STAT G8325
Gaussian Processes and Kernel Methods
Lecture Notes §05: Speed and Scaling

John P. Cunningham

Department of Statistics
Columbia University

Outline

Administrative interlude

Practical realities of kernel methods

Inducing point methods [QCR05]; [SG07]

Variational inference

Variational inducing point methods [Tit09]

Outline

Administrative interlude

Practical realities of kernel methods

Inducing point methods [QCR05]; [SG07]

Variational inference

Variational inducing point methods [Tit09]

Progress...

Week	Lectures	Content
4	Oct 5,7	Kernels
5	Oct 12,14	Speed and scaling part 1: reduced-rank processes
		• Reading: [QCR05]; [SG07]; [Tit09]
		• Optional additional reading: [GT15]; [RW06, ch. 8]; [TLG14]
6	Oct 19	Speed and scaling part 2: special structure

- Project brainstorming list available on courseworks.
- Make an appointment with me in the next week.
- Homeworks will become more and more project oriented.

Outline

Administrative interlude

Practical realities of kernel methods

Inducing point methods [QCR05]; [SG07]

Variational inference

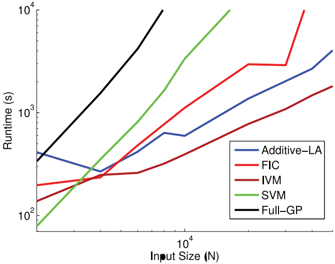
Variational inducing point methods [Tit09]

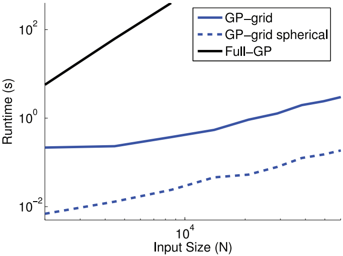
Fundamental fact about nonparametric techniques

- The number of parameters grows with the amount of data.
- In gp (and mostly in kernel methods):

$$y^*|y \sim \mathcal{N}\left(K_{y^*y}K_{yy}^{-1}(y - m_y) \;\;,\;\; K_{y^*y^*} - K_{y^*y}K_{yy}^{-1}K_{y^*y}^{\top}\right)$$

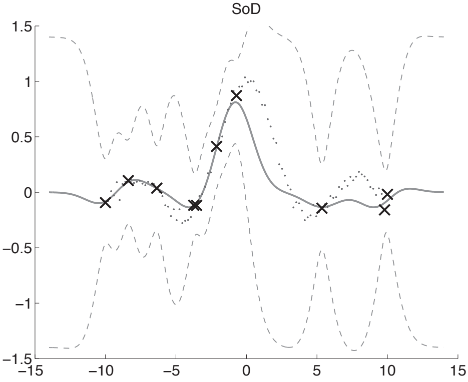
- Storing and inverting $K_{yy} \in \mathbb{R}^{n \times n}$ costs $\mathcal{O}(n^2)$ and $\mathcal{O}(n^3)$.
- Practically speaking, these operations become impossible fast:





The simplest idea: subset of data

- If $\mathcal{O}(n^3)$ is too big, randomly choose $m < n$ points, and proceed.
- Unsurprisingly, this technique does not work particularly well.



- Notice that an item of focus will be the uncertainty estimates...

Outline

Administrative interlude

Practical realities of kernel methods

Inducing point methods [QCR05]; [SG07]

Variational inference

Variational inducing point methods [Tit09]

Inducing points

- To proceed, we define a set of inducing points $u = [u_1, ..., u_m]$.
- These are jointly gaussian with the latent gp f, such that:

$$p(f_*, f) = \int p(f, f_*, u) du = \int p(f, f_*|u) p(u) du \quad , \quad \text{where } u \sim \mathcal{N}(0, K_{uu}),$$

