

Expectation Propagation

Ricardo Andrade Pacheco



Gaussian Process Winter School
16 September 2014

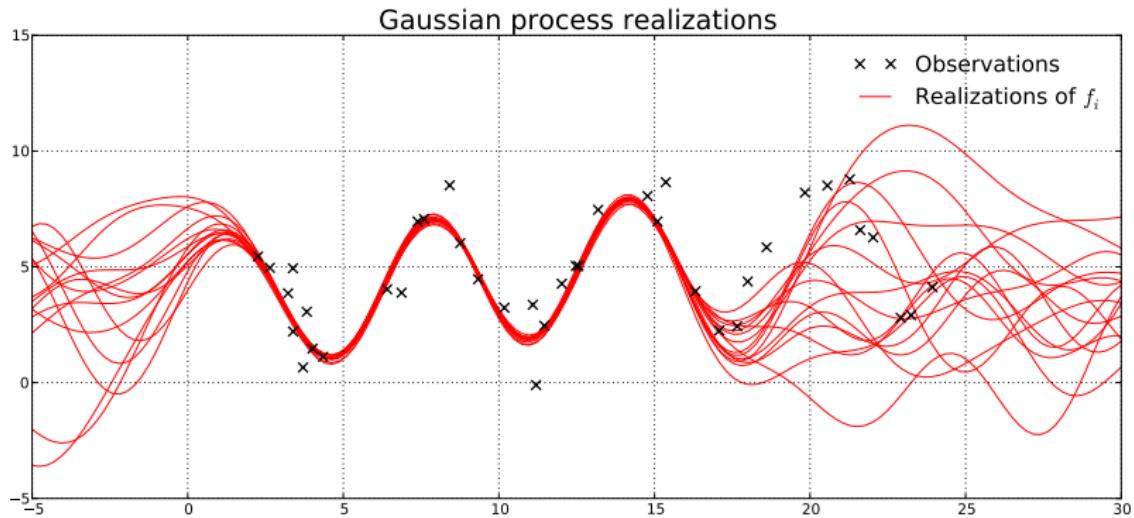
Outline

- ▶ 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:

$$\text{Gaussian prior: } \mathbf{f} \sim \mathcal{GP}(\mathbf{m}, \mathbf{K}_{\mathbf{ff}})$$

$$\text{Gaussian likelihood: } \prod_{i=1}^n p(y_i | f_i) \sim \mathcal{N}(\mathbf{y} | \mathbf{f}, \sigma_i^2 \mathbf{I})$$

$$\text{Gaussian posterior: } p(\mathbf{f} | \mathbf{y}) \propto \mathcal{N}(\mathbf{f} | \mathbf{m}, \mathbf{K}_{\mathbf{ff}}) \mathcal{N}(\mathbf{y} | \mathbf{f}, \sigma_i^2 \mathbf{I})$$

In this talk

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

Classification: $y \in \{C_1, \dots, C_k\}$

Count process: $y \in \mathbb{N}$

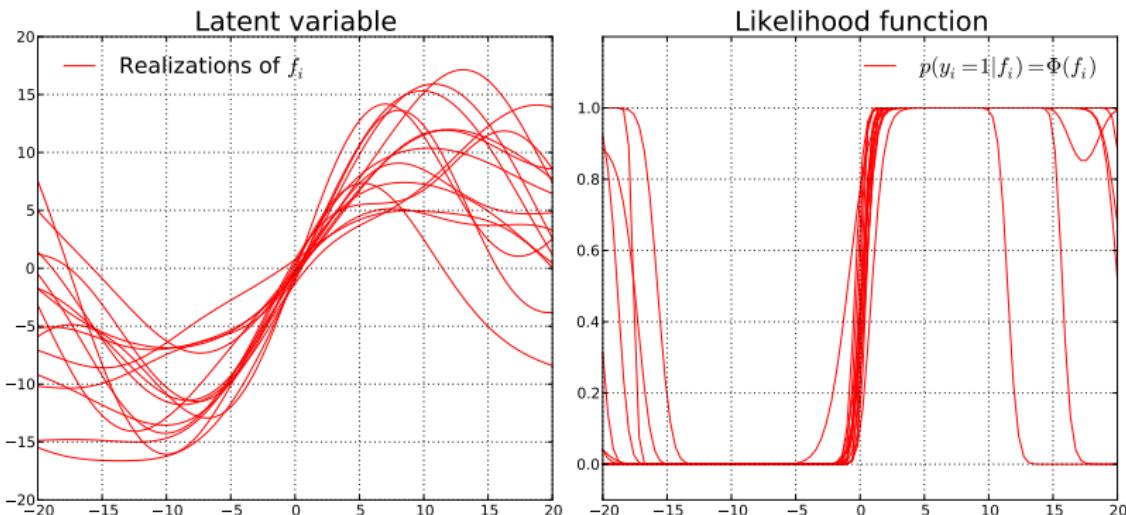
Other assumptions: $y \in [0, 1]$

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 transformation 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.

