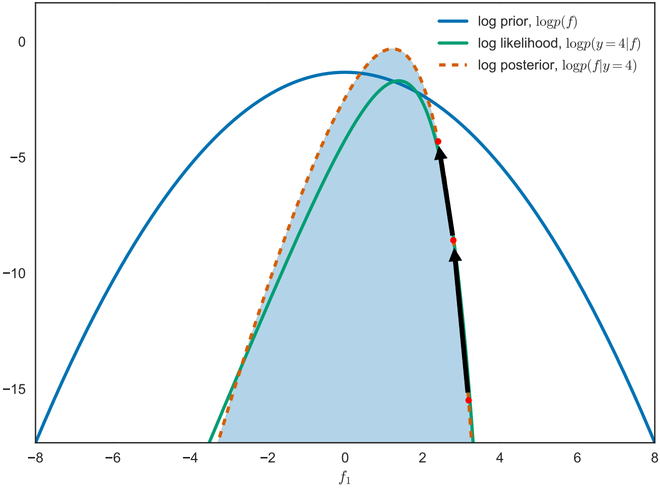
Visualization of Laplace



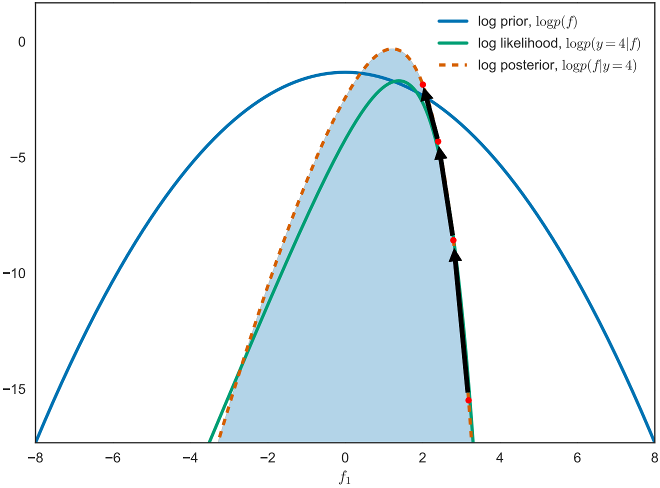
Visualization of Laplace



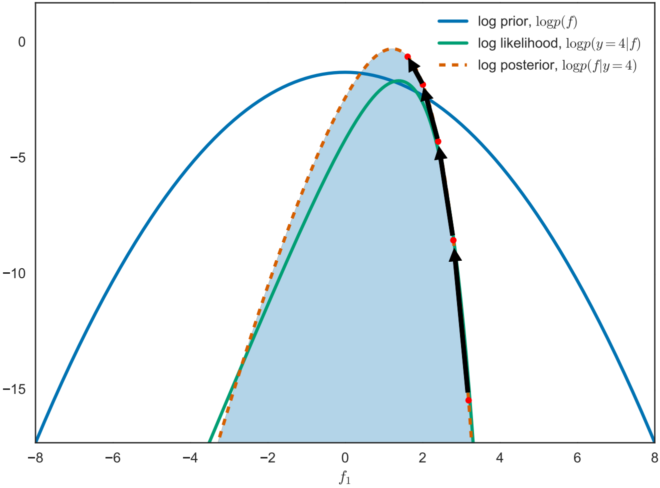
Visualization of Laplace



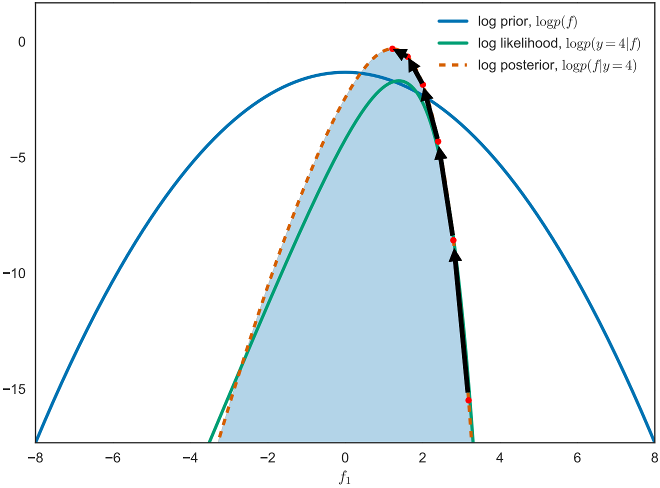
Visualization of Laplace



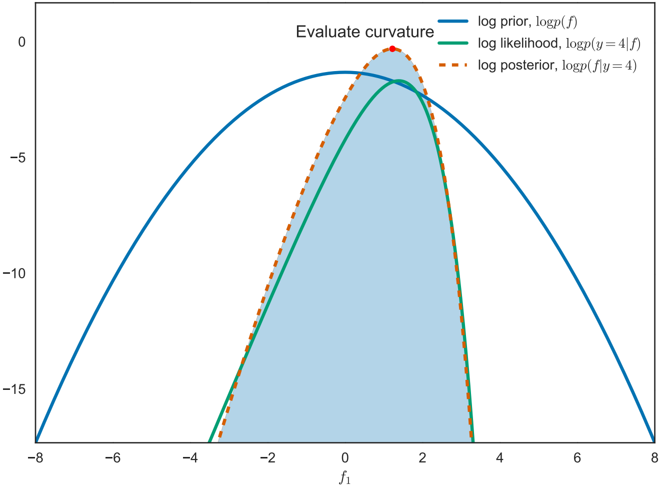
Visualization of Laplace



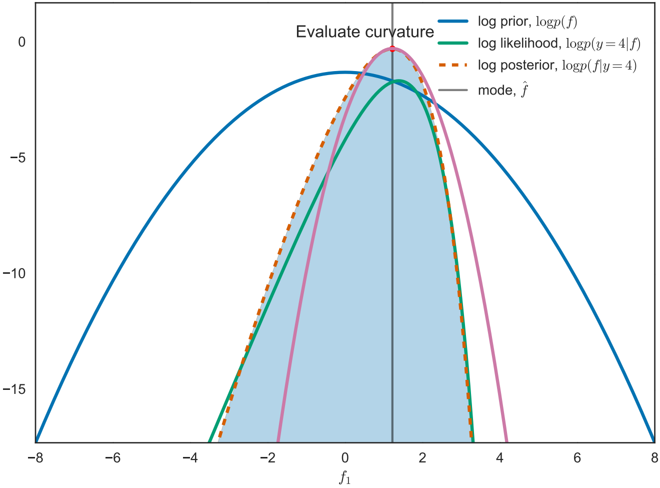
Visualization of Laplace



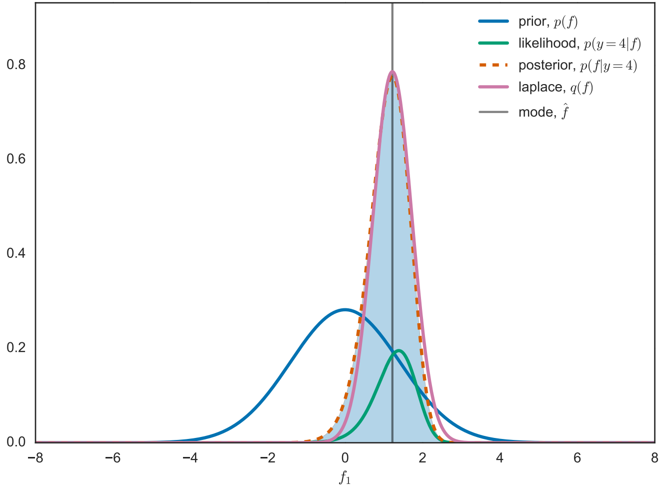
Visualization of Laplace



Visualization of Laplace



Visualization of Laplace



