Expectation Propagation

Ricardo Andrade

So far...

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{matrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{matrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right)$$

Nice properties of the Gaussian distribution:

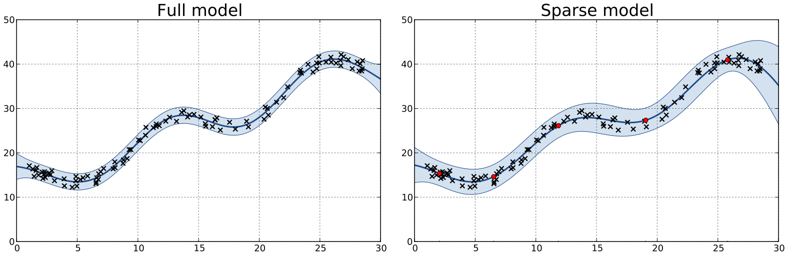
- Marginals $\mathbf{y}_1$ and $\mathbf{y}_2$ are Gaussian.
- The sum of Gaussians is Gaussian.
- The conditional of $\mathbf{y}_i$ given $\mathbf{y}_j$ is Gaussian.
- The product of Gaussians is an un-normalized Gaussian:

$$\mathbf{y}_1 \mathbf{y}_2 \sim \mathbf{Z}\mathcal{N}\left(\left(\boldsymbol{\Sigma}_{11}^{-1} + \boldsymbol{\Sigma}_{22}^{-1} \right)^{-1} \left(\boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\mu}_2 \right), \left(\boldsymbol{\Sigma}_{11}^{-1} + \boldsymbol{\Sigma}_{22}^{-1} \right)^{-1} \right)$$

So far...

Sparsity:

The covariance between any two points $\mathbf{x}_i$ and $\mathbf{x}_j$ is induced through the dependence of $\mathbf{x}_i$ and $\mathbf{x}_j$ on $\{\mathbf{z}_k\}_{k=1}^{m}$.



So far...

Our observations $\mathbf{y}$ are a distorted version of a process $\mathbf{f}$:

$$\mathbf{y} = \mathbf{f}(\mathbf{X}) + \epsilon$$

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

$$p(\mathbf{y} \,|\, \mathbf{f}) = \mathcal{N}(\mathbf{f}, \sigma^2 \mathbf{I})$$

Motivation

Simple analytical solution for Gaussian likelihoods:

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

What if the likelihood is not Gaussian?

$$\begin{aligned}
\text{Count process:} \quad & \mathbf{y} \in \mathbb{N} \\
\text{Classification:} \quad & \mathbf{y} \in \{C_1, ..., C_k\} \\
\text{Other assumptions:} \quad & \mathbf{y} \in [0, 1]
\end{aligned}$$

General case

Exact (intractable) posterior:

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

EP posterior approximation:

$$q(\mathbf{f} \,|\, \mathbf{y}) = \frac{\prod_{i=1}^{K} t_i(f_i)}{Z_{EP}}$$

Fully factorized Gaussian approximation

Consider the special case:

- $p(y_i \,|\, f_i) \approx t_i(f_i) \propto \mathcal{N}(f_i \,|\, \tilde{\mu}_i, \tilde{\sigma}_i^2)$, with $i = 1, \ldots, n$.
- $p(\mathbf{f}) \sim \mathcal{N}(\mathbf{f} \,|\, \mathbf{0}, \mathbf{K}_{nn})$. Not approximation needed.

EP posterior approximation:

$$q(\mathbf{f} \,|\, \mathbf{y}) = \frac{p(\mathbf{f}) \prod_{i=1}^{n} t(f_i)}{Z_{EP}} = \mathcal{N}(\mathbf{f} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Site approximations

Assume:

- Initial approximations given: $t_j(f_j)$ is given for $j \neq i$.
- Interest in finding $t_i(f_i) \approx p(y_i|f_i)$.

$$p(y_i|f_i)p(\mathbf{f}) \prod_{j \neq i} t_j(f_j) \approx p(\mathbf{f}) \prod_{j=1}^{n} t_j(f_j)$$

$$p(y_i|f_i) \int p(\mathbf{f}) \prod_{j \neq i} t_j(f_j) \, \mathrm{d} f_{j \neq i} \approx \int p(\mathbf{f}) \prod_{j=1}^{n} t_j(f_j) \, \mathrm{d} f_{j \neq i}$$

$$p(y_i|f_i) q_{-i}(f_i) \approx \mathcal{N}(f_i \,|\, \hat{\mu}_i, \hat{\sigma}_i^2) \hat{Z}_i$$

