MOGPTK notes

Taco de Wolff

December 11, 2023

1 Introduction

1.1 Notation

We state that

$$y = f(\boldsymbol{x}) + \epsilon$$

with

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

where

$$f(\boldsymbol{x}) \sim \mathcal{GP}(m(\boldsymbol{x}), K(\boldsymbol{x}, \boldsymbol{x}'))$$

We write $K_{ab} = K(a, b)$ and (for inducing point models) $Q_{ab} = K_{au}K_{uu}^{-1}K_{ub}$ with u the inducing points.

1.2 Bayes' theorem

Bayes' theorem states that

$$p(\boldsymbol{f}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f})}{p(\boldsymbol{y})}$$

with $p(\boldsymbol{y}|\boldsymbol{f})$ the likelihood and $p(\boldsymbol{y}$ the evidence or marginal likelihood.

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f})df \tag{1}$$

1.3 Matrix inversion lemma

$$(Z + UWV^T)^{-1} = Z^{-1} - Z^{-1}U(W^{-1} + V^T Z^{-1}U)^{-1}V^T Z^{-1}$$

1.4 Matrix determinant lemma

$$|Z + UWV^T| = |W^{-1} + V^T Z^{-1} U| \, |W| \, |Z|$$

1.5 Gaussian linear transformation

$$\begin{aligned} &\int \mathcal{N}(\boldsymbol{y}|A\boldsymbol{z}, B) \cdot \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \Sigma) d\boldsymbol{z} \\ &= \int \mathcal{N}(\boldsymbol{y}|A\boldsymbol{\mu}, A\Sigma A^T + B) \cdot \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \Sigma) d\boldsymbol{z} \\ &= \mathcal{N}(\boldsymbol{y}|A\boldsymbol{\mu}, A\Sigma A^T + B) \end{aligned} \tag{2}$$

1.6 Marginal Gaussian properties

Given the joint Gaussian distribution

$$\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} A & C \\ C^T & B \end{bmatrix} \right) \tag{3}$$

we have

$$\begin{aligned} \boldsymbol{x} &\sim \mathcal{N}(\boldsymbol{\mu}_x, A) \\ \boldsymbol{x}|\boldsymbol{y} &\sim \mathcal{N}(\boldsymbol{\mu}_x + CB^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y), A - CB^{-1}C^T) \\ \boldsymbol{y}|\boldsymbol{x} &\sim \mathcal{N}(\boldsymbol{\mu}_y + C^T A^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_x), B - C^T A^{-1} C) \end{aligned} \tag{4}$$

As stated by Bishop page 93, this is equivalent to given

$$\begin{aligned} \boldsymbol{x} &\sim \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_x, A) \\ \boldsymbol{y}|\boldsymbol{x} &\sim \mathcal{N}(\boldsymbol{y}|Z\boldsymbol{x} + \boldsymbol{z}, L) \end{aligned} \tag{5}$$

then

$$\begin{aligned} \boldsymbol{y} &\sim \mathcal{N}(\boldsymbol{y}|Z\boldsymbol{\mu}_x + \boldsymbol{z}, L + ZAZ^T) \\ \boldsymbol{x}|\boldsymbol{y} &\sim \mathcal{N}(\boldsymbol{x}|\Sigma(Z^T L^{-1}(\boldsymbol{y} - \boldsymbol{z}) + A^{-1}\boldsymbol{\mu}_x), \Sigma) \end{aligned} \tag{6}$$

with

$$\Sigma = (A^{-1} + Z^T L^{-1} Z)^{-1}$$

1.7 Kullback-Leibler

For $\mathrm{KL}(Q \,||\, P)$ where Q and P are both Gaussian

$$Q \sim \mathcal{N}(\boldsymbol{\mu}_q, \Sigma_q) \qquad P \sim \mathcal{N}(\boldsymbol{\mu}_p, \Sigma_p)$$

the Kullback-Leibler divergences reduces down to

$$\mathrm{KL}(Q \mid\mid P) = \frac{1}{2}\left(\mathrm{Tr}(\Sigma_p^{-1}\Sigma_q) + (\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)^T\Sigma_p^{-1}(\boldsymbol{\mu}_p - \boldsymbol{\mu}_q) - k + \log\frac{|\Sigma_p|}{|\Sigma_q|}\right) \tag{7}$$

where k is the dimension of the distributions.

Note that if we have $L_pL_p^T = \Sigma_p$ and $L_qL_q^T = \Sigma_q$, then we can rewrite the trace as

$$\mathrm{Tr}(\Sigma_p^{-1}\Sigma_q) = \mathrm{Tr}\left((L_p^{-1}L_q)^{\circ 2}\right)$$

where $A^{\circ 2}$ is the element-wise square of matrix A. Also note that $|\Sigma_p| = |L_pL_p^T| = |L_p|\,|L_p^T| = |L_p|^2$.

1.8 Gaussian quadratures

Using Gauss-Hermite quadratures, we can approximate infinite integrals as sums of m terms, where m can be chosen. Higher m will be a more accurate approximation but more costly to calculate. We can state that

$$\int g(x)e^{-x^2}dx \approx \sum_{j=1}^{m} w_j g(t_j)$$

where position $\boldsymbol{t}$ and weight $\boldsymbol{w}$ are specific to an nth-degree quadrature.

1.9 Integrals

$$\int_{-\infty}^{\infty} xe^{-ax^2}dx = 0$$

$$\int_{-\infty}^{\infty} e^{-(ax^2+bx+c)}dx = \sqrt{\frac{\pi}{a}}e^{\frac{b^2}{4a}-c}$$

$$\int_{0}^{\infty} x^n e^{-ax^2}dx = \frac{\Gamma\left(\frac{n+1}{2}\right)}{2\left(a^{\frac{n+1}{2}}\right)}$$

$$\int_{0}^{\infty} xe^{-ax^2}dx = \frac{1}{2a}$$

2 Exact

Using a Gaussian likelihood and a prior

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{f} &\sim \mathcal{N}(\boldsymbol{f},\ \sigma^2 I) \\ \boldsymbol{f} &\sim \mathcal{N}(\mathbf{0},\ K_{ff}), \end{aligned} \tag{8}$$

the marginal likelihood of Eq. 1 is tractable. Using Eq. 4 we obtain

$$\boldsymbol{y} \sim \mathcal{N}(\mathbf{0}, K_{ff} + \sigma^2 I)$$

Objective Maximize log marginal likelihood:

$$\log p(\boldsymbol{y}) = -\frac{1}{2}\boldsymbol{y}^T(K_{ff} + \sigma^2 I)^{-1}\boldsymbol{y} - \frac{1}{2}\log|K_{ff} + \sigma^2 I| - \frac{n}{2}\log 2\pi$$

Prediction Given the noisy joint distribution

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{f_*} \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K_{ff} + \sigma^2 I & K_{f*} \\ K_{*f} & K_{**} \end{bmatrix}\right) \tag{9}$$

with the predictive distribution defined as

$$\boldsymbol{f}_*|\boldsymbol{y} \sim \mathcal{N}(K_{*f}(K_{ff} + \sigma^2 I)^{-1}\boldsymbol{y},\ K_{**} - K_{*f}(K_{ff} + \sigma^2 I)^{-1}K_{f*})$$

We can verify the non-diagonal terms by noting that $cov(\boldsymbol{y}, \boldsymbol{f}_*) = cov(\boldsymbol{f}, \boldsymbol{f}_*) = K_{f*}$.