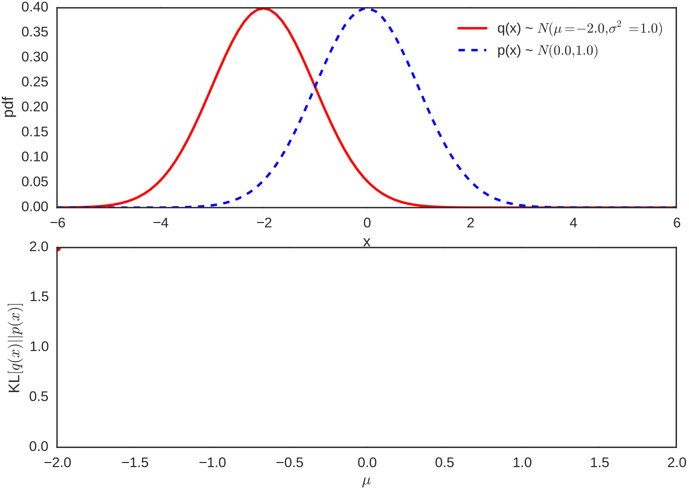
KL-method

- Make a Gaussian approximation, $q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, C)$, as similar possible to true posterior, $p(\mathbf{f}|\mathbf{y})$.
- Treat $\boldsymbol{\mu}$ and C as variational parameters, effecting quality of approximation.
- Define a divergence measure between two distributions, KL divergence, KL $(q(\mathbf{f}) \,\|\, p(\mathbf{f}|\mathbf{y}))$.
- Minimize this divergence between the two distributions (Nickisch and Rasmussen, 2008).

