VARIATIONAL INFERENCE
in Gaussian process models

James Hensman
Approximate Inference workshop, NIPS 2015
Lancaster University
A unified view of variational GP approximations

- Deals with non-Gaussian posterior
- Deals with $\mathcal{O}(n^3)$ complexity (sparse)
- The variational distribution contains a (conditionally) Gaussian process
\[ f(x) \sim \mathcal{GP}(0, k(x, x')) \]

\[ f \sim \mathcal{N}(0, K) \]

with:

\[ K_{i,j} = k(x_i, x_j) \]

\[ y_i | f_i \sim \text{Po}(y_i | e^{f_i}) \quad \text{or} \quad \text{Bin}(y_i | \sigma(f_i)) \quad \text{or} \ldots \]
DEALING WITH NON-CONJUGACY

- Local variational bounds (classification only) \(^1\)
- Expectation Propagation \(^2\)
- For classification, EP > VB \(^3\)
- Variational methods need only 2N parameters \(^4\)
- VB methods can be fast too! \(^5\)
- VB can be applied to lots of different likelihoods \(^6\)

---

\(^1\) MN Gibbs, DJC MacKay - Variational Gaussian process classifiers - IEEE TNN 2000
\(^3\) H Nickisch, CE Rasmussen - Approximations for binary Gaussian process classification - JMLR 2008
\(^4\) M. Opper and C. Archambeau – The variational Gaussian approximation revisited - Neural comp. 2009
\(^5\) E Khan, S Mohamed, KP Murphy - Fast Bayesian inference for non-conjugate Gaussian process regression- NIPS 2012
\(^6\) Nguyen and Bonilla – Automated variational inference for Gaussian process models - NIPS 2014
DEALING WITH $o(n^3)$ COMPLEXITY

- Subset-of-data methods\(^7\) \(^8\) hence ‘sparse’.
- Pseudo-inputs introduced \(^9\)
- A unifying view brings several ideas together \(^10\)
- Variational approach \(^11\) makes for better placement of pseudo/inducing points
- Variational approach can be optimized with SVI \(^12\)

\(^7\) AJ Smola, P Bartlett - Sparse greedy Gaussian process regression - NIPS 2001
\(^8\) M Seeger, C Williams - Fast forward selection to speed up sparse Gaussian process regression - AISTATS 2003
\(^9\) E Snelson, Z Ghahramani - Sparse Gaussian processes using pseudo-inputs - NIPS 2005
\(^10\) Quiñonero-Candela, CE Rasmussen - A unifying view of sparse approximate Gaussian process regression - JMLR 2005
\(^11\) M. Titsias - Variational learning of inducing variables in sparse Gaussian processes - AISTATS 2009
\(^12\) J. Hensman, N. Fusi and N. Lawrence - Gaussian Processes for Big Data - UAI 2013
A GRAPHICAL MODEL FOR GAUSSIAN PROCESSES

\[
\theta \sim p(\theta) \\
\mathbf{f}(x) \sim \mathcal{GP}(0, k(x, x'; \theta)) \\
\mathbf{f} = [f(x_1), f(x_2), \ldots, f(x_n)]^T \\
y_n \sim p(y_n | f(x_n))
\]
\[ \theta \sim p(\theta) \]
\[ f|\theta \sim \mathcal{N}(0, K) \]
\[ y_n \sim p(y_n | f(x_n)) \]
\[ f^*(x)|f, \theta \sim \mathcal{GP}(a(x)^T f, b(x, x')) \]
A different graphical model for Gaussian processes

\[
\begin{align*}
\theta & \sim p(\theta) \\
\mathbf{f} | \theta & \sim \mathcal{N}(0, \mathbf{K}) \\
y_n & \sim p(y_n | \mathbf{f}(x_n))
\end{align*}
\]
\[ \theta, u \sim q(\theta, u) \]

\[ f^*(x) \sim \mathcal{GP}(a'(x)^T u, b'(x)) \]
KL DIVERGENCE BETWEEN GAUSSIAN PROCESSES?

Intuitive version: $\mathbf{f}^*$ is a really long vector containing all points of interest.