3 Titsias

We propose a set of induction points $\boldsymbol{u}$ at locations Z and write the (augmented) joint model as

$$p(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{u}) = p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{u})p(\boldsymbol{f}|\boldsymbol{u})p(\boldsymbol{u})$$

which is equivalent to our exact model by marginalizing out $\boldsymbol{u}$. Assuming that $\boldsymbol{u}$ is a sufficient statistic for $\boldsymbol{f}$ such that $\boldsymbol{y}$ and $\boldsymbol{u}$ and independent, we obtain $p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{u}) \approx p(\boldsymbol{y}|\boldsymbol{f})$ with

$$\boldsymbol{y}|\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{f},\ \sigma^2 I)$$

with the inducing prior as

$$\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0},\ K_{uu})$$

We can write the joint Gaussian model as

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{f} \\ \boldsymbol{u} \end{bmatrix} \sim \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix} K_{ff} + \sigma^2 I & K_{ff} & K_{fu} \\ K_{ff} & K_{ff} & K_{fu} \\ K_{uf} & K_{uf} & K_{uu} \end{bmatrix}\right) \tag{10}$$

and it follows using Eq. 4 that

$$\boldsymbol{f}|\boldsymbol{u} \sim \mathcal{N}(K_{fu}K_{uu}^{-1}\boldsymbol{u},\ K_{ff} - K_{fu}K_{uu}^{-1}K_{uf})$$

Objective The exact marginal likelihood and posterior of the joint model can marginalize out the inducing points $\boldsymbol{u}$ and become the classical exact model which prohibits the learning of the inducing locations Z. The assumption earlier that $\boldsymbol{u}$ is a sufficient statistic for $\boldsymbol{f}$ such that $\boldsymbol{y}$ and $\boldsymbol{u}$ and independent, allows to learn these inducing locations but prohibits us to calculate the posterior or marginal likelihood exactly.

Instead, we introduce a variational distribution q that approaches the full posterior as

$$p(\boldsymbol{f}, \boldsymbol{u}|\boldsymbol{y}) \approx q(\boldsymbol{f}, \boldsymbol{u}) = p(\boldsymbol{f}|\boldsymbol{u})q(\boldsymbol{u}) \tag{11}$$

Since we cannot optimize the marginal likelihood $p(\boldsymbol{y})$ as its distribution is unknown, we optimize the evidence lower bound (ELBO) which can be derived by writing out the Kullback-Leibler divergence between $q(\boldsymbol{f}, \boldsymbol{u})$ and

$p(\boldsymbol{f}, \boldsymbol{u}|\boldsymbol{y})$. In other words, we try to minimize the divergence between the true posterior and the variational posterior.

$$\begin{aligned}
\mathrm{KL}(q(\boldsymbol{f}, \boldsymbol{u}) \,||\, p(\boldsymbol{f}, \boldsymbol{u}|\boldsymbol{y})) &= \iint q(\boldsymbol{f}, \boldsymbol{u}) \log \frac{q(\boldsymbol{f}, \boldsymbol{u})}{p(\boldsymbol{f}, \boldsymbol{u}|\boldsymbol{y})} d\boldsymbol{f} d\boldsymbol{u} \\
&= \iint q(\boldsymbol{f}, \boldsymbol{u}) \log \frac{q(\boldsymbol{f}, \boldsymbol{u}) p(\boldsymbol{y})}{p(\boldsymbol{f}, \boldsymbol{u}, \boldsymbol{y})} d\boldsymbol{f} d\boldsymbol{u} \\
&= \iint q(\boldsymbol{f}, \boldsymbol{u}) \left(\log \frac{q(\boldsymbol{f}, \boldsymbol{u})}{p(\boldsymbol{f}, \boldsymbol{u}, \boldsymbol{y})} + \log p(\boldsymbol{y}) \right) d\boldsymbol{f} d\boldsymbol{u} \\
&= \iint q(\boldsymbol{f}, \boldsymbol{u}) \log \frac{q(\boldsymbol{f}, \boldsymbol{u})}{p(\boldsymbol{f}, \boldsymbol{u}, \boldsymbol{y})} d\boldsymbol{f} d\boldsymbol{u} + \log p(\boldsymbol{y})
\end{aligned} \tag{12}$$

As the KL-divergence cannot be calculated, instead of minimizing it directly we maximize the ELBO. We continue

$$\begin{aligned}
\log p(\boldsymbol{y}) &= \mathrm{KL}(q(\boldsymbol{f}, \boldsymbol{u}) \,||\, p(\boldsymbol{f}, \boldsymbol{u}|\boldsymbol{y})) - \iint q(\boldsymbol{f}, \boldsymbol{u}) \log \frac{q(\boldsymbol{f}, \boldsymbol{u})}{p(\boldsymbol{f}, \boldsymbol{u}, \boldsymbol{y})} d\boldsymbol{f} d\boldsymbol{u} \\
&= \mathrm{KL}(q(\boldsymbol{f}, \boldsymbol{u}) \,||\, p(\boldsymbol{f}, \boldsymbol{u}|\boldsymbol{y})) + \iint q(\boldsymbol{f}, \boldsymbol{u}) \log \frac{p(\boldsymbol{f}, \boldsymbol{u}, \boldsymbol{y})}{q(\boldsymbol{f}, \boldsymbol{u})} d\boldsymbol{f} d\boldsymbol{u} \\
&\geq \iint q(\boldsymbol{f}, \boldsymbol{u}) \log \frac{p(\boldsymbol{f}, \boldsymbol{u}, \boldsymbol{y})}{q(\boldsymbol{f}, \boldsymbol{u})} d\boldsymbol{f} d\boldsymbol{u} = \mathrm{ELBO} \\
&= \iint p(\boldsymbol{f}|\boldsymbol{u}) q(\boldsymbol{u}) \log \frac{p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{u}) p(\boldsymbol{f}|\boldsymbol{u}) p(\boldsymbol{u})}{p(\boldsymbol{f}|\boldsymbol{u}) q(\boldsymbol{u})} d\boldsymbol{f} d\boldsymbol{u} \\
&= \iint p(\boldsymbol{f}|\boldsymbol{u}) q(\boldsymbol{u}) \log \frac{p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{u}) p(\boldsymbol{u})}{q(\boldsymbol{u})} d\boldsymbol{f} d\boldsymbol{u} \\
&= \int q(\boldsymbol{u}) \left(\int p(\boldsymbol{f}|\boldsymbol{u}) \log p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{u}) d\boldsymbol{f} + \frac{p(\boldsymbol{u})}{q(\boldsymbol{u})} \right) d\boldsymbol{u} \\
&= \log \mathcal{N}(\boldsymbol{0},\ Q_{ff} + \sigma^2 I) - \frac{1}{2\sigma^2} \mathrm{Tr}(K_{ff} - Q_{ff})
\end{aligned} \tag{13}$$

See Appendix A in Titsias' 2009 technical report for a derivation of the last line. The ELBO can be written out as

$$\begin{aligned}
\mathrm{ELBO} &= \log \mathcal{N}(\boldsymbol{0},\ Q_{ff} + \sigma^2 I) - \frac{1}{2\sigma^2} \mathrm{Tr}(K_{ff} - Q_{ff}) \\
&= -\frac{n}{2} \log 2\pi - \frac{1}{2} \log \left| Q_{ff} + \sigma^2 I \right| - \frac{1}{2} \boldsymbol{y}^T (Q_{ff} + \sigma^2 I)^{-1} \boldsymbol{y} - \frac{1}{2\sigma^2} \mathrm{Tr}(K_{ff} - Q_{ff})
\end{aligned} \tag{14}$$