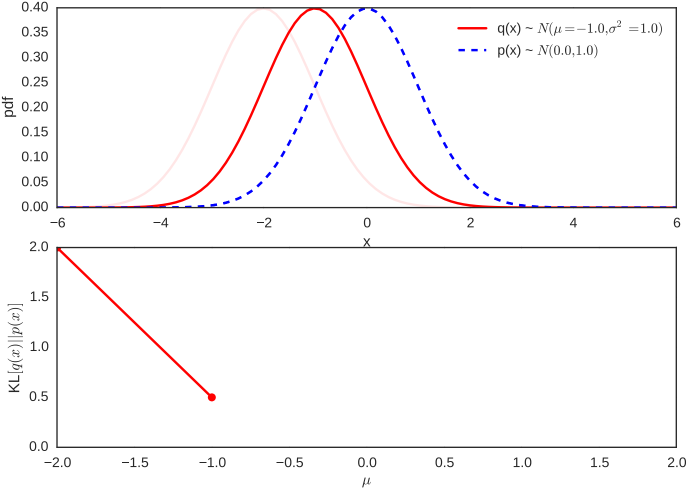
KL divergence

- General for any two distributions $q(\mathbf{x})$ and $p(\mathbf{x})$.
- $\text{KL}\left(q(\mathbf{x})\,\|\,p(\mathbf{x})\right)$ 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, $p(\mathbf{x})$.
- $\text{KL}\left(q(\mathbf{x})\,\|\,p(\mathbf{x})\right) = \left\langle \log \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\rangle_{q(\mathbf{x})}$
- Always 0 or positive, not symmetric.
- Lets look at how it changes with response to changes in the approximating distribution.