- We are interested in the usual things, like the posterior:

$$f_*|y \sim \mathcal{N}\left(K_{f_* f}(K_{ff} + \sigma_\epsilon^2 I)^{-1} y \quad , \quad K_{f_* f_*} - K_{f_* f}(K_{ff} + \sigma_\epsilon^2 I)^{-1} K_{f f_*}\right)$$

- The **critical** conditional independence assumption:

$$p(f_*, f) \approx q(f_*, f) = \int q(f_*|u) q(f|u) p(u) du$$

- *Training and test points are conditionally independent, given inducing points.*

Inducing points

$$p(f_*, f) \approx q(f_*, f) = \int q(f_*|u)q(f|u)p(u)du$$

- *Training and test points are conditionally independent, given inducing points.*
- u induce dependency between the training and test latents f and f_*, where:

$$\begin{aligned} p(f|u) &= \mathcal{N}\left(K_{fu}K_{uu}^{-1}u \;\;,\;\; K_{ff} - K_{fu}K_{uu}^{-1}K_{uf}\right) \\ p(f_*|u) &= \mathcal{N}\left(K_{f_*u}K_{uu}^{-1}u \;\;,\;\; K_{f_*f_*} - K_{f_*u}K_{uu}^{-1}K_{uf_*}\right) \end{aligned}$$

- $K_{ff} - K_{fu}K_{uu}^{-1}K_{uf} \triangleq K_{ff} - Q_{ff}$... Q_{ff} is information u passed to f.
- Most all methods choose $p(f|u)$ and $p(f_*|u)$, but unchanged are:

$$p(y|f) = \mathcal{N}(f, \sigma_\epsilon I) \qquad \text{and} \qquad p(u) = \mathcal{N}(0, K_{uu}).$$

Q is somewhat fundamental to this setup

- We also know the marginals:

$$p(u) = \mathcal{N}(0, K_{uu}) \quad , \quad p(f) = \mathcal{N}(0, K_{ff}) \quad , \quad p(f_*) = \mathcal{N}(0, K_{f_* f_*}).$$

- What is $cov(f, f_*)$ under the model $q(f_*|u)q(f|u)p(u)$?

$$\begin{aligned}
cov(f, f_*) &= E(f f_*^\top) - E(f)E(f_*)^\top \\
&= \int\int f f_*^\top q(f, f_*) df df_* \\
&= \int\int\int f f_*^\top q(f, f_*, u) df df_* du \\
&= \int\int\int f f_*^\top q(f|u) q(f_*|u) q(u) df df_* du \\
&= \int \left(\int f q(f|u) df\right) \left(\int f_* q(f_*|u) df_*\right)^\top q(u) du \\
&= \int \left(K_{fu} K_{uu}^{-1} u\right) \left(K_{f_* u} K_{uu}^{-1} u\right)^\top q(u) du \\
&= \int K_{fu} K_{uu}^{-1} u u^\top K_{uu}^{-1} K_{u f_*} q(u) du \\
&= K_{fu} K_{uu}^{-1} K_{u f_*} \\
&= Q_{f f_*}.
\end{aligned}$$

- Somewhat odd: define conditional and recover the effective prior.

Hypothetical full inducing point setup

- With our definition of the conditionals f, f_*:

$$\begin{aligned} p(f|u) &= \mathcal{N}\left(K_{fu}K_{uu}^{-1}u \ , \ K_{ff} - K_{fu}K_{uu}^{-1}K_{uf}\right) \\ p(f_*|u) &= \mathcal{N}\left(K_{f_*u}K_{uu}^{-1}u \ , \ K_{f_*f_*} - K_{f_*u}K_{uu}^{-1}K_{uf_*}\right) \end{aligned}$$

- We can then consider the *effective prior*:

$$p(f, f_*) = \mathcal{N}\left(0, \begin{bmatrix} K_{ff} & Q_{ff_*} \\ Q_{f_*f} & K_{f_*f_*} \end{bmatrix}\right) = \mathcal{N}\left(0, \begin{bmatrix} K_{ff} & K_{fu}K_{uu}^{-1}K_{uf_*} \\ K_{f_*u}K_{uu}^{-1}K_{uf} & K_{f_*f_*} \end{bmatrix}\right)$$