Prediction We approximate the posterior as

$$\begin{aligned} p(\boldsymbol{f}_*|\boldsymbol{y}) &= \int p(\boldsymbol{f}_*, \boldsymbol{u}|\boldsymbol{y}) d\boldsymbol{u} \\ &\approx \int q(\boldsymbol{f}_*, \boldsymbol{u}) d\boldsymbol{u} \\ &= \int p(\boldsymbol{f}_*|\boldsymbol{u}) q(\boldsymbol{u}) d\boldsymbol{u} = q(\boldsymbol{f}_*) \end{aligned} \tag{15}$$

remember that

$$\boldsymbol{f}_*|\boldsymbol{u} \sim \mathcal{N}(K_{*u}K_{uu}^{-1}\boldsymbol{u},\ K_{**} - K_{*u}K_{uu}^{-1}K_{u*})$$

we take $q(\boldsymbol{u})$ as (see Appendix A of Titsias' 2009 technical report)

$$q(\boldsymbol{u}) = \mathcal{N}\left(\frac{K_{uu}}{\sigma^2}\left(K_{uu} + \frac{K_{uf}K_{fu}}{\sigma^2}\right)^{-1} K_{uf}\boldsymbol{y},\ K_{uu}\left(K_{uu} + \frac{K_{uf}K_{fu}}{\sigma^2}\right)^{-1} K_{uu}\right)$$

Using the Gaussian linear transformation of Eq.2, the predictive distribution results in

$$q(\boldsymbol{f}_*) = \int p(\boldsymbol{f}_*|\boldsymbol{u}) q(\boldsymbol{u}) d\boldsymbol{u} = \mathcal{N}(\boldsymbol{\mu}_*,\ \Sigma_*)$$

where

$$\begin{aligned} \boldsymbol{\mu}_* &= \frac{1}{\sigma^2} K_{*u}(K_{uu} + \frac{1}{\sigma^2}K_{uf}K_{fu})^{-1}K_{uf}\boldsymbol{y} \\ &= \frac{1}{\sigma^2} K_{*u}K_{uu}^{-\frac{1}{2}}(\frac{1}{\sigma^2}K_{uu}^{-\frac{1}{2}}K_{uf}K_{fu}K_{uu}^{-\frac{1}{2}} + I)^{-1}K_{uu}^{-\frac{1}{2}}K_{uf}\boldsymbol{y} \end{aligned} \tag{16}$$

and

$$\begin{aligned} \Sigma_* &= K_{**} - Q_{*f}(Q_{ff} + \sigma^2 I)^{-1}Q_{f*} \\ &= K_{**} - Q_{**} + K_{*u}(K_{uu} + \frac{1}{\sigma^2}K_{uf}K_{fu})^{-1}K_{u*} \\ &= K_{**} - Q_{**} + K_{*u}K_{uu}^{-\frac{1}{2}}(\frac{1}{\sigma^2}K_{uu}^{-\frac{1}{2}}K_{uf}K_{fu}K_{uu}^{-\frac{1}{2}} + I)^{-1}K_{uu}^{-\frac{1}{2}}K_{u*} \end{aligned} \tag{17}$$

4 Hensman

We introduce a variational distribution that approximates the posterior as

$$p(\boldsymbol{f}|\boldsymbol{y}) \approx q(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{\mu}_q,\ \Sigma_q)$$

and we take the likelihood $p(\boldsymbol{y}|\boldsymbol{f})$ to be a known distribution that is not necessarily Gaussian.

Objective We will want to minimize the divergence between the exact posterior and the variational distribution, that is to minimize the Kullback-Leibler divergence defined as

$$\begin{aligned}
\mathrm{KL}(q(\boldsymbol{f}) \,||\, p(\boldsymbol{f}|\boldsymbol{y})) &= \int q(\boldsymbol{f}) \log \frac{q(\boldsymbol{f})}{p(\boldsymbol{f}|\boldsymbol{y})} d\boldsymbol{f} \\
&= \int q(\boldsymbol{f}) \log \frac{q(\boldsymbol{f})p(\boldsymbol{y})}{p(\boldsymbol{f},\boldsymbol{y})} d\boldsymbol{f} \\
&= \int q(\boldsymbol{f}) \left(\log \frac{q(\boldsymbol{f})}{p(\boldsymbol{f})} - \log(p(\boldsymbol{y}|\boldsymbol{f})) + \log p(\boldsymbol{y}) \right) d\boldsymbol{f} \\
&= \int q(\boldsymbol{f}) \log \frac{q(\boldsymbol{f})}{p(\boldsymbol{f})} d\boldsymbol{f} - \int q(\boldsymbol{f}) \log p(\boldsymbol{y}|\boldsymbol{f}) d\boldsymbol{f} + \log p(\boldsymbol{y}) \\
&= -\mathrm{ELBO} + \log p(\boldsymbol{y})
\end{aligned} \tag{18}$$

Since we cannot calculate the KL-divergence, we maximize the evidence lower bound (ELBO) instead in order to approximate our objective of maximizing $p(\boldsymbol{y})$.

$$\begin{aligned}
\mathrm{ELBO} &= \int q(\boldsymbol{f}) \log p(\boldsymbol{y}|\boldsymbol{f}) d\boldsymbol{f} - \int q(\boldsymbol{f}) \log \frac{q(\boldsymbol{f})}{p(\boldsymbol{f})} d\boldsymbol{f} \\
&= \int q(\boldsymbol{f}) \log p(\boldsymbol{y}|\boldsymbol{f}) d\boldsymbol{f} - \mathrm{KL}(q(\boldsymbol{f}) \,||\, p(\boldsymbol{f}))
\end{aligned} \tag{19}$$

The second term is the Kullback-Leibler divergence between two known Gaussians (see Variational model above) as we remember that $p(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{0},\ K_{ff})$. The first term can be calculated by remembering that

$$\boldsymbol{y}|\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{f},\ \sigma^2 I)$$

As the likelihood is not necessarily Gaussian, we use Gaussian quadratures to calculate the integral.