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

EP: general case

Exact (intractable) posterior:

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

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 $i = 1, \dots, n$.
- ▶ $p(\mathbf{f}) \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{\mathbf{ff}})$. No 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 $t_i(f_i) \approx p(y_i|f_i)$.

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

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

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

Minimization of the KL divergence

$$\hat{\mu}_i, \hat{\sigma}_i = \operatorname{argmin}_{\hat{\mu}_i, \hat{\sigma}_i} \text{KL} \left(p(y_i | f_i) q_{\setminus i}(f_i) \| \mathcal{N} (f_i | \hat{\mu}_i, \hat{\sigma}_i^2) \hat{Z} \right)$$

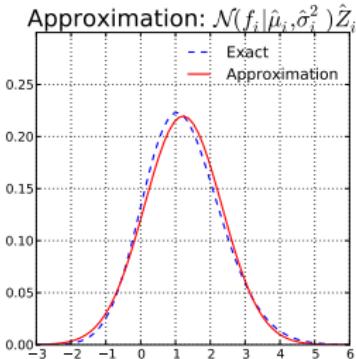
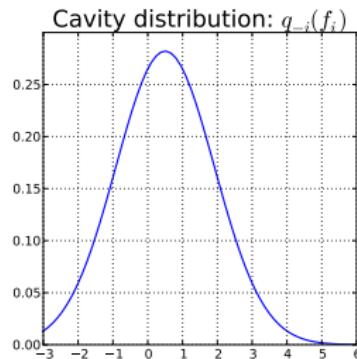
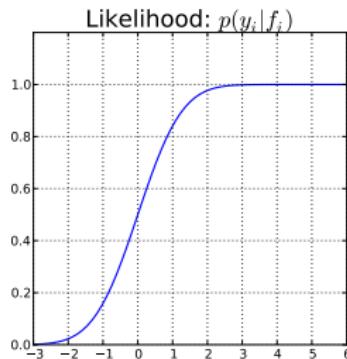
Since the approximation is Gaussian, KL is minimal when:

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

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

- ▶ $\hat{Z}_i = \int p(y_i | f_i) q_{\setminus i}(f_i) df_i$

Site approximation example



Predictions

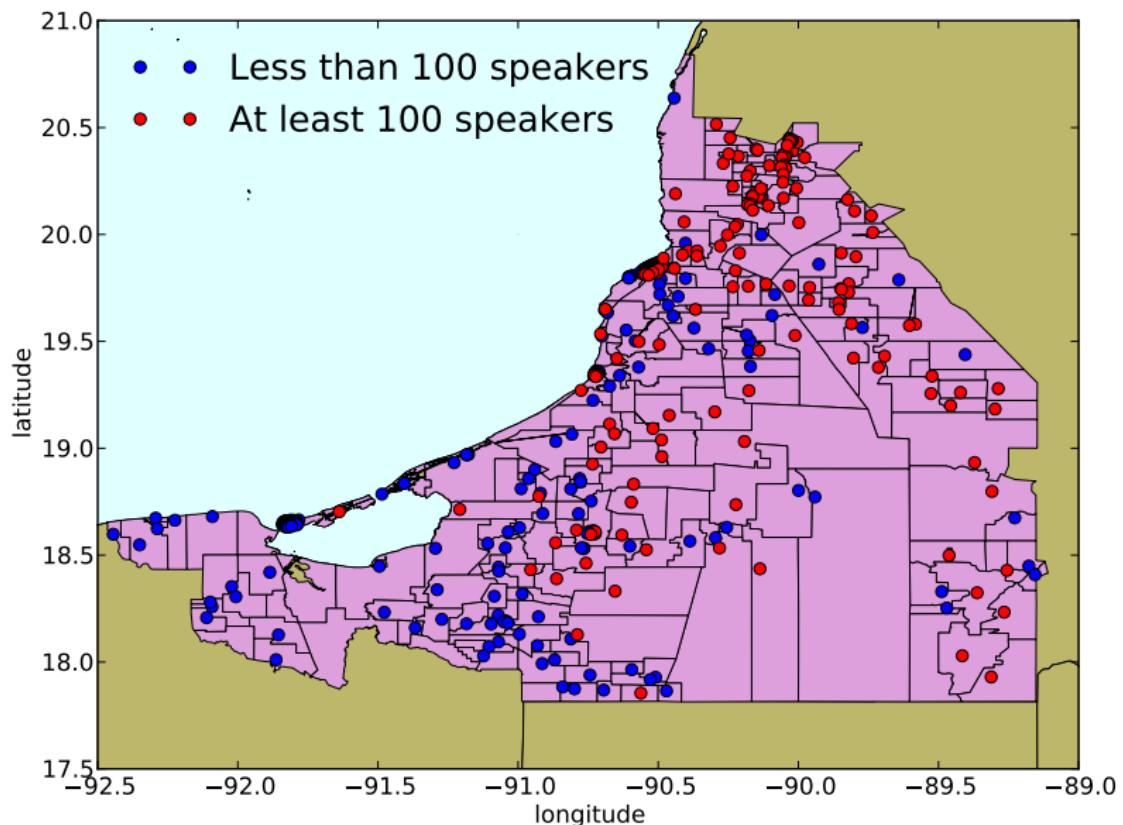
Predictive distribution of $q(f_*|\mathbf{y})$ is also Gaussian:

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

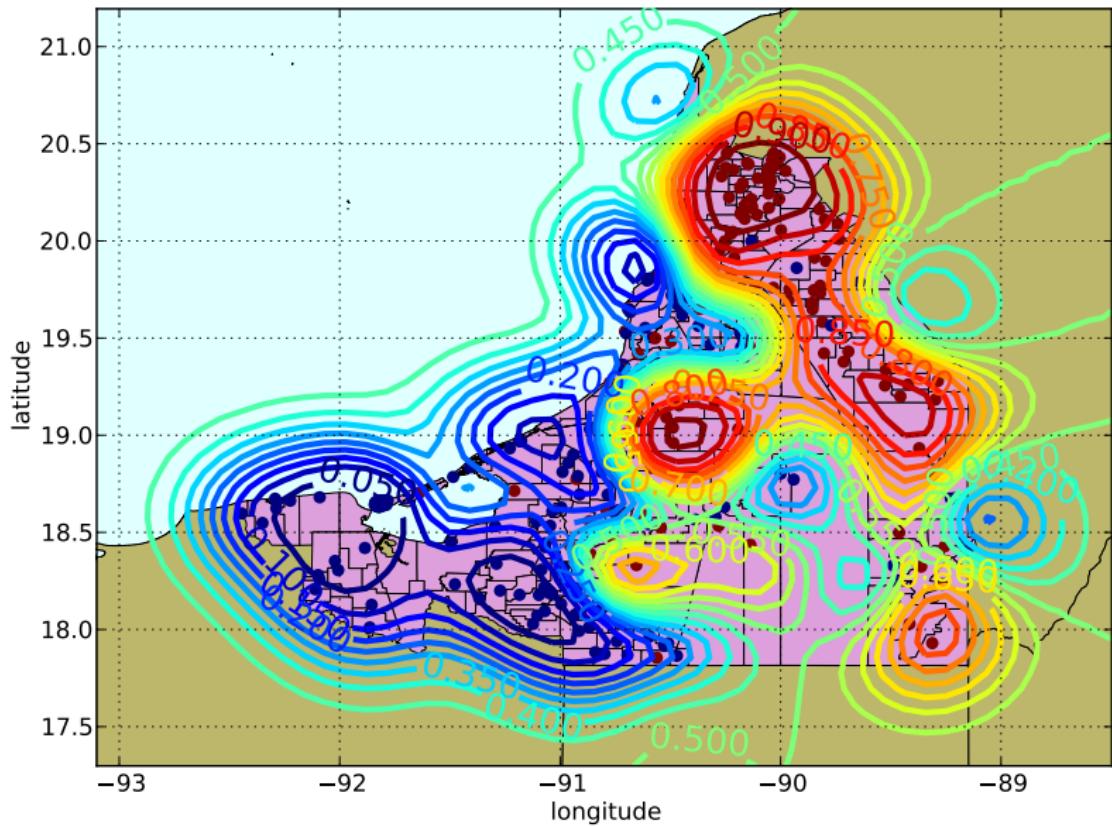
Predictive distribution of y_* might still be intractable:

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

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:

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

Sparse EP

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

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

$$\mathbf{K}_{\mathbf{ff}} \approx \mathbf{K}_{\mathbf{fu}} \mathbf{K}_{\mathbf{uu}}^{-1} \mathbf{K}_{\mathbf{uf}} + \text{diag}(\mathbf{K}_{\mathbf{ff}} - \mathbf{Q}_{\mathbf{ff}})$$

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

$$\mathbf{K}_{\mathbf{ff}} \approx \mathbf{K}_{\mathbf{fu}} \mathbf{K}_{\mathbf{uu}}^{-1} \mathbf{K}_{\mathbf{uf}}$$

EP-FITC (generalized FITC)

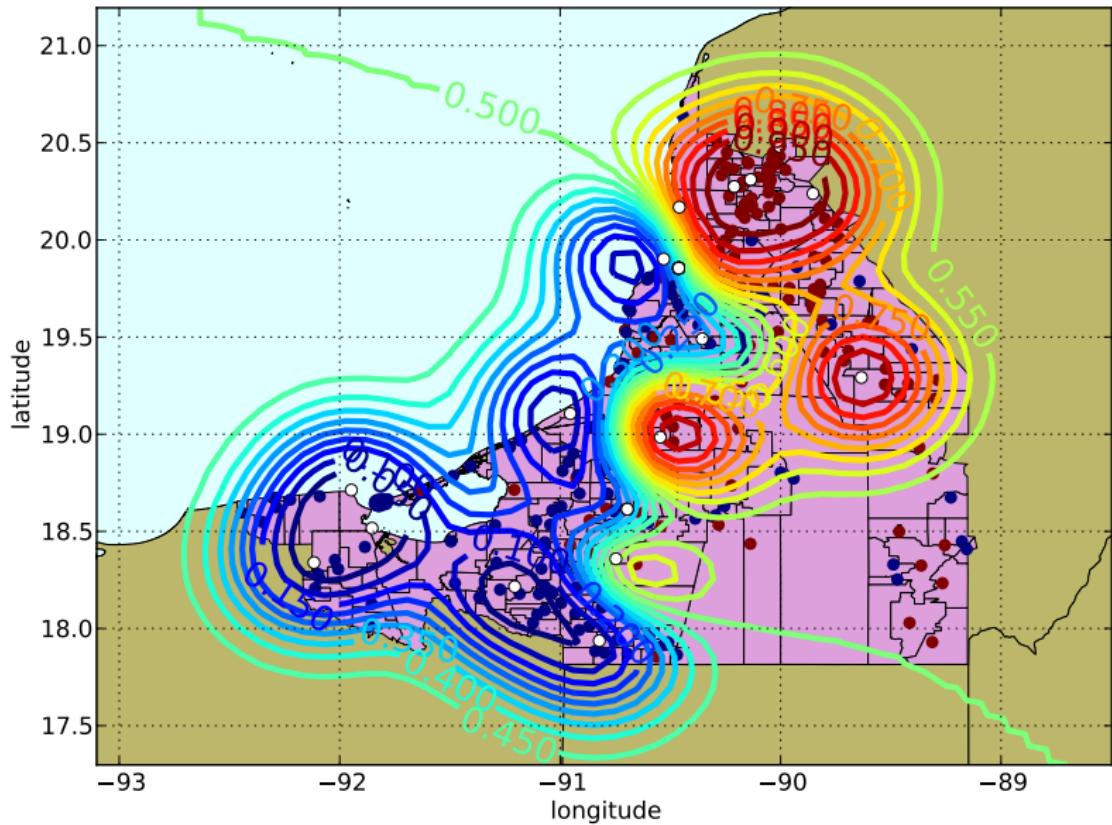
Predictions now depend on \mathbf{u} :

- ▶ $q(f_*|\mathbf{y}) = \int p(f_*|\mathbf{u})q(\mathbf{u}|\mathbf{y})d\mathbf{u}$
- ▶ $q(y_*|f_*) = \int q(y_*|f_*)q(f_*|\mathbf{y})df_*$

The following is needed:

$$p(\mathbf{u}|\mathbf{f}) \propto p(\mathbf{f}|\mathbf{u})p(\mathbf{u})$$
$$q(\mathbf{u}|\mathbf{y}) = \int p(\mathbf{u}|\mathbf{f})q(\mathbf{f}|\mathbf{y})d\mathbf{f}$$