- which leads to the same old posterior form:

$$f_*|y \sim \mathcal{N}\left(Q_{f_*f}(K_{ff} + \sigma_\epsilon^2 I)^{-1}y \ , \ K_{f_*f_*} - Q_{f_*f}(K_{ff} + \sigma_\epsilon^2 I)^{-1}Q_{ff_*}\right).$$

- Note: there is no speed up here!

Deterministic inducing conditionals

- Let f, f_* be deterministic functions of u, namely:

$$\begin{aligned} q(f|u) &= \mathcal{N}\left(K_{fu}K_{uu}^{-1}u\ ,\ 0\ \right) \\ q(f_*|u) &= \mathcal{N}\left(K_{f_*u}K_{uu}^{-1}u\ ,\ 0\ \right) \end{aligned}$$

where the notation $\mathcal{N}(\cdot, 0) = \delta$.

- We can then consider the *effective prior*:

$$q_{DIC}(f, f_*) = \mathcal{N}\left(0, \begin{bmatrix} Q_{ff} & Q_{ff_*} \\ Q_{f_*f} & Q_{f_*f_*} \end{bmatrix}\right) = \mathcal{N}\left(0, \begin{bmatrix} K_{fu}K_{uu}^{-1}K_{uf} & K_{fu}K_{uu}^{-1}K_{uf_*} \\ K_{f_*u}K_{uu}^{-1}K_{uf} & K_{f_*u}K_{uu}^{-1}K_{uf_*} \end{bmatrix}\right)$$

- which leads to the same old posterior form:

$$f_*|y \sim \mathcal{N}\left(Q_{f_*f}(Q_{ff} + \sigma_\epsilon^2 I)^{-1}y\ \ ,\ \ Q_{f_*f_*} - Q_{f_*f}(Q_{ff} + \sigma_\epsilon^2 I)^{-1}Q_{ff_*}\right).$$

- A degenerate and non stationary gp with $k(x, x') = k(x, x_u)K_{uu}^{-1}k(x_u, x')$.
- Cost reduction to $\mathcal{O}(nm^2)$...

...via the matrix inversion lemma on $(Q_{ff} + \sigma_\epsilon^2 I)^{-1}$.

Deterministic *training* conditionals

- Let only f be a deterministic functions of u, namely:

$$\begin{aligned} q(f|u) &= \mathcal{N}\left(K_{fu}K_{uu}^{-1}u\ ,\ 0\ \right) \\ q(f_*|u) &= p(f_*|u) = \mathcal{N}\left(K_{f_*u}K_{uu}^{-1}u\ \ ,\ \ K_{f_*f_*} - K_{f_*u}K_{uu}^{-1}K_{uf_*}\right) \end{aligned}$$

- Again we consider the *effective prior*:

$$q_{DTC}(f, f_*) = \mathcal{N}\left(0, \begin{bmatrix} Q_{ff} & Q_{ff_*} \\ Q_{f_*f} & K_{f_*f_*} \end{bmatrix}\right) = \mathcal{N}\left(0, \begin{bmatrix} K_{fu}K_{uu}^{-1}K_{uf} & K_{fu}K_{uu}^{-1}K_{uf_*} \\ K_{f_*u}K_{uu}^{-1}K_{uf} & K_{f_*f_*} \end{bmatrix}\right)$$

- Make sure you understand:

$$q_{DTC}(f_*) = \mathcal{N}(0, K_{f_*f_*}) = \int p(f_*|u)p(u)du \int p(f_*|u)\mathcal{N}(0, K_{uu})du.$$

...the joint $p(f_*, u) = \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K_{f_*f_*} & K_{f_*u} \\ K_{uf_*} & K_{uu} \end{bmatrix}\right)$.