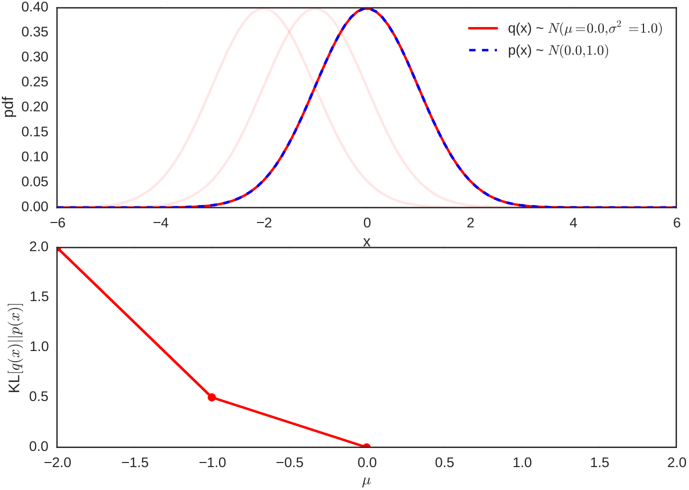
KL varying mean



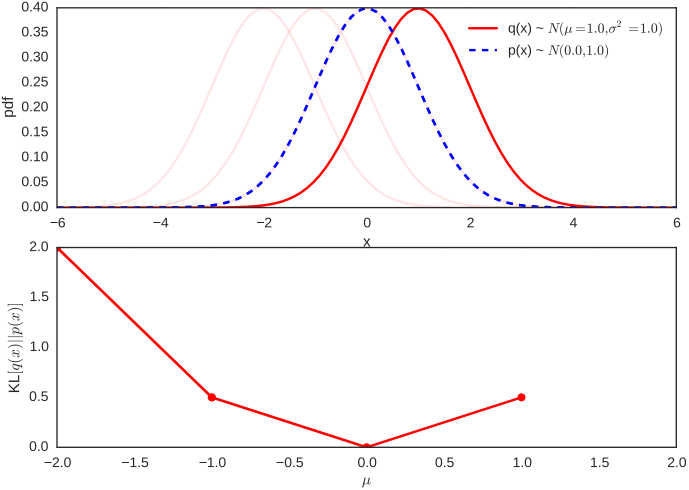
KL varying mean



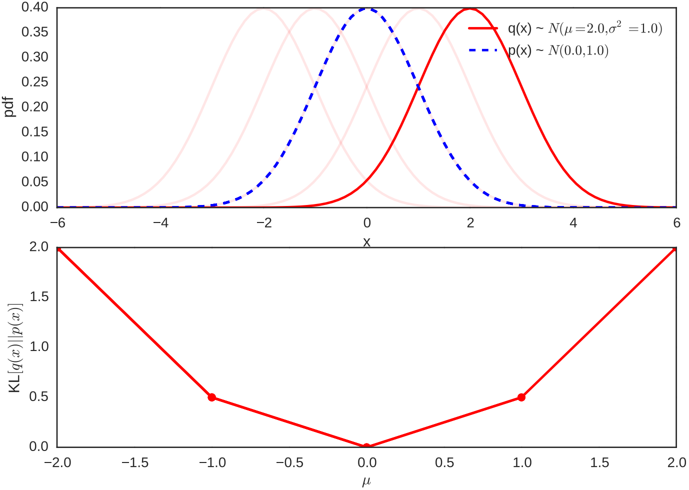
KL varying mean



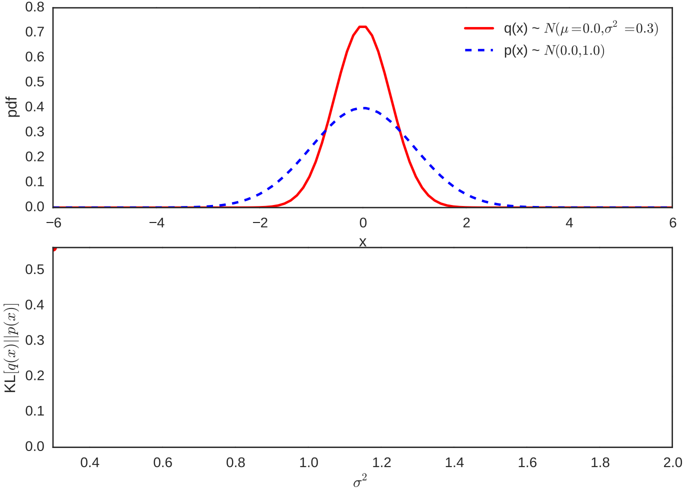
KL varying mean



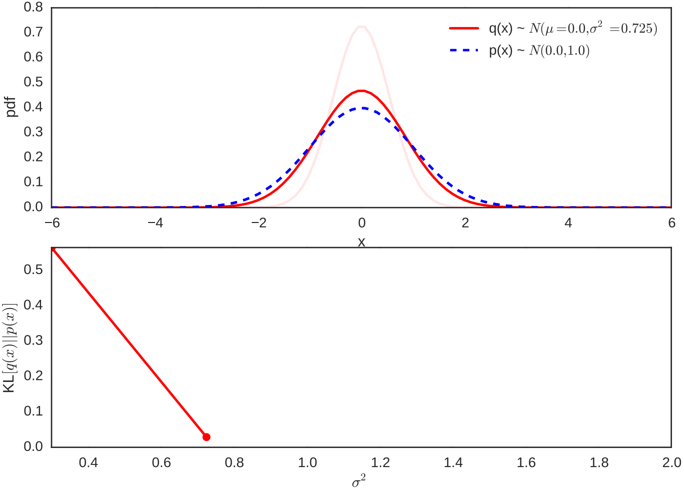
KL varying mean



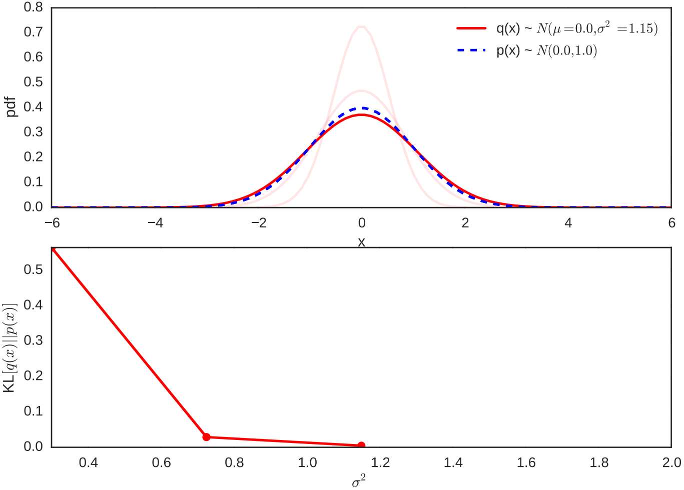
KL varying variance



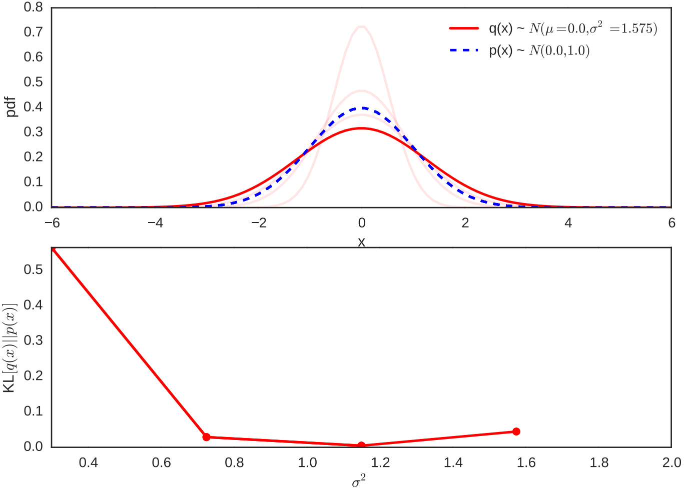
KL varying variance



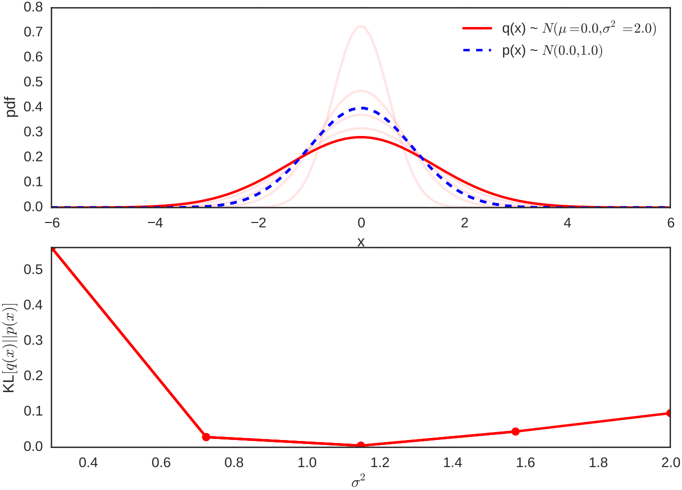
KL varying variance



KL varying variance



KL varying variance



KL-method derivation

- Assume Gaussian approximate posterior, $q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, C)$.
- 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, $p(\mathbf{f}|\mathbf{y})$.

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

KL-method derivation

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

- Tractable terms give lower bound on $\log p(\mathbf{y})$ as $\mathrm{KL}\left(q(\mathbf{f}) \,\|\, p(\mathbf{f}|\mathbf{y})\right)$ always positive.
