SPARSE VARIATIONAL GP

for classification

James Hensman

May 21, 2015

University of Sheffield

JOINT WORK

- Nicolo' Fusi
- Alex Matthews
- Neil Lawrence
- Zoubin Ghahramani

REVIEW: VARIATIONAL SPARSE GP

$$p(\mathbf{y}, \mathbf{f}) = p(\mathbf{y} \,|\, \mathbf{f})p(\mathbf{f})$$

AUGMENTING THE MODEL

$$p(\mathbf{y}, \mathbf{f}) = p(\mathbf{y} \,|\, \mathbf{f})p(\mathbf{f})$$

$$p(\mathbf{y}, \mathbf{f}, \mathbf{u}) = p(\mathbf{y} \,|\, \mathbf{f})p(\mathbf{f})p(\mathbf{u} \,|\, \mathbf{f})$$

Let the form of the approximate posterior be

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

THE BOUND:

$$\mathcal{L} = \mathbb{E}_{\mathrm{q}(\mathbf{f})}\left[\log \mathrm{p}(\mathbf{y} \,|\, \mathbf{f})\right] - \mathcal{KL}\left[\mathrm{q}(\mathbf{u}) || \mathrm{p}(\mathbf{u})\right]$$

For a Gaussian likelihood, it is possible to find an analytical solution for $q(\mathbf{u})$:

$$\mathcal{L} = \log \mathcal{N}(\mathbf{y} \mid \mathbf{0}, \mathbf{Q}_{\mathrm{ff}} + \sigma^2 \mathbf{I}) - \frac{1}{2\sigma^2} \mathrm{tr}\left(\mathbf{K}_{\mathrm{ff}} - \mathbf{Q}_{\mathrm{ff}}\right)$$

with

$$\mathbf{Q}_{\mathrm{ff}} = \mathbf{K}_{\mathrm{fu}} \mathbf{K}_{\mathrm{uu}}^{-1} \mathbf{K}_{\mathrm{fu}}^{\top}$$

Titsias 2009

…but making the representation of $\mathrm{q}(\mathbf{u})$ explicit leads to a stochastic optimization algorithm

$$\mathrm{q}(\mathbf{u}) = \mathcal{N}(\mathbf{u} \,|\, \mathbf{m}, \mathbf{S})$$

$$\mathcal{L} = \log \mathcal{N}(\mathbf{y} \,|\, \mathbf{A}\mathbf{m}, \sigma^2 \mathbf{I}) - \frac{1}{2\sigma^2} \operatorname{tr}\left(\mathbf{K}_{ff} - \mathbf{A}\mathbf{S}\mathbf{A}^\top\right) - \mathcal{KL}\left[\mathrm{q}(\mathbf{u}) || \mathrm{p}(\mathbf{u})\right]$$

with

$$\mathbf{A} = \mathbf{K}_{fu} \mathbf{K}_{uu}{}^{-1}$$

Hensman et al 2013

PREDICTION

The nonparametric nature of the model is recovered at predict time.

The nonparametric nature of the model is recovered at predict time.

$$p(\mathbf{f}^\star \mid \mathbf{y}) = \int p(\mathbf{f}^\star \mid \mathbf{f}) p(\mathbf{f} \mid \mathbf{y}) \, d\mathbf{f}$$

The nonparametric nature of the model is recovered at predict time.

$$p(\mathbf{f}^\star \mid \mathbf{y}) = \int p(\mathbf{f}^\star \mid \mathbf{f}) p(\mathbf{f} \mid \mathbf{y}) \, d\mathbf{f}$$

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

Minimize the KL:

$$\mathcal{KL}\left[q(\mathbf{u})p(\mathbf{f} \mid \mathbf{u}) || p(\mathbf{u}, \mathbf{f} \mid \mathbf{y})\right]$$

Minimize the KL:

$$\mathcal{KL}\left[q(\mathbf{u})p(\mathbf{f}\,|\,\mathbf{u})||p(\mathbf{u},\mathbf{f}\,|\,\mathbf{y})\right]$$

Which minimizes the KL beween the posterior processes

$$\int p(\mathbf{f}^\star\,|\,\mathbf{u})q(\mathbf{u})\,\mathrm{d}\mathbf{u} \;\;||\;\; p(\mathbf{f}^\star\,|\,\mathbf{y})$$