- Not a gp!

...treats the training and test points differently.

Fully independent (training) conditionals

- Now assume the f, f_* are again stochastic, but fully independent given u:

$$q(f|u) \;=\; \prod_{i=1}^{n} p(f_i|u) \;=\; \prod_{i=1}^{n} \mathcal{N}\left(K_{f_i u} K_{uu}^{-1} u \;,\; K_{f_i f_i} - K_{f_i u} K_{uu}^{-1} K_{u f_i}\right).$$

(and same for $q(f_*|u)$).

- Again we consider the effective prior:

$$q_{FIC}(f, f_*) = \mathcal{N}\left(0, \begin{bmatrix} Q_{ff} + diag\left(K_{ff} - Q_{ff}\right) & Q_{ff_*} \\ Q_{f_* f} & Q_{f_* f_*} + diag\left(K_{f_* f_*} - Q_{f_* f_*}\right) \end{bmatrix}\right)$$

- FIC $\rightarrow$ this assumption is made for both $q(f|u)$ and $q(f_*|u)$.
- FITC $\rightarrow$ this assumption is made for training conditionals $q(f|u)$ only.
- These techniques are probably the most heavily used sparse gp methods.
- FIC is a gp with $k(x, x') = k_{DIC}(x, x') + \mathbb{1}(x = x')\left(k(x, x') - k_{DIC}(x, x')\right)$.
- FITC is not a gp.

Partially independent (training) conditionals

- Same setup as FIC and FITC, but assume blockwise partial independence...

$$q(f|u) \quad = \quad \prod_{\text{blocks } s} p(f_s|u) \quad = \quad \prod_{\text{blocks } s} \mathcal{N}\left(K_{f_s u} K_{uu}^{-1} u\ ,\ K_{f_s f_s} - K_{f_s u} K_{uu}^{-1} K_{u f_s}\right).$$

(and same for $q(f_*|u)$).

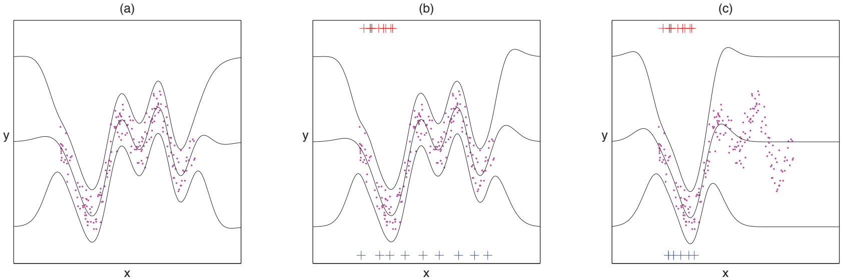
- Again we consider the effective prior:

$$q_{PIC}(f, f_*) = \mathcal{N}\left(0, \begin{bmatrix} Q_{ff} + blkdiag\left(K_{ff} - Q_{ff}\right) & Q_{ff_*} \\ Q_{f_* f} & Q_{f_* f_*} + blkdiag\left(K_{f_* f_*} - Q_{f_* f_*}\right) \end{bmatrix}\right)$$

- PIC $\rightarrow$ this assumption is made for both $q(f|u)$ and $q(f_*|u)$.
- PITC $\rightarrow$ this assumption is made for training conditionals $q(f|u)$ only.
- Neither PIC nor PITC are gp models.

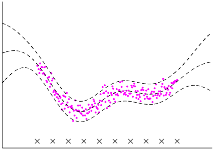
Important: cost, inducing point locations, nonconjugacy

- The cost is now reduced to $\mathcal{O}(nm^2)$, which is much less than cubic.
- Approximation quality still depends on locations x_{u_i} of each point u_i.