Prediction From the joint model

$$\begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K_{ff} & K_{f*} \\ K_{*f} & K_{**} \end{bmatrix}\right) \tag{20}$$

it follows using Eq. 4 that

$$\boldsymbol{f}_* | \boldsymbol{f} \sim \mathcal{N}(K_{*f} K_{ff}^{-1} \boldsymbol{f},\ K_{**} - K_{*f} K_{ff}^{-1} K_{f*})$$

so that our predictive distribution can be written as

$$\begin{aligned} p(\boldsymbol{f}_* | \boldsymbol{y}) &= \int p(\boldsymbol{f}_* | \boldsymbol{f}, \boldsymbol{y}) p(\boldsymbol{f} | \boldsymbol{y}) d\boldsymbol{f} \\ &= \int p(\boldsymbol{f}_* | \boldsymbol{f}) p(\boldsymbol{f} | \boldsymbol{y}) d\boldsymbol{f} \\ &\approx \int p(\boldsymbol{f}_* | \boldsymbol{f}) q(\boldsymbol{f}) d\boldsymbol{f} = q(\boldsymbol{f}_*) \end{aligned} \tag{21}$$

Using the Gaussian linear transformation of Eq.2, the predictive distribution results in

$$q(\boldsymbol{f}_*) = \mathcal{N}(K_{*f} K_{ff}^{-1} \boldsymbol{\mu}_q,\ K_{**} - Q_{**} + K_{*f} K_{ff}^{-1} \Sigma_q K_{ff}^{-1} K_{f*})$$

Reparametrization In general, this model is difficult to optimize since optimizing the kernel parameters and the variational parameters are optimized as separate terms. We can reduce the optimization space and improve training results by reparametrizing $\boldsymbol{\mu}_q \rightarrow L\boldsymbol{\mu}_q$ and $\Sigma_q \rightarrow L\Sigma_q L^t$ where $LL^T = K_{ff}$. It follows that

$$q(\boldsymbol{f}) = \mathcal{N}(L\boldsymbol{\mu}_q,\ L\Sigma_q L^T)$$

$$q(\boldsymbol{f}_*) = \mathcal{N}(K_{*f} L^{-T} \boldsymbol{\mu}_q,\ K_{**} - Q_{**} + K_{*f} L^{-T} \Sigma_q L^{-1} K_{f*})$$

The Kullback-Leibler divergence part of the ELBO simplifies to

$$\mathrm{KL}(q(\boldsymbol{f}) \,||\, p(\boldsymbol{f})) = \mathrm{KL}(\mathcal{N}(\boldsymbol{\mu}_q,\ \Sigma_q) \,||\, \mathcal{N}(0,\ I))$$

5 Sparse Hensman

See the Variational model for a basis, however now we introduce $\boldsymbol{u}$ as our inducing variables. Remember, we use the variational parameters $\boldsymbol{\mu}_q$ and Σ_q to specify our prior

$$q(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{\mu}_q, \Sigma_q)$$

Objective Our derivation is very similar to the variational model above (see Eq. 19), but our ELBO now incorporates the inducing variables and becomes

$$\begin{aligned}
\text{ELBO} &= \iint q(\boldsymbol{f}, \boldsymbol{u}) \log p(\boldsymbol{y}|\boldsymbol{f}) d\boldsymbol{f} d\boldsymbol{u} - \text{KL}(q(\boldsymbol{f}, \boldsymbol{u}) \,||\, p(\boldsymbol{f}, \boldsymbol{u})) \\
&= \int \left(\int q(\boldsymbol{f}, \boldsymbol{u}) d\boldsymbol{u} \right) \log p(\boldsymbol{y}|\boldsymbol{f}) d\boldsymbol{f} - \iint q(\boldsymbol{f}, \boldsymbol{u}) \log \frac{q(\boldsymbol{f}, \boldsymbol{u})}{p(\boldsymbol{f}, \boldsymbol{u})} d\boldsymbol{f} d\boldsymbol{u} \\
&= \int q(\boldsymbol{f}) \log p(\boldsymbol{y}|\boldsymbol{f}) d\boldsymbol{f} - \iint q(\boldsymbol{f}, \boldsymbol{u}) \log \frac{p(\boldsymbol{f}|\boldsymbol{u}) q(\boldsymbol{u})}{p(\boldsymbol{f}|\boldsymbol{u}) p(\boldsymbol{u})} d\boldsymbol{f} d\boldsymbol{u} \\
&= \int q(\boldsymbol{f}) \log p(\boldsymbol{y}|\boldsymbol{f}) d\boldsymbol{f} - \int \left(\int q(\boldsymbol{f}, \boldsymbol{u}) d\boldsymbol{f} \right) \log \frac{q(\boldsymbol{u})}{p(\boldsymbol{u})} d\boldsymbol{u} \\
&= \int q(\boldsymbol{f}) \log p(\boldsymbol{y}|\boldsymbol{f}) d\boldsymbol{f} - \int q(\boldsymbol{u}) \log \frac{q(\boldsymbol{u})}{p(\boldsymbol{u})} d\boldsymbol{u} \\
&= \int q(\boldsymbol{f}) \log p(\boldsymbol{y}|\boldsymbol{f}) d\boldsymbol{f} - \text{KL}(q(\boldsymbol{u}) \,||\, p(\boldsymbol{u}))
\end{aligned} \tag{22}$$

The second term is the Kullback-Leibler divergence between two known Gaussians (see Variational model above) as we remember that $p(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{0},\ K_{uu})$. The first term can be calculated by remembering that

$$\boldsymbol{y}|\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{f},\ \sigma^2 I)$$

and

$$q(\boldsymbol{f}) = \int p(\boldsymbol{f}|\boldsymbol{u}) q(\boldsymbol{u}) d\boldsymbol{u} \tag{23}$$

Then from our joint model

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{f} \end{bmatrix} \sim \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix} K_{uu} & K_{uf} \\ K_{fu} & K_{ff} \end{bmatrix} \right) \tag{24}$$

and using Eq. 4 we obtain

$$p(\boldsymbol{f}|\boldsymbol{u}) = \mathcal{N}(K_{fu} K_{uu}^{-1} \boldsymbol{u},\ K_{ff} - K_{fu} K_{uu}^{-1} K_{uf})$$

Now using the Gaussian linear transformation rule on Eq. 23, we obtain our predictive distribution

$$q(\boldsymbol{f}) = \mathcal{N}(K_{fu}K_{uu}^{-1}\boldsymbol{\mu}_q,\ K_{ff} - Q_{ff} + K_{fu}K_{uu}^{-1}\Sigma_q K_{uu}^{-1}K_{uf})$$

As the likelihood is not necessarily Gaussian, we use Gaussian quadratures to calculate the integral, see the Variational model above.

Prediction From $q(\boldsymbol{f})$ above, our predictive distribution is

$$q(\boldsymbol{f}_*) = \mathcal{N}(K_{*u}K_{uu}^{-1}\boldsymbol{\mu}_q,\ K_{**} - Q_{**} + K_{*u}K_{uu}^{-1}\Sigma_q K_{uu}^{-1}K_{u*})$$

Reparametrization In general, this model is difficult to optimize since optimizing the kernel parameters and the variational parameters are optimized as separate terms. We can reduce the optimization space and improve training results by reparametrizing $\boldsymbol{\mu}_q \rightarrow L\boldsymbol{\mu}_q$ and $\Sigma_q \rightarrow L\Sigma_q L^t$ where $LL^T = K_{uu}$. It follows that

$$q(\boldsymbol{u}) = \mathcal{N}(L\boldsymbol{\mu}_q,\ L\Sigma_q L^T)$$

$$q(\boldsymbol{f}) = \mathcal{N}(K_{fu}L^{-T}\boldsymbol{\mu}_q,\ K_{ff} - Q_{ff} + K_{fu}L^{-T}\Sigma_q L^{-1}K_{uf})$$

$$q(\boldsymbol{f}_*) = \mathcal{N}(K_{*u}L^{-T}\boldsymbol{\mu}_q,\ K_{**} - Q_{**} + K_{*u}L^{-T}\Sigma_q L^{-1}K_{u*})$$

The Kullback-Leibler divergence part of the ELBO simplifies to

$$\mathrm{KL}(q(\boldsymbol{u}) \,||\, p(\boldsymbol{u})) = \mathrm{KL}(\mathcal{N}(\boldsymbol{\mu}_q,\ \Sigma_q) \,||\, \mathcal{N}(0,\ I))$$

