

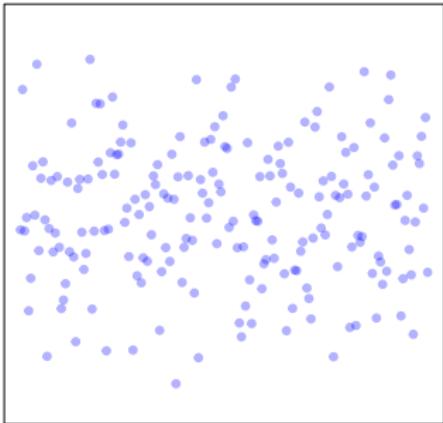
Unbiased computations for tractable and scalable learning of Gaussian processes

Maurizio Filippone

Department of Data Science, EURECOM
maurizio.filippone@eurecom.fr

April 24th, 2017

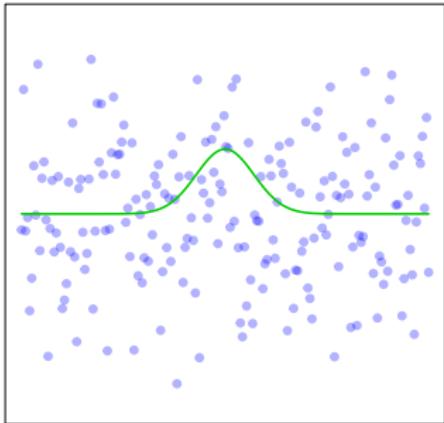
Gaussian Processes - Prior over Functions



$$K = \begin{matrix} & \text{---} \\ \text{---} & \end{matrix}$$
A diagram illustrating a covariance matrix K . It consists of a square grid of blue squares, each representing an element of the matrix. The grid is bounded by dashed lines. A dashed diagonal line runs from the top-left square to the bottom-right square, representing the main diagonal of the matrix.

$K =$

Gaussian Processes - Prior over Functions

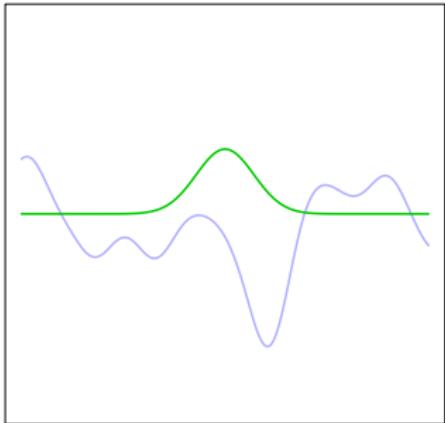


$$K = \begin{matrix} & \text{---} \\ \text{---} & \end{matrix}$$

A diagram illustrating a covariance matrix K . It shows a square grid with dashed lines representing the boundaries. Blue squares are placed at the intersections of the first few rows and columns, forming a lower-triangular pattern. A dashed diagonal line extends from the top-left corner to the bottom-right corner, indicating the symmetry of the matrix.

$$K =$$

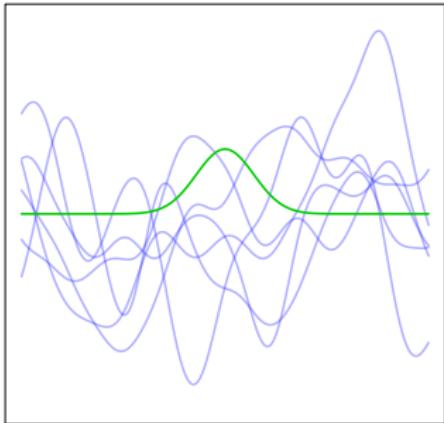
Gaussian Processes - Prior over Functions



$$K = \begin{matrix} & \begin{matrix} \text{---} & \end{matrix} \\ \begin{matrix} \text{---} \\ \text{---} \end{matrix} & \begin{matrix} \text{---} & \text{---} \\ \text{---} & \text{---} \end{matrix} \end{matrix}$$

A diagram illustrating the covariance matrix K of a Gaussian process. It shows a 4x4 grid of smaller 2x2 blocks, each containing four blue squares. Dashed lines connect the centers of these blocks, forming a larger 4x4 grid pattern, representing the correlation structure between data points.