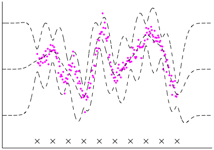


- Consider these extra model hyperparameters...
- Use all model selection tools from §02 (again ML-II is most common).
- Finally, new prior → nonconjugacy is no problem (or, the same problem).

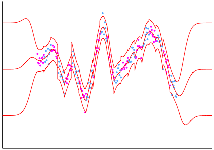
Local vs. Global



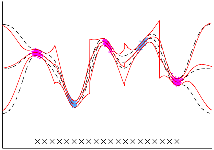
(a) Long lengthscale — FI(T)C



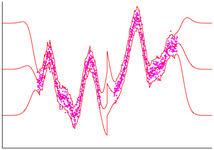
(b) Short lengthscale — FI(T)C



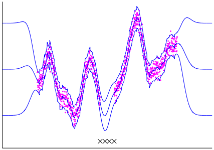
(c) Short lengthscale — local GPs



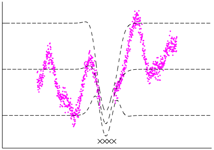
(d) Clumped training inputs



(e) local GPs



(f) PIC



(g) FI(T)C

Figure 1: 1D comparison of global, local and combined sparse GP approximations. Mean predictions and two standard deviation error lines are plotted, as black dashed lines for FI(T)C, red solid lines for local GPs, and blue solid lines for PIC. For FI(T)C and PIC the x positions of the inducing inputs are marked by black crosses. In (c) and (d) the local training blocks are demarcated by alternating the colors of the data points. In (e) and (f) the blocks are not marked for clarity, because they are very small.

Outline

Administrative interlude

Practical realities of kernel methods

Inducing point methods [QCR05]; [SG07]

Variational inference

Variational inducing point methods [Tit09]

What we want to calculate our usual quantities

- predictive distribution:

$$p(y^*|y) = \int p(y^*|f^*)p(f^*|y)df^*$$

- predictive posterior:

$$p(f^*|y) = \int p(f^*|f)p(f|y)df$$

- data posterior:

$$p(f|y) = \frac{\prod_i p(y_i|f_i)p(f)}{p(y)}$$

- *None* of which is tractable to compute.

Structured approximate inference

- predictive distribution:

$$p(y^*|y) = \int p(y^*|f^*)q(f^*|y)df^*$$

- predictive posterior:

$$q(f^*|y) = \int p(f^*|f)q(f|y)df$$

- data posterior:

$$q(f|y) \approx p(f|y) = \frac{\prod_i p(y_i|f_i)p(f)}{p(y)}$$

- Structured (gaussian) approximations: [KR05]; [RMC09]; [RW06, ch. 3; 5.5].
- Variational inference another approach (very important, and growing).

The key variational idea

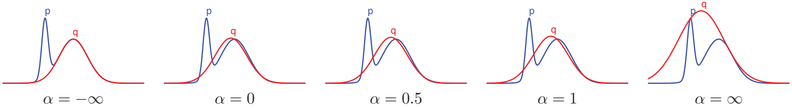
- Approximate inference → optimization by introducing variational parameters:

$$q(f|y) \quad \triangleq \quad \arg\min_{q \in \mathcal{F}} D_{KL}(q||p(f|y)) \quad = \quad \arg\min E_q \left(\log \frac{q(f)}{p(f|y)} \right)$$

  where $\mathcal{F}$ is a tractable approximating family, e.g. $\mathcal{F} = \prod_{i=1}^{n} \mathcal{N}(f_i, \mu_i, \sigma_i^2)$.

- Recall some facts about KL divergence:
  - Special case ($\alpha = 0$) of the α-divergence:

$$D_\alpha(p||q) = \frac{1}{\alpha(1-\alpha)} \int \alpha p + (1-\alpha) q - p^\alpha q^{1-\alpha}$$



  - big q, big p good; big q, small p bad; small q, big p who cares.
  - cf. minimizing $KL(p||q)$ ($\alpha = 1$)...

