Expectation Propagation

Ricardo Andrade Pacheco

Gaussian Process Winter School
14 January 2014

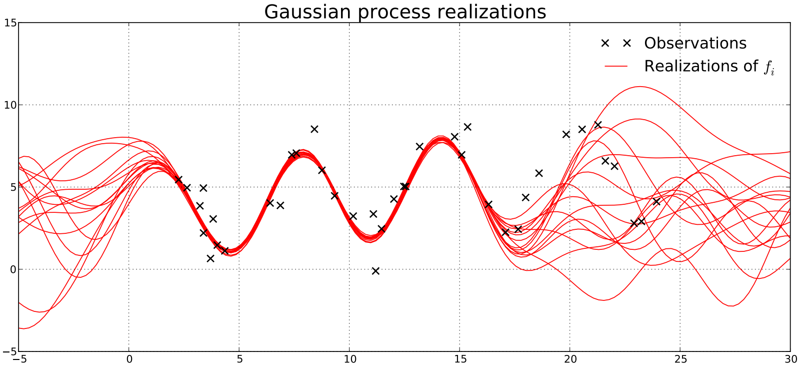
Outline

- Motivation
- Expectation propagation
- Sparse expectation propagation

GP Regression

Observations y_i are a distorted version of a process f_i:

$$y_i = f_i(\mathbf{x}_i) + \epsilon_i, \text{ with } \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$



GP Regression

Analytical tractability of the posterior distribution is assured:

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

In this talk

Assume Gaussian assumption is not longer adequate, e.g.:

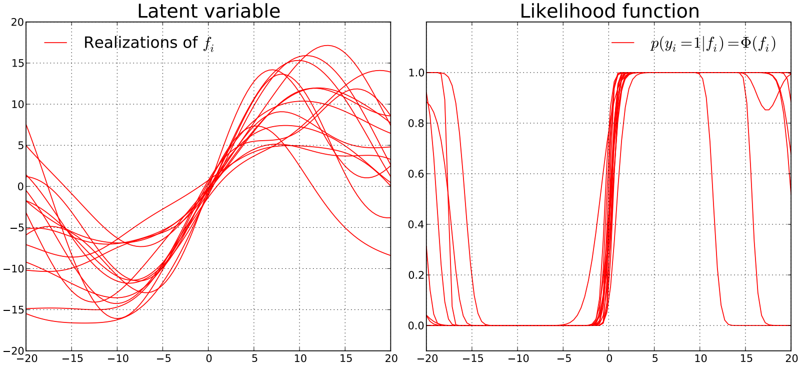
Classification:	$\mathbf{y} \in \{C_1, ..., C_k\}$
Count process:	$\mathbf{y} \in \mathbb{N}$
Other assumptions:	$\mathbf{y} \in [0, 1]$

Example: binary classification

- We are interested in modelling binary outcomes.
- Assume:

$$y_i = \begin{cases} 1, & \text{with probability } p_i \\ 0, & \text{with probability } 1 - p_i \end{cases}$$

- Model $p(y_i|f_i)$ as a monotonic tranformation of f_i:



Non-linear response functions

Non-Gaussian likelihood:

$$p(y_i|f_i) = \Phi(f_i)$$

Exact computation of the posterior is no longer possible analytically.

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

EP: 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:

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

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

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

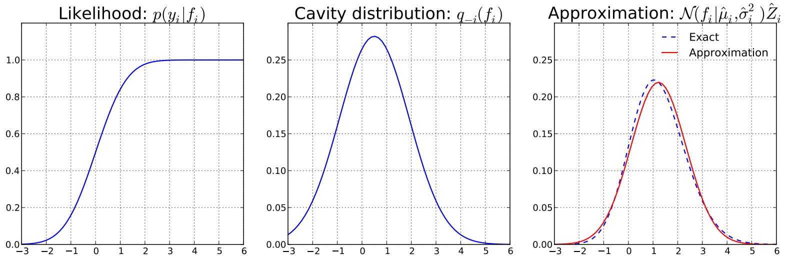
Since the approximation is Gaussian, KL is minimal when:

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

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

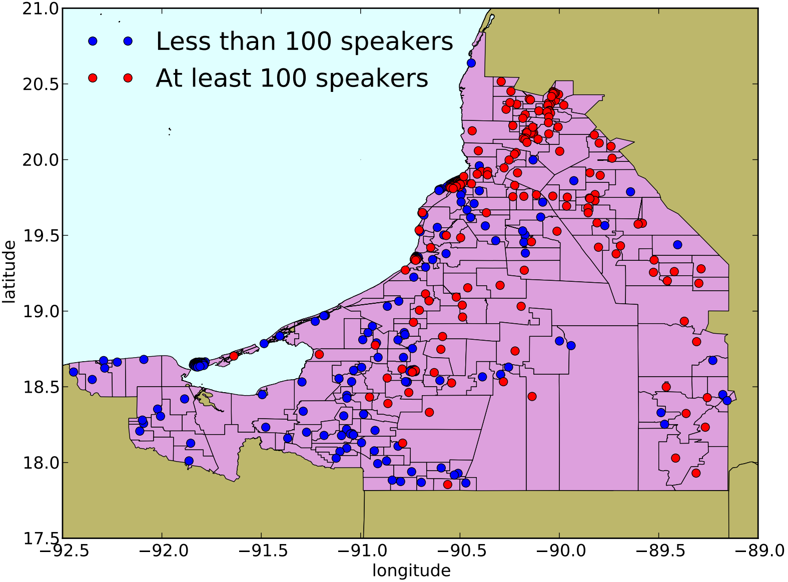
Predictive distribution of $q(f_* \,|\, \mathbf{y})$ is also Gaussian:

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

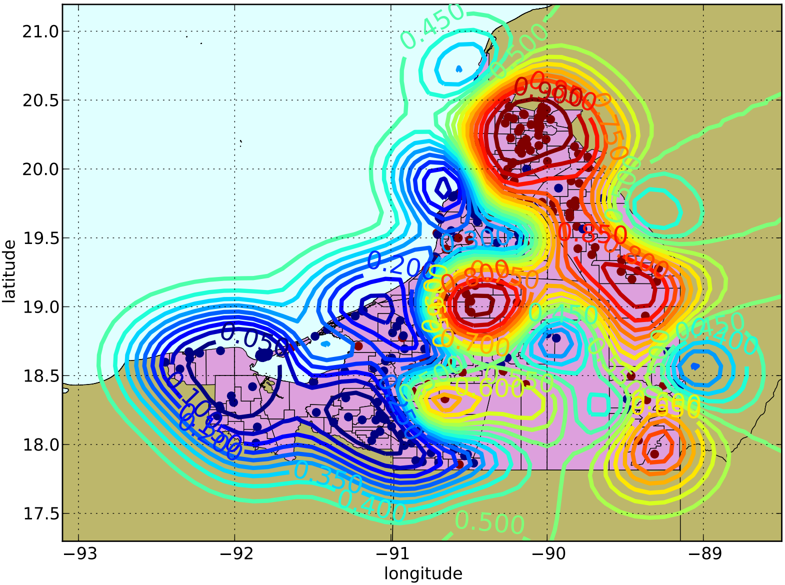
Predictive distribution of y_* might still be intractable:

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

Example: People who speak an indigenous language



Example: People who speak an indigenous language



Posterior variance update

Complexity is dominated by the computation of the posterior covariance:

$$\boldsymbol{\Sigma} = \left(\mathbf{K}_{nn}^{-1} + \tilde{\boldsymbol{\Sigma}}^{-1}\right)^{-1}$$

Sparse EP

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

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

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

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

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

EP-FITC (generalized FITC)

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

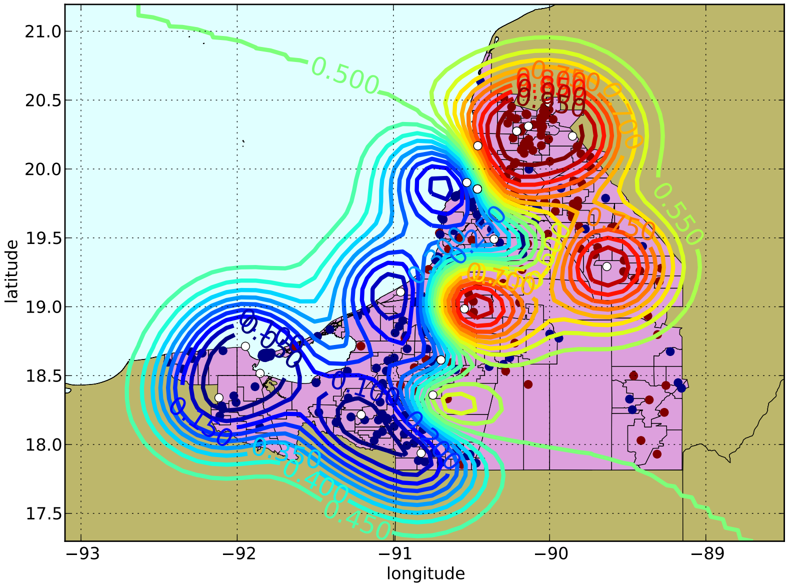
- $q(f_* \mid \mathbf{y}) = \int p(f_* \mid \mathbf{u}) q(\mathbf{u} \mid \mathbf{y}) \, \mathrm{d}\mathbf{u}$
- $q(y_* \mid \mathbf{y}) = \int q(y_* \mid f_*) q(f_* \mid \mathbf{y}) \, \mathrm{d}f_*$

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-FITC (generalized FITC)

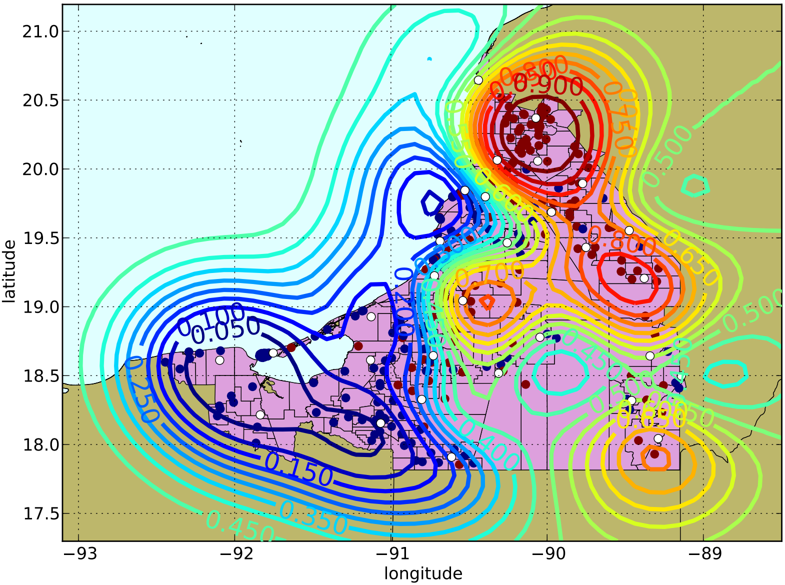


EP-DTC

Compatible with sparse variational approach:

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

Sparse variational + EP-DTC



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.