Minimization of the KL divergence

$$\min \mathrm{KL}\left(p(y_i|f_i)q_{-i}(f_i) \| \mathcal{N}(f_i \mid \hat{\mu}_i, \hat{\sigma}_i^2)\hat{Z}\right)$$

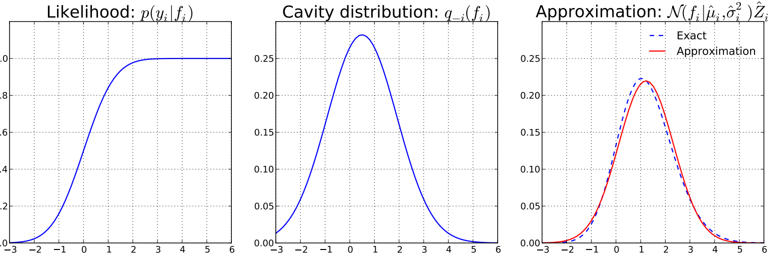
Since the approximation is Gaussian, KL is minimal when:

- $\hat{\mu}_i = \langle f_i \rangle_{p(y_i|f_i)q_{-i}(f_i)}$
- $\hat{\sigma}_i^2 = \langle f_i \rangle^2_{p(y_i|f_i)q_{-i}(f_i)} - \tilde{\mu}_i^2$

Since the approximation is un-normalized, we need that:

- $\hat{Z}_i = \int p(y_i|f_i)q_{-i}(f_i)\,\mathrm{d}f_i$

Site approximation example



Predictions

Predictive distribution of $q(f_* \,|\, \mathbf{y})$ is also Gaussian:

- $\langle f_* \,|\, \mathbf{y} \rangle_{q(f_* \,|\, \mathbf{y})} = \mathbf{k}_*^\top \left(\mathbf{K}_{nn} + \tilde{\boldsymbol{\Sigma}} \right)^{-1} \tilde{\boldsymbol{\mu}}$
- $\langle f_*^2 \,|\, \mathbf{y} \rangle_{q(f_* \,|\, \mathbf{y})} - \langle f_* \,|\, \mathbf{y} \rangle^2_{q(f_* \,|\, \mathbf{y})} = k_{**} - \mathbf{k}_*^\top \left(\mathbf{K}_{nn} + \tilde{\boldsymbol{\Sigma}} \right)^{-1} \mathbf{k}_*$

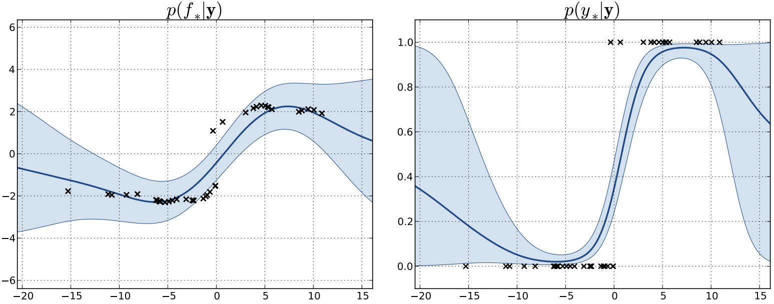
Predictive distribution of y_* might still be intractable:

$$q(y_* | f_*) = \int p(y_* | f_*) q(f_* | \mathbf{y}) \, \mathrm{d} f_*$$

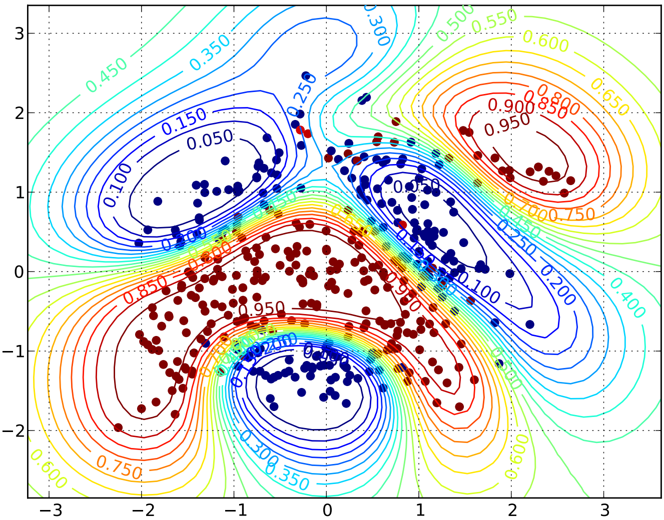
Example 1: binary classification

Let y_i be a binary variable for $i = 1, \ldots, n$.
Then we can define $p(y_i = 1 \mid f_i)$ with a squashing function over f_i, i.e.:

- $p(y_i = 1 \mid f_i) = \frac{1}{1+e^{-f_i}}$
- $p(y_i = 1 \mid f_i) = \Phi(f_i)$



Example 2: banana data set



Posterior moments update

Complexity is dominated by the computation of the posterior covariance $\mathbf{\Sigma} = \left(\mathbf{K}_{nn}^{-1} + \tilde{\mathbf{\Sigma}}^{-1}\right)^{-1}$:

- Rank-one updates are possible after a careful re-formulation: $\mathcal{O}(n^2)$ per factor.

Sparse EP

$q(\mathbf{f}|\mathbf{y})$ is computed as before, but an sparse approximation is used instead of the exact covariance $\mathbf{K}_{nn}$.

FITC approximation: $\mathcal{O}(nm^2)$

$$\mathbf{K}_{nn} \approx \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn} + \mathrm{diag}(\mathbf{K}_{nn} - \mathbf{Q}_{nn})$$

DTC approximation: $\mathcal{O}(nm^2)$

$$\mathbf{K}_{nn} \approx \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}$$

EP-FITC (aka generalized FITC)

Predictions now depend on $\mathbf{u}$:

- $q(f_* \mid \mathbf{y}) = \int p(f_* \mid \mathbf{u}) q(\mathbf{u} \mid \mathbf{y}) \, \mathrm{d}\mathbf{u}$
- $q(y_* \mid \mathbf{y}) = \int q(y_* \mid f_*) q(f_* \mid \mathbf{y}) \, \mathrm{d}f_*$

The following is needed:

$$p(\mathbf{u} \mid \mathbf{f}) \propto p(\mathbf{f} \mid \mathbf{u}) p(\mathbf{u})$$

$$q(\mathbf{u} \mid \mathbf{y}) = \int p(\mathbf{u} \mid \mathbf{f}) q(\mathbf{f} \mid \mathbf{y}) \, \mathrm{d}\mathbf{f}$$

EP-DTC

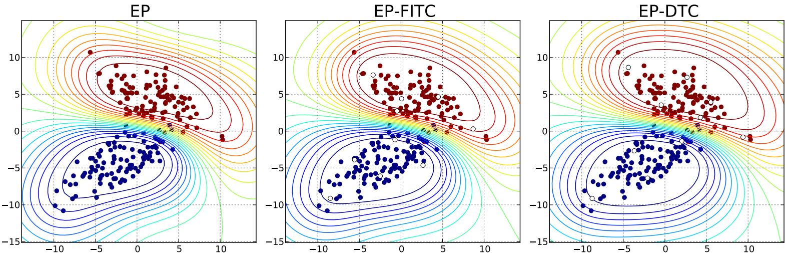
Compatible with sparse variational approach:

$$\mathcal{L} = \log \mathcal{N}\left(\tilde{\boldsymbol{\mu}}|0, \mathbf{Q}_{nn} + \tilde{\boldsymbol{\Sigma}}\right) - \frac{1}{2}\operatorname{Tr}\left((\mathbf{K}_{nn} - \mathbf{Q}_{nn})\tilde{\boldsymbol{\Sigma}}^{-1}\right) - Z_{EP}.$$

Penalty term vs. diagonal term inside the covariance:

- Updates are simpler than in EP-FITC.
- As in the regression case, optimization of $\mathbf{Z}$ is simpler than in FITC.

EP variants



References

[1] Thomas Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc., 2001.

[2] Andrew Naish-Guzman and Sean Holden. The generalized FITC approximation. *Advances in Neural Information Processing Systems*, 20:1057–1064, 2008.

[3] Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate Gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, 2005.

[4] Matthias Seeger. Expectation propagation for exponential families. Technical report, University of California at Berkeley, 2005.

[5] Michalis K. Titsias. Variational learning of inducing variables in sparse Gaussian processes. *Journal of Machine Learning Research*, 5:567–574, 2009.

[6] Christopher K. I. Williams and Carl Edward Rasmussen. *Gaussian processes for Machine Learning*. MIT Press, 2006.