$\alpha = 1 \rightarrow$ moment matching; $\alpha = 0 \rightarrow$ local correctness

The key variational idea

- Approximate inference → optimization by introducing variational parameters:

$$
\begin{aligned}
q(f|y) &\triangleq \arg\min_{q \in \mathcal{F}} D_{KL}(q||p(f|y)) \\
&= \arg\min E_q \left(\log \frac{q(f)}{p(f|y)} \right) \\
&= \arg\min E_q \left(\log q(f) \right) - E_q \left(\log p(f|y) \right) \\
&= \arg\min E_q \left(\log q(f) \right) - E_q \left(\log p(f, y) \right) + \log p(y) \\
&= \arg\min E_q \left(\log q(f) \right) - E_q \left(\log p(f, y) \right) \\
&\triangleq \arg\max \mathcal{L}(q)
\end{aligned}
$$

- We call $\mathcal{L}(q) = E_q \left(\log p(f, y) \right) - E_q \left(\log q(f) \right)$ the *ELBO* because:

$$
\begin{aligned}
\log p(y) &= \log \int p(f, y) df \\
&= \log \int p(f, y) \frac{q(f)}{q(f)} df \\
&= \log E_q \left(\frac{p(f, y)}{q(f)} \right) df \\
&\geq \mathcal{L}(q),
\end{aligned}
$$

- ...as in, evidence (marginal likelihood) lower bound ...(or energy + entropy).
- Note: to truly understand what VB (and EP) is doing, see [WJ08].

Mean field variational inference

- Assume independence, e.g. $q(f) = \prod_i q_i(f_i)$. Conveniently:

$$\begin{aligned} \mathcal{L}(q) &= -D_{KL}\left(q_i(f_i)||\frac{1}{Z}\exp\left\{E_{q_{-i}}\left(\log p(f,y)\right)\right\}\right) - E_{q_{-i}}\left(\log q_{-i}(f_{-i})\right) + \log Z \\ &\propto -D_{KL}\left(q_i(f_i)||\frac{1}{Z}\exp\left\{E_{q_{-i}}\left(\log p(f,y)\right)\right\}\right), \end{aligned}$$

See [FR12] for details.

- ...thus, iteratively minimizing marginal KL divergences $\rightarrow$ coordinate ascent.
- MFVB is local and overconfident, but really useful.
- Common mistake: posterior marginals are **not** well captured, generally.
- Exponential family distributions often make MFVB easy to implement.
- Editorializing: VB is broader than MF, like EP is more general than Gaussian.

Outline

Administrative interlude

Practical realities of kernel methods

Inducing point methods [QCR05]; [SG07]

Variational inference

Variational inducing point methods [Tit09]

Recall inducing points

- Inducing points u are jointly gaussian with the latent gp f, such that:

$$p(f_*, f) = \int p(f, f_*, u) du = \int p(f, f_*|u) p(u) du \quad , \quad \text{where } u \sim \mathcal{N}(0, K_{uu}),$$

- We are interested in the usual things, like the posterior:

$$f_*|y \sim \mathcal{N}\left(K_{f_*f}(K_{ff} + \sigma_\epsilon^2 I)^{-1} y \quad , \quad K_{f_*f_*} - K_{f_*f}(K_{ff} + \sigma_\epsilon^2 I)^{-1} K_{ff_*}\right)$$

- The **critical** conditional independence assumption:

$$p(f_*, f) \approx q(f_*, f) = \int q(f_*|u) q(f|u) p(u) du$$

- *Training and test points are conditionally independent, given inducing points.*

Bringing together sparse and variational inference

- This sentence should now make sense:

  We introduce a variational formulation for sparse approximations that jointly infers the inducing inputs and the kernel hyperparameters by maximizing a lower bound of the true log marginal likelihood.