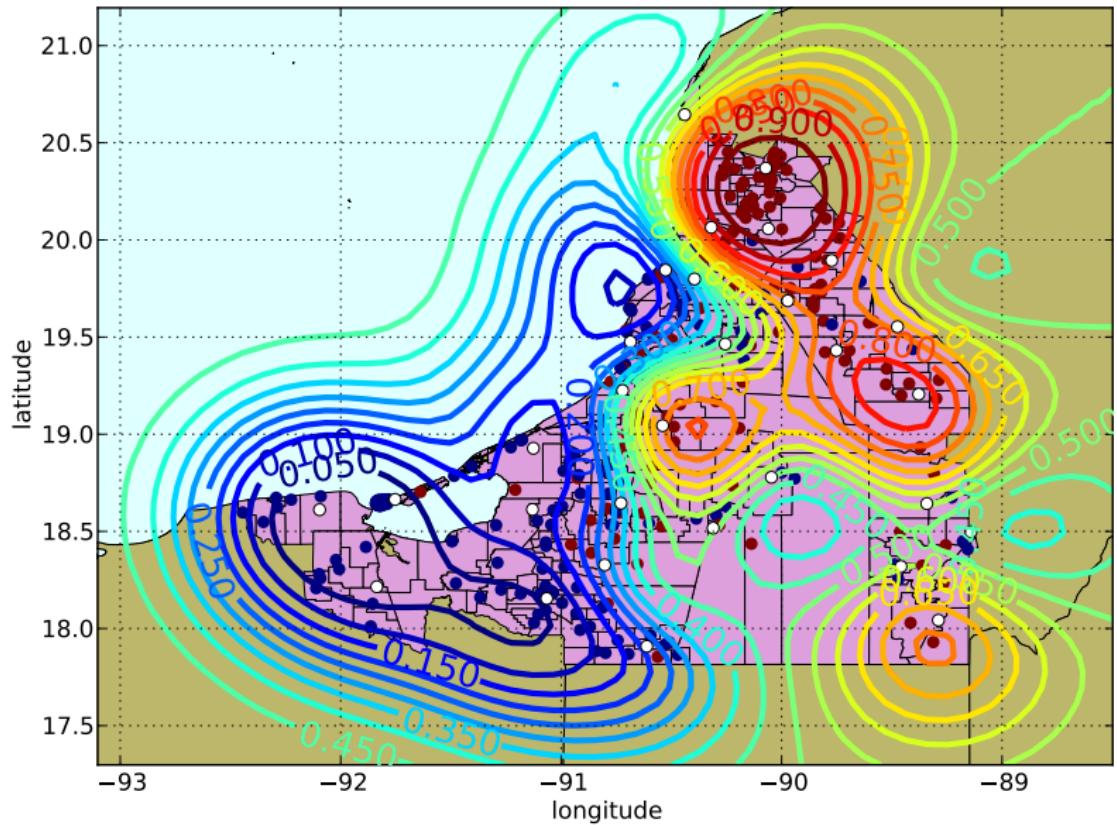
EP-FITC (generalized FITC)



Compatible with sparse variational approach:

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

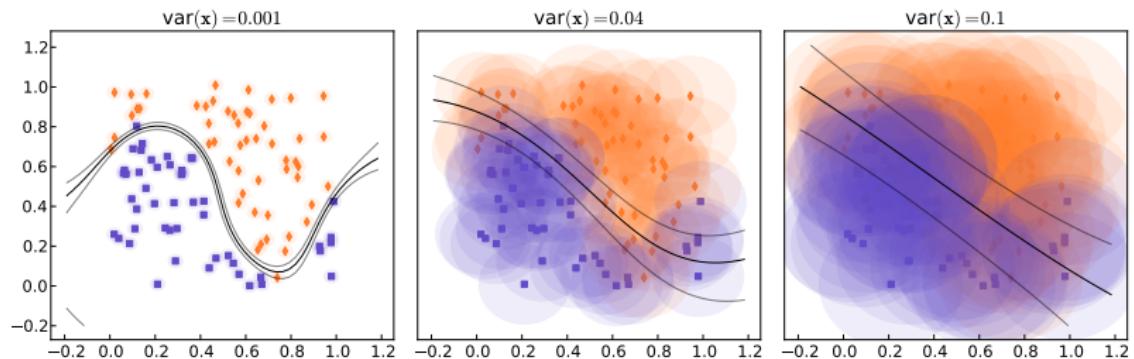
Sparse variational + EP-DTC



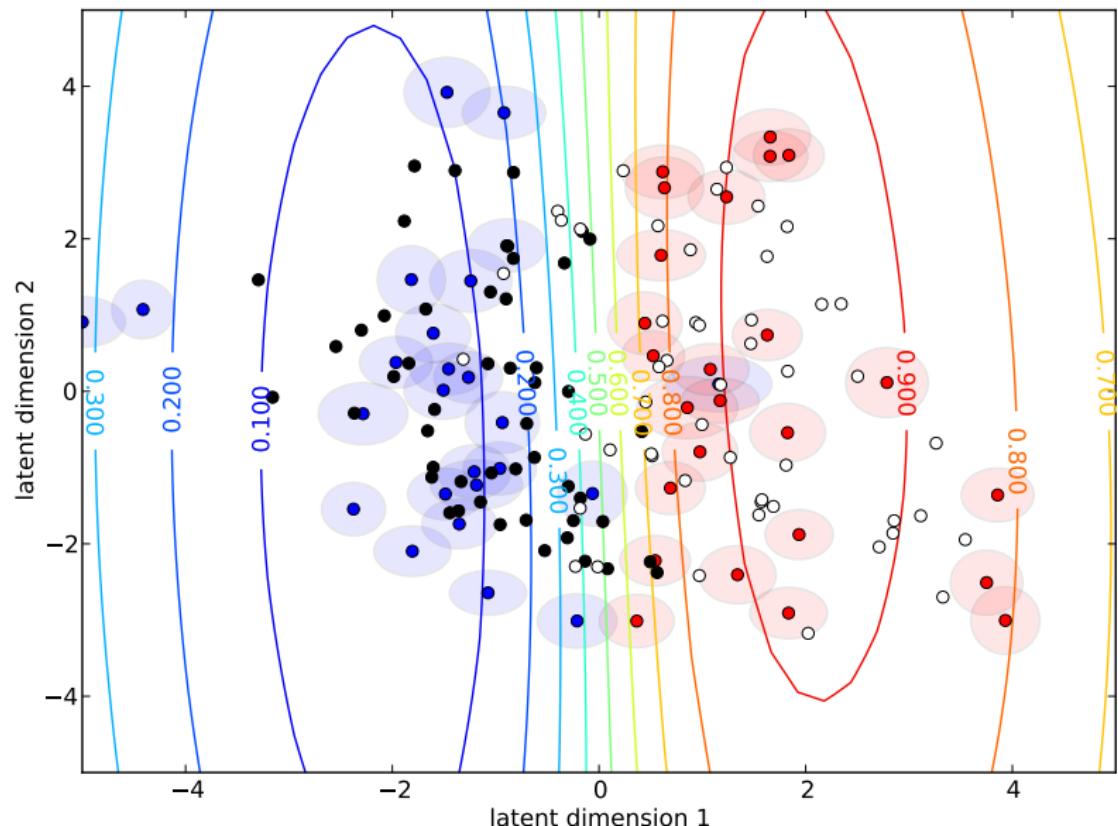
Going further

- ▶ Use a probability distribution on \mathbf{X} .
- ▶ Dimensionality reduction of non-Gaussian data.

Classification with uncertain inputs



Classification in a lower dimensional space (USPS: 3 vs 5)



References

- [1] Ricardo Andrade-Pacheco, James Hensman, and Neil D. Lawrence. Hybrid discriminative-generative approaches with Gaussian processes. In *Proceedings of the Seventeenth International Workshop on Artificial Intelligence and Statistics*, volume 33. JMLR W&CP 33, 2014.
- [2] 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.
- [3] Andrew Naish-Guzman and Sean Holden. The generalized FITC approximation. *Advances in Neural Information Processing Systems*, 20:1057–1064, 2008.
- [4] National Institute of Statistics and Geography (INEGI). Population census, 2010.
- [5] 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.
- [6] Matthias Seeger. Expectation propagation for exponential families. Technical report, University of California at Berkeley, 2005.
- [7] Michalis K. Titsias. Variational learning of inducing variables in sparse Gaussian processes. *Journal of Machine Learning Research*, 5:567–574, 2009.
- [8] Christopher K. I. Williams and Carl Edward Rasmussen. *Gaussian processes for Machine Learning*. MIT Press, 2006.