6 Appendix: likelihoods

The ELBO contains the following term

$$\int \log p(\boldsymbol{y}|\boldsymbol{f}) q(\boldsymbol{f}) d\boldsymbol{f}$$

where $q(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{\mu},\ \Sigma)$ and the likelihood $p(\boldsymbol{y}|\boldsymbol{f})$ can be of any distribution. We rewrite the integral to allow for solving using Gaussian quadratures.

$$\begin{aligned}
&\int \log p(\boldsymbol{y}|\boldsymbol{f}) q(\boldsymbol{f}) d\boldsymbol{f} \\
&= \sum_{i=1}^{n} \int \log p(y_i|f_i) q(f_i) df_i \\
&= \sum_{i=1}^{n} \int \log p(y_i|f_i) \frac{1}{\sqrt{2\pi\Sigma_{ii}}} e^{-\frac{1}{2}(f_i-\mu_i)^2\Sigma_{ii}^{-1}} df_i \\
&= \sum_{i=1}^{n} \int \log p(y_i|\mu_i + \sqrt{2\Sigma_{ii}}x_i) \frac{1}{\sqrt{\pi}} e^{-x_i^2} dx_i \\
&\approx \sum_{i=1}^{n} \sum_{j=1}^{m} w_j g_i(t_j)
\end{aligned} \tag{25}$$

where $x_i = \frac{1}{\sqrt{2\Sigma_{ii}}}(f_i - \mu_i)$ and $dx_i = \frac{1}{\sqrt{2\Sigma_{ii}}} df_i$ so that our function g is defined as

$$g_i(t_j) = \frac{1}{\sqrt{\pi}} \log p(y_i|\mu_i + \sqrt{2\Sigma_{ii}}t_j)$$

6.1 Gaussian

Given the following Gaussian likelihood

$$p(y_i|f_i) = \mathcal{N}(y_i|f_i,\ \sigma^2 I)$$

the function g becomes

$$g_i(t_j) = \frac{1}{\sqrt{\pi}} \left(-\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_i - \mu_i - \sqrt{2\Sigma_{ii}}t_j)^2 \right)$$

Exact The exact solution exists as

$$\begin{aligned}
&\sum_{i=1}^{n} \int \log p(y_i|f_i) q(f_i) df_i \\
&= \sum_{i=1}^{n} \int \left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_i - f_i)^2 \right) \frac{1}{\sqrt{2\pi\Sigma_{ii}}} e^{-\frac{1}{2}\frac{(f_i-\mu_i)^2}{\Sigma_{ii}}} df_i \\
&= \sum_{i=1}^{n} \left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\frac{1}{\sqrt{2\pi\Sigma_{ii}}} \int (y_i - f_i)^2 e^{-\frac{1}{2}\frac{(f_i-\mu_i)^2}{\Sigma_{ii}}} df_i \right) \\
&= \sum_{i=1}^{n} \left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\frac{1}{\sqrt{2\pi\Sigma_{ii}}} \int x_i^2 e^{-\frac{1}{2\Sigma_{ii}}(x_i+y_i-\mu_i)^2} dx_i \right)
\end{aligned} \tag{26}$$

We can expand the integral in the second term using known definite integrals of exponentials, defining $a = 1/(2\Sigma_{ii})$, $b = 2(y_i - \mu_i)/(2\Sigma_{ii})$, and $c = (y_i - \mu_i)^2/(2\Sigma_{ii})$, as

$$\begin{aligned}
&\int x_i^2 e^{-\frac{1}{2\Sigma_{ii}}(x_i+y_i-\mu_i)^2} dx_i \\
&= \int x_i^2 e^{-\frac{1}{2\Sigma_{ii}}\left(x_i^2 + x_i(2y_i-2\mu_i) + (y_i-\mu_i)^2\right)} dx_i \\
&= e^{-c} \int x_i^2 e^{-ax_i^2 - bx_i} dx_i \\
&= e^{-c} \cdot \frac{\sqrt{\pi}(2a+b^2)}{4a^{5/2}} e^{b^2/4a} && \text{(known definitive integral)} \\
&= \frac{\sqrt{\pi}\left(\frac{2}{2\Sigma_{ii}} + 4\frac{(y_i-\mu_i)^2}{4\Sigma_{ii}^2)}\right)}{4(2\Sigma_{ii})^{-5/2}} e^{\frac{4(y_i-\mu_i)^2}{4\Sigma_{ii}^2}\frac{2\Sigma_{ii}}{4} - \frac{(y_i-\mu_i)^2}{2\Sigma_{ii}}} \\
&= \frac{\sqrt{\pi}\left(\frac{1}{\Sigma_{ii}} + \frac{(y_i-\mu_i)^2}{\Sigma_{ii}^2)}\right)}{4(2\Sigma_{ii})^{-5/2}} \\
&= \sqrt{2\pi\Sigma_{ii}}\left((y_i-\mu_i)^2 + \Sigma_{ii}\right)
\end{aligned} \tag{27}$$

so that

$$\sum_{i=1}^{n} \int \log p(y_i|f_i) q(f_i) df_i = \sum_{i=1}^{n} \left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left((y_i-\mu_i)^2 + \Sigma_{ii}\right) \right)$$

6.2 Student-T

Given the following Student-T likelihood

$$p(y_i|f_i) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi\sigma^2}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{1}{\nu}\left(\frac{y_i-f_i}{\sigma}\right)^2\right)^{-\frac{\nu+1}{2}}$$

with scale σ and degrees-of-freedom ν, the function g becomes

$$\begin{aligned} g_i(t_j) = \frac{1}{\sqrt{\pi}}\Bigg[&\log\Gamma\left(\frac{\nu+1}{2}\right) - \frac{1}{2}\log(\nu\pi\sigma^2) - \log\Gamma\left(\frac{\nu}{2}\right) \\ &-\frac{\nu+1}{2}\log\left(1+\frac{1}{\nu}\left(\frac{y_i-\mu_i-\sqrt{2\Sigma_{ii}}t_j}{\sigma}\right)^2\right)\Bigg] \end{aligned} \tag{28}$$

6.3 Laplace

Given the following Laplace likelihood

$$p(y_i|f_i) = \frac{1}{2\sigma}e^{-\frac{1}{\sigma}|y_i-f_i|}$$

with scale σ, the function g becomes

$$g_i(t_j) = \frac{1}{\sqrt{\pi}}\left(-\log(2\sigma) - \frac{|y_i-\mu_i-\sqrt{2\Sigma_{ii}}t_j|}{\sigma}\right)$$