- We are, as always, interested in the posterior. Chain rule:

$$p(f_*|y) = \int p(f_*|u,f)p(f|u,y)p(u|y)dfdu.$$

Why not $p(f_*|u,f,y)$?

- Now we make the usual sparse assumption $f \perp f_*|u$, such that:

$$q(f_*) = \int p(f_*|u)p(f|u)p(u|y)du$$

...$p(f|u) = p(f|u,y)$ is a nontrivial fact that you should prove.

...also: $f \leftrightarrow f$; $y \leftrightarrow y$; $f_m \leftrightarrow u$; $X_m \leftrightarrow X_u$; $z \leftrightarrow f_*$.

- Key idea: $q(f_*) \approx p(f_*|y)$, so let $p(u|y) \triangleq q(u)$, a variational distribution!

...pause to appreciate the indirect variational posterior $q(f)$.

Difference vs previous

- Let $q(u) = \mathcal{N}(u; \mu, A)$. This induces a posterior gp:

$$q(f) = \mathcal{GP}\left(\;\; K_{\cdot u} K_{uu} \mu \;\;,\;\; k_{\cdot\cdot} - k_{\cdot u} K_{uu}^{-1} k_{u\cdot} + k_{\cdot u} \left(K_{uu}^{-1} A K_{uu}^{-1} \right) k_{u\cdot} \;\; \right)$$

Somewhat tedious, but correct...

- This now defines an approximate posterior (one step removed), and thus:

$$\begin{aligned}
\{X_u, \mu, A\} &= \arg\max \mathcal{L}(X_u, \mu, A) \\
&= \arg\max E_q\left(\log p(f, y)\right) - E_q\left(\log q(f)\right) \\
&= \arg\max \int p(f|u) q(u) \log \frac{p(y|f) p(f|u) p(u)}{p(f|u) q(u)} df\, du.
\end{aligned}$$

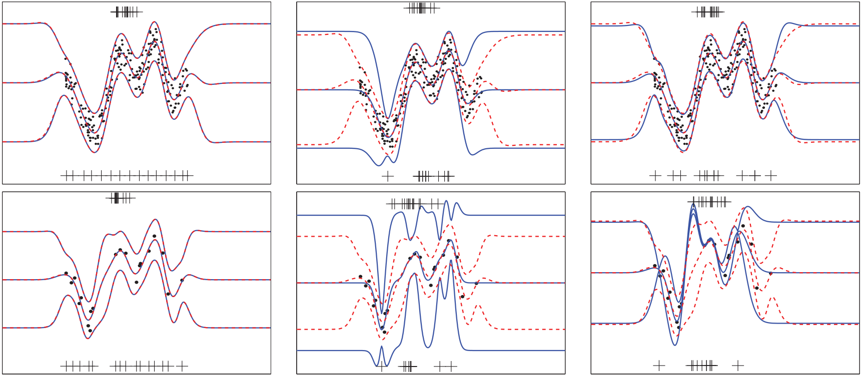
- Variational parameters μ, A can be solved analytically (see [Tit09, Supp.]):

$$\mathcal{L}(X_u) = \log \mathcal{N}\left(y; 0, K_{fu} K_{uu}^{-1} K_{uf} + \sigma_\epsilon^2 I\right) - \frac{1}{2\sigma^2} tr\left(K_{ff} - K_{fu} K_{uu}^{-1} K_{uf}\right).$$