Gaussian Processes - Prior over Functions

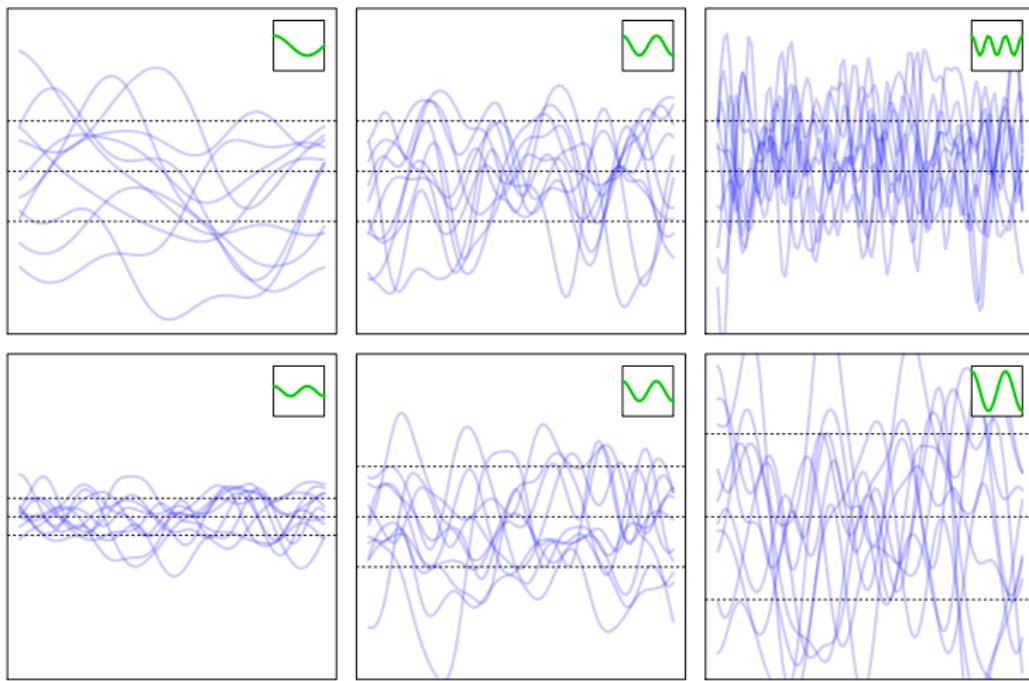


$$K = \begin{matrix} & \begin{matrix} \text{---} & \end{matrix} \\ \begin{matrix} \text{---} & \end{matrix} & \begin{matrix} \text{---} & \end{matrix} \end{matrix}$$

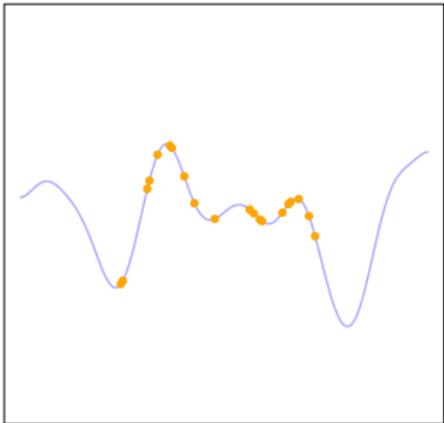
A diagram illustrating the covariance matrix K of a Gaussian process. It shows a 4x4 grid of blue squares representing the entries of the matrix. Dashed lines connect the squares in a pattern that indicates non-zero entries only on the main diagonal and the two adjacent diagonals, forming a tridiagonal matrix structure.

Gaussian Processes - Priors over Functions

- Infinite Gaussian random variables with parameterized and input-dependent covariance



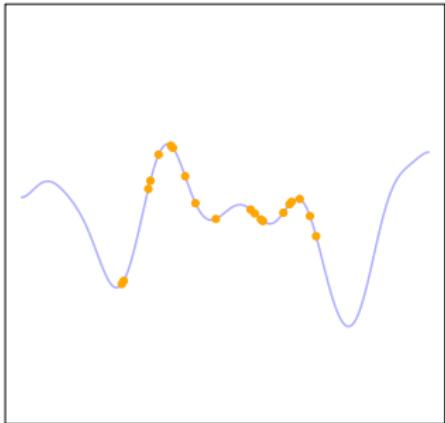
Gaussian Processes - Prior over Functions



$$K = \begin{matrix} & \star & \star \\ \star & \begin{matrix} \text{blue} & \text{orange} \\ \text{orange} & \text{blue} \end{matrix} & \begin{matrix} \text{blue} & \text{orange} \\ \text{orange} & \text{blue} \end{matrix} \\ \star & \begin{matrix} \text{blue} & \text{orange} \\ \text{orange} & \text{blue} \end{matrix} & \begin{matrix} \text{blue} & \text{orange} \\ \text{orange} & \text{blue} \end{matrix} \end{matrix}$$

A diagram illustrating the covariance matrix K for a Gaussian process. The matrix is a 3x3 grid of smaller 2x2 blocks. The diagonal blocks are labeled with stars (\star). The off-diagonal blocks are labeled with either blue or orange colors, representing the covariance between different data points. Dashed lines indicate the connections between the blocks, showing how the function values at one location influence the values at others.