Seeger 2003, Matthews et al 2015

NON-GAUSSIAN LIKELIHOODS

LIKELIHOOD FACTORIZATION

$$\mathcal{L} = \mathbb{E}_{\mathsf{q}(\mathbf{f})}\left[\log \mathsf{p}(\mathbf{y} \,|\, \mathbf{f})\right] - \mathcal{KL}\left[\mathsf{q}(\mathbf{u}) || \mathsf{p}(\mathbf{u})\right]$$

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{f})}\left[\log p(\mathbf{y} \mid \mathbf{f})\right] - \mathcal{KL}\left[q(\mathbf{u}) || p(\mathbf{u})\right]$$

$$\mathcal{L} = \sum_i \mathbb{E}_{q(\mathbf{f}_i)}\left[\log p(\mathbf{y}_i \mid \mathbf{f}_i))\right] - \mathcal{KL}\left[q(\mathbf{u}) || p(\mathbf{u})\right]$$

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

One dimensional quadrature

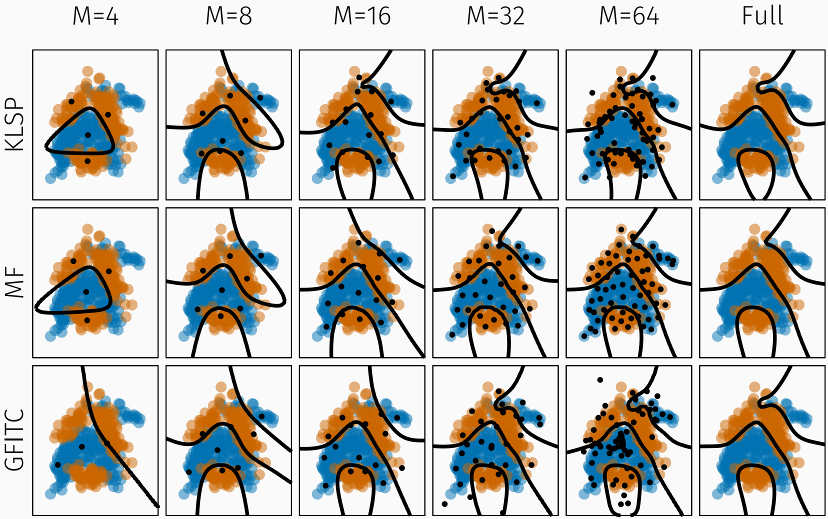
- Assume $q(\mathbf{u})$ is Gaussian
- parameterize as $q(\mathbf{u}) = \mathcal{N}(\mathbf{m}, \mathbf{L}\mathbf{L}^\top)$
- Optimize $\mathbf{Z}, \theta, \mathbf{m}, \mathbf{L}$ using preferred (stochastic?) optimizer