...recall $Q_{ff} = K_{fu} K_{uu}^{-1} K_{uf}$

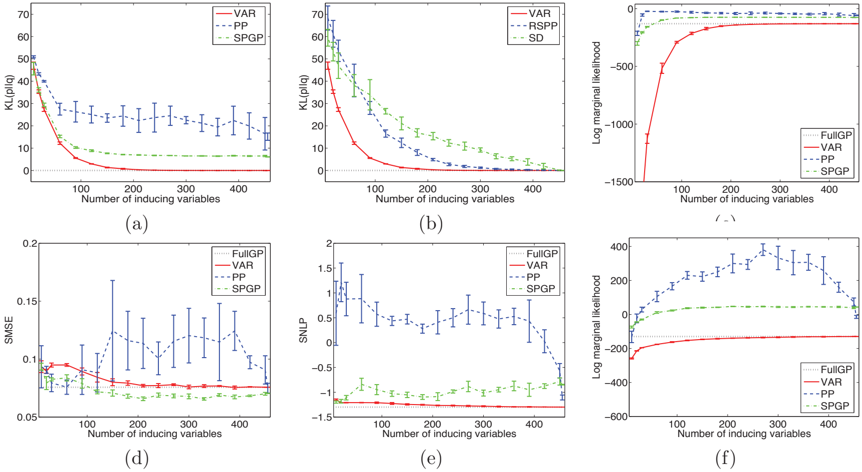
- This is the same as previously ([SG07] and friends), with a regularizer!

Results



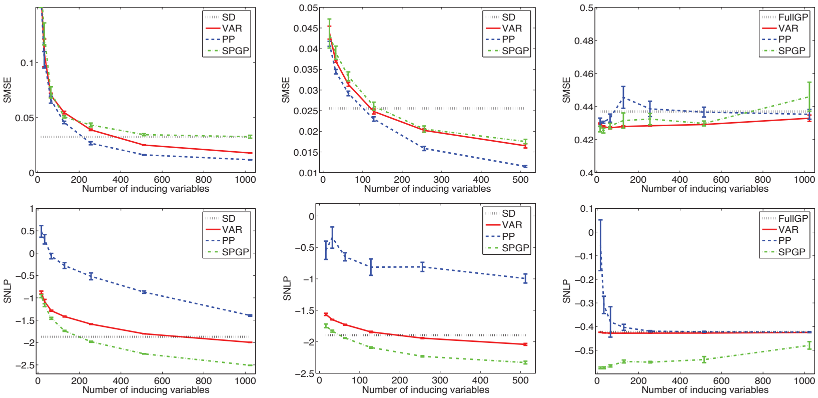
Note: here SPGP is FI(T)C from [SG07].

Results



Note: here SPGP is FI(T)C from [SG07].

Results



Note: here SPGP is FI(T)C from [SG07].

References

[FR12] Charles W Fox and Stephen J Roberts.
A tutorial on variational bayesian inference.
Artificial intelligence review, 38(2):85–95, 2012.

[GT15] Yarin Gal and Richard Turner.
Improving the gaussian process sparse spectrum approximation by representing uncertainty in frequency inputs.
In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 655–664, 2015.

[KR05] Malte Kuss and Carl Edward Rasmussen.
Assessing approximate inference for binary gaussian process classification.
The Journal of Machine Learning Research, 6:1679–1704, 2005.

[QCR05] 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.

[RMC09] Havard Rue, Sara Martino, and Nicolas Chopin.
Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations.
Journal of the royal statistical society: Series b (statistical methodology), 71(2):319–392, 2009.

[RW06] C. E. Rasmussen and C.K.I. Williams.
Gaussian Processes for Machine Learning.
MIT Press, Cambridge, 2006.

[SG07] Edward Snelson and Zoubin Ghahramani.
Local and global sparse gaussian process approximations.
In *International Conference on Artificial Intelligence and Statistics*, pages 524–531, 2007.

[Tit09] Michalis K Titsias.
Variational learning of inducing variables in sparse gaussian processes.
In *International Conference on Artificial Intelligence and Statistics*, pages 567–574, 2009.

[TLG14] Michalis Titsias and Miguel Lázaro-Gredilla.
Doubly stochastic variational bayes for non-conjugate inference.
In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1971–1979, 2014.

[WJ08] Martin J Wainwright and Michael I Jordan.
Graphical models, exponential families, and variational inference.
Foundations and Trends in Machine Learning, 1(1-2):1–305, 2008.