Exact The exact solution exists as

$$
\begin{aligned}
&\sum_{i=1}^{n} \int_{-\infty}^{\infty} \log p(y_i|f_i) q(f_i) df_i \\
&= \sum_{i=1}^{n} \int_{-\infty}^{\infty} \left(-\log(2\sigma) - \frac{|y_i - f_i|}{\sigma} \right) \frac{1}{\sqrt{2\pi\Sigma_{ii}}} e^{-\frac{1}{2}\frac{(f_i-\mu_i)^2}{\Sigma_{ii}}} df_i \\
&= \sum_{i=1}^{n} -\log(2\sigma) - \frac{1}{\sigma\sqrt{2\pi\Sigma_{ii}}} \int_{-\infty}^{\infty} |y_i - f_i| e^{-\frac{1}{2}\frac{(f_i-\mu_i)^2}{\Sigma_{ii}}} df_i \\
&= \sum_{i=1}^{n} -\log(2\sigma) + \frac{1}{\sigma\sqrt{2\pi\Sigma_{ii}}} \int_{\infty}^{-\infty} |z_i| e^{-\frac{1}{2}\frac{(y_i-z_i-\mu_i)^2}{\Sigma_{ii}}} dz_i \\
&= \sum_{i=1}^{n} -\log(2\sigma) + \frac{2}{\sigma\sqrt{2\pi\Sigma_{ii}}} \int_{0}^{-\infty} z_i e^{-\frac{1}{2}\frac{(y_i-z_i-\mu_i)^2}{\Sigma_{ii}}} dz_i \qquad (29) \\
&= \sum_{i=1}^{n} -\log(2\sigma) - \frac{2}{\sigma\sqrt{2\pi\Sigma_{ii}}} \int_{0}^{\infty} (y_i - x_i - \mu_i) e^{-\frac{1}{2}\frac{x_i^2}{\Sigma_{ii}}} dx_i \\
&= \sum_{i=1}^{n} -\log(2\sigma) - \frac{y_i}{\sigma} + \frac{\mu_i}{\sigma} + \frac{2}{\sigma\sqrt{2\pi\Sigma_{ii}}} \int_{0}^{\infty} x_i e^{-\frac{1}{2}\frac{x_i^2}{\Sigma_{ii}}} dx_i \\
&= \sum_{i=1}^{n} -\log(2\sigma) - \frac{y_i}{\sigma} + \frac{\mu_i}{\sigma} + \frac{2}{\sigma\sqrt{2\pi\Sigma_{ii}}} \frac{2\Sigma_{ii}}{2} \\
&= \sum_{i=1}^{n} -\log(2\sigma) - \frac{y_i}{\sigma} + \frac{\mu_i}{\sigma} + \frac{1}{\sigma}\sqrt{\frac{2\Sigma_{ii}}{\pi}}
\end{aligned}
$$

DOESN'T SEEM TO WORK!

6.4 Gamma

Given the following Gamma likelihood

$$p(y_i|f_i) = \frac{1}{\Gamma(k)h(f)^k} y_i^{k-1} e^{-\frac{y_i}{h(f_i)}}$$

with scale $h(f)$ and shape k, where h is the link function, the function g becomes

$$g_i(t_j) = \frac{1}{\sqrt{\pi}} \left(-\log\Gamma(k) - k\log h(f_i) + (k-1)\log(y_i) - \frac{y_i}{h(f_i)} \right)$$

Exact The exact solution exists when $h(\cdot) = e^{\cdot}$.

$$
\begin{aligned}
&\sum_{i=1}^{n} \int \log p(y_i|f_i) q(f_i) df_i \\
&= \sum_{i=1}^{n} \int \left(-\log \Gamma(k) - k f_i + (k-1)\log y_i - y_i e^{-f_i} \right) \frac{1}{\sqrt{2\pi\Sigma_{ii}}} e^{-\frac{1}{2}\frac{(f_i-\mu_i)^2}{\Sigma_{ii}}} df_i \\
&= \sum_{i=1}^{n} \int \left(-\log \Gamma(k) - k(x_i+\mu_i) + (k-1)\log y_i - y_i e^{-x_i-\mu_i} \right) \frac{1}{\sqrt{2\pi\Sigma_{ii}}} e^{-\frac{1}{2}\frac{x_i^2}{\Sigma_{ii}}} dx_i \\
&= \sum_{i=1}^{n} \left(-\log \Gamma(k) + (k-1)\log y_i - \frac{1}{\sqrt{2\pi\Sigma_{ii}}} \left(\int k(x_i+\mu_i) e^{-\frac{1}{2}\frac{x_i^2}{\Sigma_{ii}}} dx_i + \int y_i e^{-x_i-\mu_i} e^{-\frac{1}{2}\frac{x_i^2}{\Sigma_{ii}}} dx_i \right) \right) \\
&= \sum_{i=1}^{n} \left(-\log \Gamma(k) + (k-1)\log y_i - k\mu_i - \frac{1}{\sqrt{2\pi\Sigma_{ii}}} \int y_i e^{-\frac{1}{2}\frac{x_i^2}{\Sigma_{ii}} - x_i - \mu_i} dx_i \right) \\
&= \sum_{i=1}^{n} \left(-\log \Gamma(k) + (k-1)\log y_i - k\mu_i - y_i e^{\frac{1}{2}\Sigma_{ii} - \mu_i} \right)
\end{aligned}
\tag{30}
$$

6.5 Exponential

Given the following exponential likelihood

$$p(y_i|f_i) = \frac{1}{h(f_i)} e^{-\frac{1}{h(f_i)} y_i}$$

with scale $h(f_i)$ where h is the link function.

Exact The exact solution exists when $h(\cdot) = e^{\cdot}$.

$$
\begin{aligned}
&\sum_{i=1}^{n} \int \log p(y_i|f_i) q(f_i) df_i \\
&= \sum_{i=1}^{n} \int (-f_i - y_i e^{-f_i}) \frac{1}{\sqrt{2\pi\Sigma_{ii}}} e^{-\frac{1}{2}\frac{(f_i-\mu_i)^2}{\Sigma_{ii}}} df_i \\
&= \sum_{i=1}^{n} \int (-x_i - \mu_i - y_i e^{-x_i-\mu_i}) \frac{1}{\sqrt{2\pi\Sigma_{ii}}} e^{-\frac{1}{2\Sigma_{ii}}x_i^2} dx_i \\
&= \sum_{i=1}^{n} -\frac{1}{\sqrt{2\pi\Sigma_{ii}}} \left(\int x_i e^{-\frac{1}{2\Sigma_{ii}}x_i^2} dx_i + \int \mu_i e^{-\frac{1}{2\Sigma_{ii}}x_i^2} dx_i + \int y_i e^{-x_i-\mu_i} e^{-\frac{1}{2\Sigma_{ii}}x_i^2} dx_i \right) \\
&= \sum_{i=1}^{n} -\frac{1}{\sqrt{2\pi\Sigma_{ii}}} \left(0 + \mu_i \sqrt{2\pi\Sigma_{ii}} + y_i e^{-\mu_i} \int e^{-\frac{1}{2\Sigma_{ii}}x_i^2 - x_i} dx_i \right) \\
&= \sum_{i=1}^{n} -\frac{1}{\sqrt{2\pi\Sigma_{ii}}} \left(0 + \mu_i \sqrt{2\pi\Sigma_{ii}} + y_i e^{-\mu_i} \sqrt{2\pi\Sigma_{ii}} e^{\frac{1}{4}2\Sigma_{ii}} \right) \\
&= \sum_{i=1}^{n} \left(-\mu_i - y_i e^{\frac{1}{2}\Sigma_{ii} - \mu_i} \right)
\end{aligned}
\tag{31}
$$

6.6 Poisson

Given the following Poisson likelihood

$$p(y_i|f_i) = \frac{h(f_i)^{y_i} e^{-h(f_i)}}{\Gamma(1+y_i)}$$

with scale $h(f_i)$ where h is the link function.