- Adjust variational parameters $\boldsymbol{\mu}$ and $\mathbf{C}$ 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.
- $\langle \log p(\mathbf{y}|\mathbf{f}) \rangle_{q(\mathbf{f})}$ 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 $\mathbf{C} = (\mathbf{K}_{\mathbf{ff}} - 2\Lambda)^{-1}$ by noting that the bound is constant in off diagonal terms of $\mathbf{C}$.

Expectation Propagation

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

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

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

- 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,

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

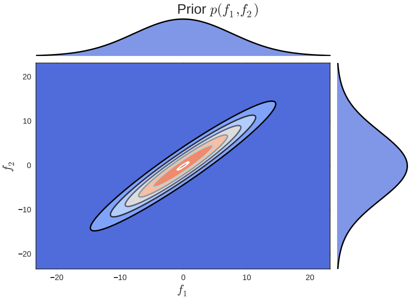
$$\rightarrow \int p(\mathbf{f}) \prod_{j \neq i} t_j(\mathbf{f}_j)\, d\mathbf{f}_{j \neq i} \triangleq q_{-i}(\mathbf{f}_i)$$

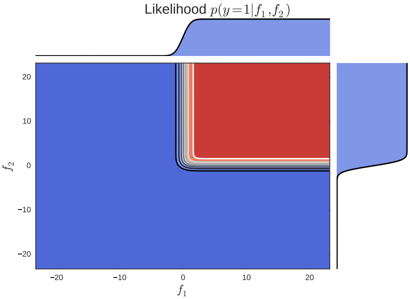
Step 2. $p(\mathbf{y}_i|\mathbf{f}_i) q_{-i}(\mathbf{f}_i) \triangleq \hat{q}(\mathbf{f}_i)$