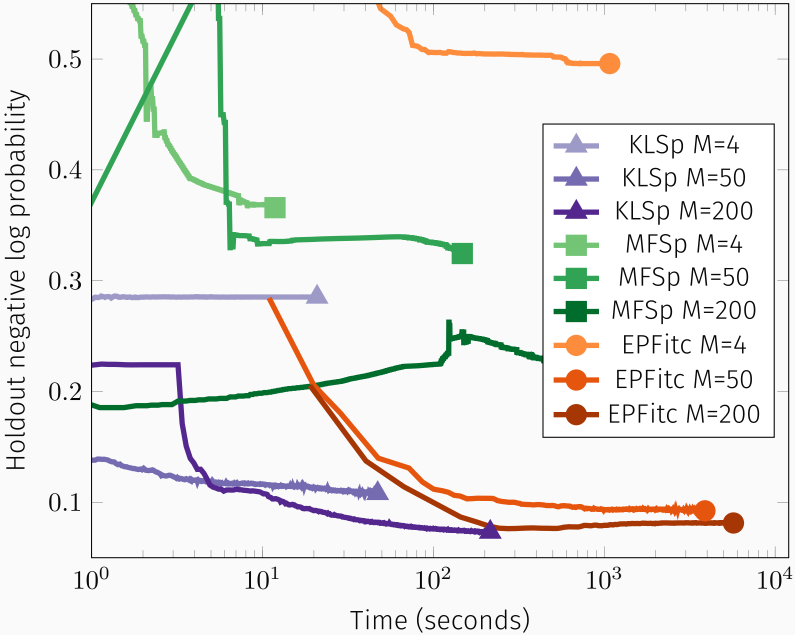
VarDTC + mean field

- Assume latent noise $p(\mathbf{t} \mid \mathbf{y})p(\mathbf{y} \mid \mathbf{f})p(\mathbf{f})$
- Use Sparse GP regression for $\mathbf{f}$
- Use mean-field approximation for latent $\mathbf{y}$
- $+$ Minimizes a KL
- $-$ Nastly factorizing assumption
- $-$ No stochastics

EP + FITC

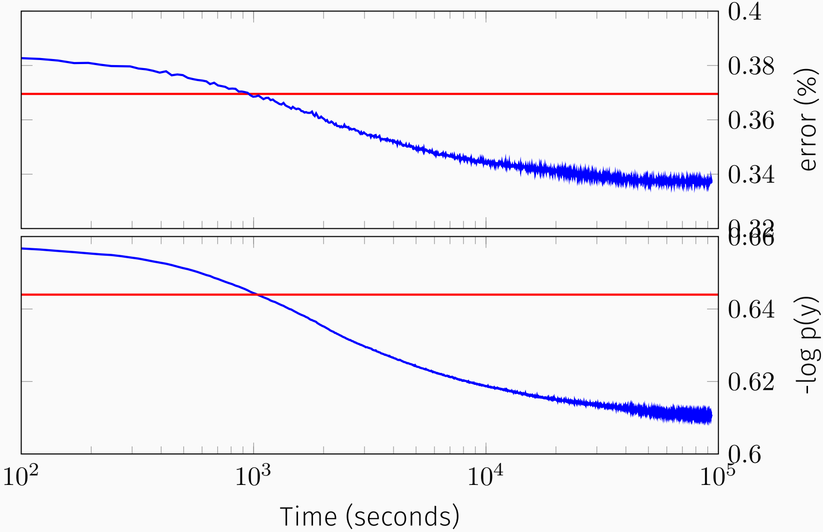
- Use FITC approximation to the covariance
- Do EP on this approximation
- $+$ EP
- $-$ Stochastics?


M=4
M=8
M=16
M=32
M=64
Full
KLSP
MF
GFITC



Holdout negative log probability
0.5
0.4
0.3
0.2
0.1
10^0
10^1
10^2
10^3
10^4
Time (seconds)
KLSp M=4
KLSp M=50
KLSp M=200
MFSp M=4
MFSp M=50
MFSp M=200
EPFitc M=4
EPFitc M=50
EPFitc M=200

AIRLINE DELAY CLASSIFICATION



FREE FORM VARIATIONAL METHOD

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{f})}\left[\log p(\mathbf{y} \mid \mathbf{f})\right] - \mathcal{KL}\left[q(\mathbf{u}) || p(\mathbf{u})\right]$$

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{f})}\left[\log p(\mathbf{y} \mid \mathbf{f})\right] - \mathcal{KL}\left[q(\mathbf{u}) || p(\mathbf{u})\right]$$

$$\log \hat{q}(\mathbf{u}) = \mathbb{E}_{p(\mathbf{f} \mid \mathbf{u})}\left[\log p(\mathbf{y} \mid \mathbf{f})\right] + \log p(\mathbf{u}) + C$$

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{f})}\left[\log p(\mathbf{y} \mid \mathbf{f})\right] - \mathcal{KL}\left[q(\mathbf{u}) || p(\mathbf{u})\right]$$

$$\log \hat{q}(\mathbf{u}) = \mathbb{E}_{p(\mathbf{f} \mid \mathbf{u})}\left[\log p(\mathbf{y} \mid \mathbf{f})\right] + \log p(\mathbf{u}) + C$$

$$\log \hat{q}(\mathbf{u}, \theta) = \mathbb{E}_{p(\mathbf{f} \mid \mathbf{u}, \theta)}\left[\log p(\mathbf{y} \mid \mathbf{f})\right] + \log p(\mathbf{u} \mid \theta) + \log p(\theta) + C$$

Questions?