Rigorous version: Matthews et al.$^{13}$

---

Let’s ignore $\theta$ for now

$\infty$

$f^*(x) \rightarrow u \rightarrow z$

$\infty$

$y \rightarrow f^*(x) \rightarrow x$
Where are the \( f \) in the approximation?
Where are the $u$ in the model?
\[ \text{ELBO} = \mathbb{E}_{q(f^*, f, u, \theta)} \left[ \log \frac{p(y | f)p(f | u, \theta)p(f^* | f, u, \theta)p(u | \theta)p(\theta)}{q(f | u, \theta)q(f^* | f, u, \theta)q(u | \theta)q(\theta)} \right] \]
\[
\text{ELBO} = \mathbb{E}_{q(f^*, f, u, \theta)} \left[ \log \frac{p(y | f)p(f | u, \theta)p(f^* | f, u, \theta)p(u | \theta)p(\theta)}{q(f | u, \theta)q(f^* | f, u, \theta)q(u | \theta)q(\theta)} \right]
\]
\[
\text{ELBO} = \mathbb{E}_{\mathbf{q}(f^*, f, u, \theta)} \left[ \log \frac{p(y | f)p(f | u, \theta)p(f^* | f, u, \theta)p(u | \theta)p(\theta)}{q(f | u, \theta)q(f^* | f, u, \theta)q(u | \theta)q(\theta)} \right]
\]
Strategy 1: Gaussian \(^{14}\)

Let \(q(u, \theta) = \mathcal{N}(u|m, LL^T)\delta(\theta - \hat{\theta})\)

Optimize wrt \(m, L, \hat{\theta}\) (and \(Z\!\)?)

Strategy 2: Free-form \(^{15}\)

Given the limited size of \(Z\) (and thus \(u\)), write down the optimal, intractable, form for \(q(u, \theta)\), and sample from it using HMC.

---

\(^{14}\) Hensman, A Matthews, Z Ghahramani - Scalable Variational Gaussian Process Classification - AISTATS 2015

\(^{15}\) Hensman, AGG Matthews, M Filippone - MCMC for Variationally Sparse Gaussian Processes - NIPS 2015
The objective function (which minimizes the KL between the q-process and the p-process) is

$$\mathcal{L} = \sum_i \mathbb{E}_{q(f_i)}[\log p(y_i|f_i)] - KL[q(u)||p(u)]$$
Left: three k-means centers used to initialize the inducing point positions. Center: the positions of the same inducing points after optimization. Right: difference.

Data: N=60,000, D=784

Accuracy: 98.04%
The ‘perfect’ distribution $\hat{q}(u, \theta)$ which minimises the KL divergence (with no further restrictions) is

$$\log \hat{q}(u, \theta) = \mathbb{E}_{p(f | u)}[\log p(y | f)] + \log p(u, \theta) + \text{const.}$$

Sampling $\hat{q}$ costs $\mathcal{O}(NM^2)$. 
The posterior of the rates for the coal mining disaster data.
VB+MCMC

MCMC

lengthscale

variance
THE EFFECT OF INDUCING POINTS SELECTION

![Graph showing the effect of inducing points selection. The y-axis represents \( \log p(y^*)_{[MCMC]} \) and the x-axis represents the number of inducing points. Two sets of data points are shown, labeled \( Z_{\text{optimized}} \) and \( Z_{\text{k-means}} \).]
SPECIAL CASES AND GENERALIZATIONS

- Exact inference (Gaussian likelihood, $\mathbf{Z} = \mathbf{X}$)
- Subset-of-data methods (e.g. IVM $^{16}$)
- Inter-domain approximations $^{17}$
- Black box likelihoods $^{18}$
- Log Gaussian Cox processes $^{19}$

$^{16}$Lawrence, Seeger and Herbrich - The Informative Vector Machine - NIPS 2003
$^{17}$Alvarez, Rosasco and Lawrence - Kernels for vector valued functions, a review - foundational trends in ML 2011
$^{18}$Dezfouli and Bonilla - Gaussian Process Models with Black-Box Likelihoods - NIPS 2015
$^{19}$Lloyd et al - Variational Inference for Gaussian Process Modulated Poisson Processes - ICML 2015
THANKS FOR LISTENING