Exact The exact solution exists when $h(\cdot) = e^{\cdot}$.

$$
\begin{aligned}
&\sum_{i=1}^{n} \int \log p(y_i|f_i) q(f_i) df_i \\
&= \sum_{i=1}^{n} \int \left(f_i y_i - e^{f_i} - \log \Gamma(1+y_i) \right) \frac{1}{\sqrt{2\pi\Sigma_{ii}}} e^{-\frac{1}{2}\frac{(f_i-\mu_i)^2}{\Sigma_{ii}}} df_i \\
&= \sum_{i=1}^{n} \int \left((x_i+\mu_i) y_i - e^{x_i+\mu_i} - \log \Gamma(1+y_i) \right) \frac{1}{\sqrt{2\pi\Sigma_{ii}}} e^{-\frac{1}{2\Sigma_{ii}} x_i^2} dx_i \\
&= \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi\Sigma_{ii}}} \left(\int (x_i+\mu_i) y_i e^{-\frac{1}{2\Sigma_{ii}} x_i^2} dx_i - \int e^{x_i+\mu_i} e^{-\frac{1}{2\Sigma_{ii}} x_i^2} dx_i - \int \log \Gamma(1+y_i) e^{-\frac{1}{2\Sigma_{ii}} x_i^2} dx_i \right) \\
&= \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi\Sigma_{ii}}} \left(\mu_i y_i \sqrt{2\pi\Sigma_{ii}} - \int e^{-\frac{1}{2\Sigma_{ii}} x_i^2 + x_i + \mu_i} dx_i - \log \Gamma(1+y_i) \sqrt{2\pi\Sigma_{ii}} \right) \\
&= \sum_{i=1}^{n} \left(\mu_i y_i - e^{\frac{1}{2}\Sigma_{ii} + \mu_i} - \log \Gamma(1+y_i) \right)
\end{aligned}
\tag{32}
$$

7 Appendix: predictive distribution

Given our predictive posterior $p(\boldsymbol{f}_*|\boldsymbol{y}) \approx q(\boldsymbol{f}_*) \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, we can obtain a prediction of $\boldsymbol{y}_*$ using

$$p(\boldsymbol{y}_*|\boldsymbol{y}) = \int p(\boldsymbol{y}_*|\boldsymbol{f}_*)q(\boldsymbol{f}_*)d\boldsymbol{f}_*$$

where $p(\boldsymbol{y}_*|\boldsymbol{f}_*)$ is our likelihood.

As our predictive distribution is usually not Gaussian, we can calculate its mean and variance by

$$\begin{aligned} \boldsymbol{\mu}_* = \mathbb{E}[Y_*] &= \int \boldsymbol{y}_* p(\boldsymbol{y}_*|\boldsymbol{y}) d\boldsymbol{y}_* \\ \boldsymbol{\sigma}_*^2 = \mathrm{Var}[Y_*] &= \int (\boldsymbol{y}_* - \boldsymbol{\mu}_*)^2 p(\boldsymbol{y}_*|\boldsymbol{y}) d\boldsymbol{y}_* = \mathbb{E}[Y_*^2] - \mathbb{E}[Y_*]^2 \end{aligned} \tag{33}$$

Both can be solved using Gaussian quadratures as follows, with $f_j = \mu_j + \sqrt{2\Sigma_{jj}}x_j$.

$$\begin{aligned} \mu_{i*} &= \int y_i \int p(y_{i*}|f_i)p(f_i|y_i)df_i dy_i \\ &\approx \int y_i \left[\Sigma_{j=1}^m \frac{w_j}{\sqrt{\pi}} p(y_{i*}|\mu_j + \sqrt{2\Sigma_{jj}}x_j)\right] dy_i \\ &= \Sigma_{j=1}^m \frac{w_j}{\sqrt{\pi}} \int y_i p(y_{i*}|f_j) dy_i \\ \sigma_{i*}^2 &\approx \Sigma_{j=1}^m \frac{w_j}{\sqrt{\pi}} \int y_i^2 p(y_{i*}|f_j) dy_i - \mu_{i*}^2 \end{aligned} \tag{34}$$

We now find the solution of the integrals in μ_{i*} and σ_{i*}^2 for various likelihoods.

7.1 Gaussian

$$\begin{aligned} \int y_i p(y_{i*}|f_j) dy_i &= \int y_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - f_j)^2} dy_i \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} f_j \sqrt{\frac{\pi}{\frac{1}{2\sigma^2}}} \\ &= f_j \end{aligned} \tag{35}$$

$$
\begin{aligned}
\int y_i^2 p(y_{i*}|f_j)dy_i &= \int y_i^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - f_j)^2} dy_i \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}f_j^2} \int y_i^2 e^{-\frac{1}{2\sigma^2}y_i^2 + \frac{1}{\sigma^2}f_j y_i} dy_i \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}f_j^2} \sqrt{\pi} \frac{\frac{1}{\sigma^2} + \frac{1}{\sigma^4}f_j^2}{4(\frac{1}{2\sigma^2})^{5/2}} e^{\frac{\frac{1}{\sigma^4}f_j^2}{4\frac{1}{2\sigma^2}}} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}} \sqrt{2\pi\sigma^2}(\sigma^2 + f_j^2) e^{\frac{1}{2\sigma^2}f_j^2} \\
&= \sigma^2 + f_j^2
\end{aligned}
\tag{36}
$$

Exact

$$
\begin{aligned}
\int p(\boldsymbol{y}_*|\boldsymbol{f}_*)q(\boldsymbol{f}_*)d\boldsymbol{f}_* &= \int \mathcal{N}(\boldsymbol{y}_*|\boldsymbol{0}, \sigma^2)\mathcal{N}(\boldsymbol{f}_*|\boldsymbol{\mu}, \Sigma)d\boldsymbol{f}_* \\
&= \mathcal{N}(\boldsymbol{\mu}, \Sigma + \sigma^2)
\end{aligned}
\tag{37}
$$

where $\boldsymbol{\mu}_* = \boldsymbol{\mu}$ and $\Sigma_* = \Sigma + \sigma^2$.

7.2 Student-T

Given $t(x)$ the probability density function of a scaled Student-T distribution (where $x = y_i - f_i$ and with scale parameter σ and degrees-of-freedom ν), we note that the density function of the Student-T distribution is

$$
t(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma(\nu/2)\sqrt{\nu\pi\sigma^2}} \left(1 + \frac{x^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}
$$

and its derivative to x is given as

$$
\begin{aligned}
\frac{d}{dx}t(x) &= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma(\nu/2)\sqrt{\nu\pi\sigma^2}} \left(-\frac{\nu+1}{2}\left(1 + \frac{x^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}-1} \frac{2x}{\nu\sigma^2}\right) \\
&= -\frac{\nu+1}{2}\left(1 + \frac{x^2}{\nu\sigma^2}\right)^{-1} \frac{2x}{\nu\sigma^2} t(x) \\
&= -\frac{(\nu+1)}{\nu\sigma^2 + x^2} x t(x)
\end{aligned}
\tag{38}
$$

rewriting and then using the product rule, we continue

$$
\begin{aligned}
x t(x) &= -\frac{\nu\sigma^2 + x^2}{(\nu+1)} \frac{d}{dx} t(x) \\
&= \frac{d}{dx}\left(\frac{\nu\sigma^2 + x^2}{(\nu-1)} t(x)\right)
\end{aligned}
\tag{39}
$$