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

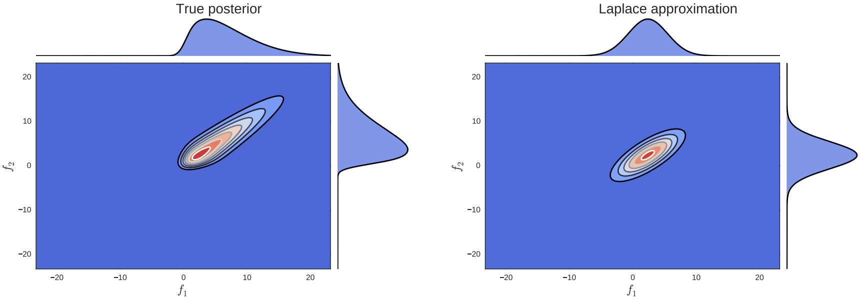
Comparing posterior approximations





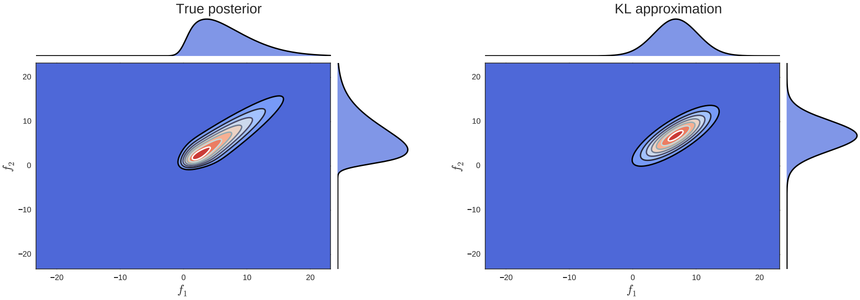
- Gaussian prior between two function values $\{\mathbf{f}_1, \mathbf{f}_2\}$, at $\{\mathbf{x}_1, \mathbf{x}_2\}$ respectively.
- Bernoulli likelihood, $\mathbf{y}_1 = 1$ and $\mathbf{y}_2 = 1$.

Comparing posterior approximations



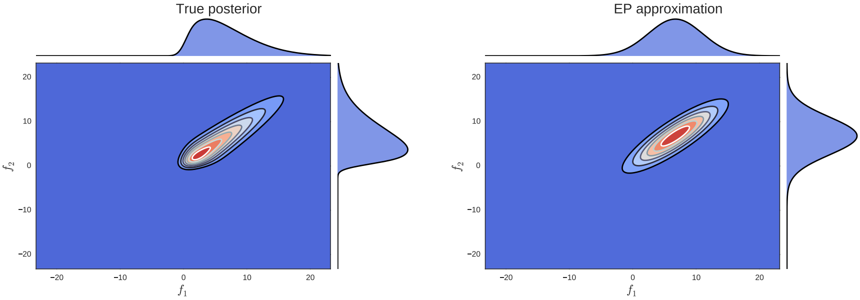
- $p(\mathbf{f}|\mathbf{y}) \propto \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f})}{p(\mathbf{y})}$
- True posterior is non-Gaussian.
- Laplace approximates with a Gaussian at the mode of the posterior.

Comparing posterior approximations



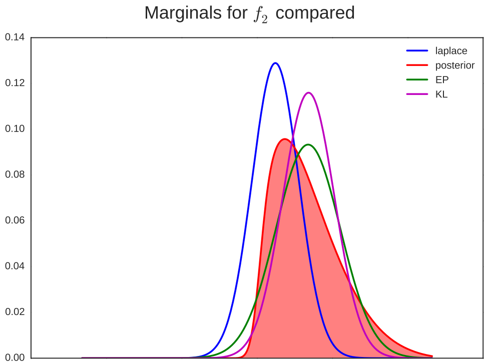
- True posterior is non-Gaussian.
- KL approximate with a Gaussian that has minimal KL divergence, KL $(q(\mathbf{f}) \,||\, p(\mathbf{f}|\mathbf{y}))$.
- This leads to distributions that avoid regions in which $p(\mathbf{f}|\mathbf{y})$ is small.
- It has a large penality for assigning density where 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 KL method.

Comparing posterior marginal approximations



- Laplace: Poor approximation.
- KL: 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 - KL

KL method

- 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”

Questions

Thanks for listening.

Any questions?

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/`.