Now we can calculate the mean by changing variable with $x_i = y_i - f_j$ and noting that the definitive integral of the cumulative density function of $t(x)$ is one. We assume that $\nu > 1$, so that

$$\begin{aligned}\int y_i p(y_{i*}|f_j)dy_i &= \int y_i t(y_i - f_j)dy_i \\ &= \int x_i t(x_i)dx_i + f_j \int t(x_i)dx_i \\ &= \left[\frac{\nu\sigma^2 + x^2}{\nu - 1}t(x)\right]_{-\infty}^{\infty} + f_j \\ &= 0 + f_j\end{aligned} \tag{40}$$

For the variance we can do the same

$$\begin{aligned}\frac{d^2}{dx^2}t(x) &= -\frac{d}{dx}\frac{(\nu+1)}{\nu\sigma^2 + x^2}xt(x) \\ &= \frac{\nu+1}{\left(\nu\sigma^2 + x^2\right)^2}\left((\nu+2)x^2 - \nu\sigma^2\right)t(x)\end{aligned} \tag{41}$$

rewriting and then using the product rule, we continue

$$\begin{aligned}x^2 t(x) &= \frac{\nu\sigma^2}{\nu+2}t(x) + \frac{\left(\nu\sigma^2 + x^2\right)^2}{(\nu+1)(\nu+2)}\frac{d}{dx}\left(-\frac{(\nu+1)}{\nu\sigma^2 + x^2}xt(x)\right) \\ &= \frac{\nu\sigma^2}{\nu+2}t(x) + \frac{d}{dx}\left(-\frac{\nu\sigma^2 + x^2}{\nu+2}xt(x)\right) + \frac{4}{\nu+2}x^2 t(x) \\ &= \frac{\nu\sigma^2}{\nu-2}t(x) + \frac{d}{dx}\left(\frac{\nu\sigma^2 + x^2}{2-\nu}xt(x)\right)\end{aligned} \tag{42}$$

Now we can calculate the variance, by changing variable with $x_i = y_i - f_j$ and assuming that $\nu > 2$, we obtain

$$\begin{aligned}\int y_i^2 p(y_{i*}|f_j)dy_i &= \int y_i^2 t(y_i - f_j)dy_i \\ &= \int x_i^2 t(x_i)dx_i + 2f_j \int x_i t(x_i)dx_i + f_j^2 \int t(x_i)dx_i \\ &= \int \frac{\nu\sigma^2}{\nu-2}t(x_i) + \left[\frac{\nu\sigma^2 + x_i^2}{2-\nu}x_i t(x_i)\right]_{-\infty}^{\infty} + 0 + f_j^2 \\ &= \frac{\nu\sigma^2}{\nu-2} + 0 + f_j^2\end{aligned} \tag{43}$$

7.3 Laplace

$$\begin{aligned}
\int y_i p(y_{i*}|f_j)dy_i &= \int y_i \frac{1}{2\sigma} e^{-\frac{|y_i - f_j|}{\sigma}} dy_i \\
&= \frac{1}{2\sigma}\int (x_i + f_j) e^{-\frac{|x_i|}{\sigma}} dx_i \\
&= \frac{1}{2\sigma}\int x_i e^{-\frac{|x_i|}{\sigma}} dx_i + \frac{1}{2\sigma} f_j \int e^{-\frac{|x_i|}{\sigma}} dx_i \\
&= 0 + \frac{1}{2\sigma} f_j 2 \int_0^\infty e^{-\frac{x_i}{\sigma}} dx_i \\
&= \frac{1}{2\sigma} f_j 2\sigma = f_j
\end{aligned} \tag{44}$$

$$\begin{aligned}
\int y_i^2 p(y_{i*}|f_j)dy_i &= \int y_i^2 \frac{1}{2\sigma} e^{-\frac{|y_i - f_j|}{\sigma}} dy_i \\
&= \frac{1}{2\sigma}\int (x_i + f_j)^2 e^{-\frac{|x_i|}{\sigma}} dx_i \\
&= \frac{1}{2\sigma}\left(\int x_i^2 e^{-\frac{|x_i|}{\sigma}} dx_i + 2f_j \int x_i e^{-\frac{|x_i|}{\sigma}} dx_i + f_j^2 \int e^{-\frac{|x_i|}{\sigma}} dx_i\right) \\
&= \frac{1}{2\sigma}\left(2\int_0^\infty x_i^2 e^{-\frac{x_i}{\sigma}} dx_i + 0 + f_j^2 2\sigma\right) \\
&= \frac{1}{2\sigma}\left(4\sigma^3 + f_j^2 2\sigma\right) \\
&= 2\sigma^2 + f_j^2
\end{aligned} \tag{45}$$

7.4 Gamma

$$\begin{aligned}
\int y_i p(y_{i*}|f_j)dy_i &= \int y_i \frac{(y_i - f_i)^{k-1}}{\Gamma(k)\sigma^k} e^{-(y_i - f_i)/\sigma} dy_i \\
&= \frac{1}{\Gamma(k)\sigma^k}\int (x_i + f_j) x_i^{k-1} e^{-x_i/\sigma} dx_i \\
&= \frac{1}{\Gamma(k)\sigma^k}\left(\int x_i^k e^{-x_i/\sigma} dx_i + f_j \int x_i^{k-1} e^{-x_i/\sigma} dx_i\right) \\
&= \text{TODO}
\end{aligned} \tag{46}$$

$$\begin{aligned}
\int y_i^2 p(y_{i*}|f_j)dy_i &= \int y_i^2 \frac{(y_i - f_i)^{k-1}}{\Gamma(k)\sigma^k} e^{-(y_i - f_i)/\sigma} dy_i \\
&= \text{TODO}
\end{aligned} \tag{47}$$

7.5 Bernoulli

With ϕ the link function and $y_i \in \{0, 1\}$, we have

$$\begin{aligned}\int y_i p(y_{i*}|f_j)dy_i &= \int y_i \phi(f_j)^{y_i}(1-\phi(f_j))^{1-y_i}dy_i \\ &= \int \phi(f_j)dy_i \\ &= \phi(f_j)\end{aligned} \tag{48}$$

$$\int y_i^2 p(y_{i*}|f_j)dy_i = \phi(f_j) \tag{49}$$

Inverse probit

$$\phi(x) = \frac{1}{2}\left(1 + \text{erf}(x/\sqrt{2})\right)$$

The integral of the error function with a Gaussian density function is known and used below.

$$\begin{aligned}\int p(y_i|f_j)q(f_j)df_j &= \int \frac{1}{2}\left(1+\text{erf}(f_j/\sqrt{2})\right)\frac{1}{\sqrt{2\pi\Sigma_{ii}}}e^{-\frac{1}{2\Sigma_{ii}}(f_j-\mu_i)^2}df_j \\ &= \frac{1}{2} + \frac{1}{2}\int \text{erf}(f_j/\sqrt{2})\frac{1}{\sqrt{2\pi\Sigma_{ii}}}e^{-\frac{1}{2\Sigma_{ii}}(f_j-\mu_i)^2}df_j \\ &= \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{\mu_i/\sqrt{2}}{\sqrt{1+2\Sigma_{ii}/2}}\right) \\ &= \phi\left(\frac{\mu_i}{\sqrt{1+\Sigma_{ii}}}\right)\end{aligned} \tag{50}$$

Logistic

$$\phi(x) = \frac{1}{1+e^{-x}}$$