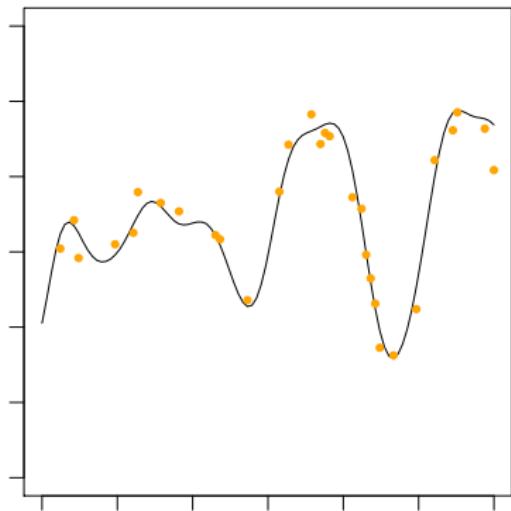
Gaussian Processes - Prior over Functions



$$K = \underbrace{\begin{matrix} & & \\ & & \\ & & \\ & & \\ & & \end{matrix}}_n \quad \underbrace{\begin{matrix} & & \\ & & \\ & & \\ & & \\ & & \end{matrix}}_n$$

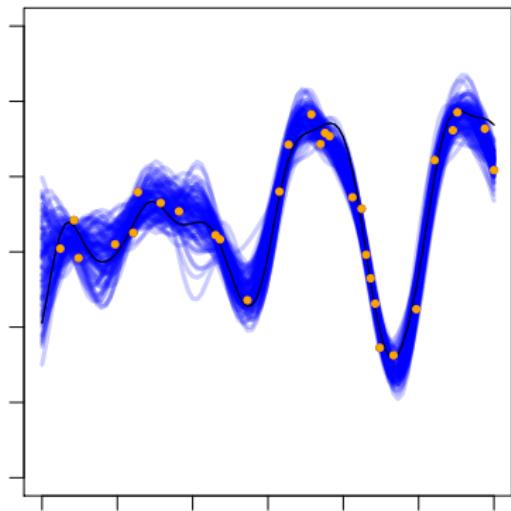
Gaussian Processes - Prior over Functions

- Regression example

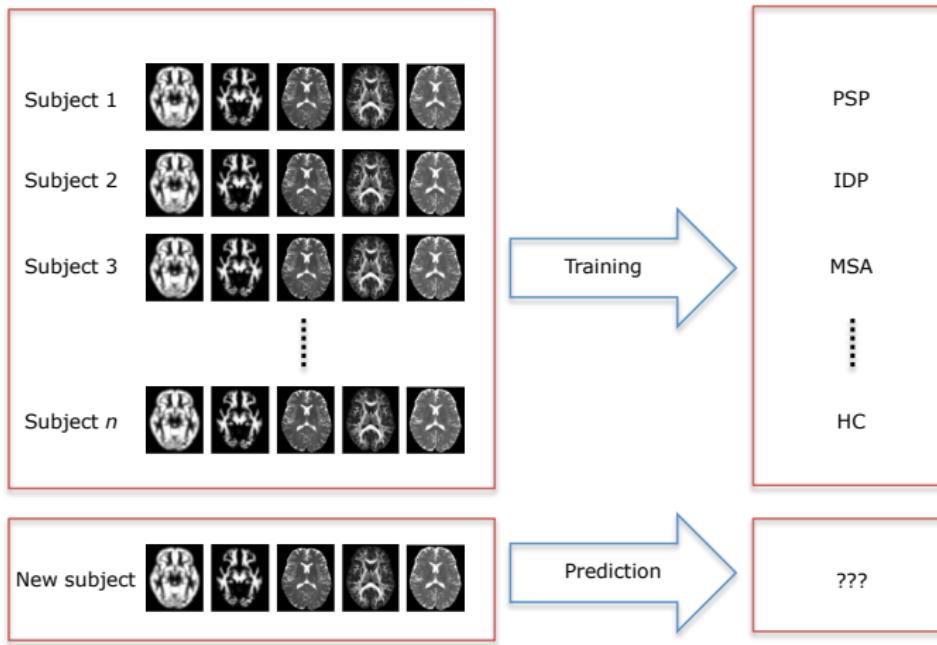


Gaussian Processes - Prior over Functions

- Regression example



Motivating Application



HC - Healthy control

MSA - Multiple system atrophy

PSP - Progressive Supranuclear Palsy

IDP - Idiopathic Parkinson's disease

Multiclass classification with multiple sources

- Multiclass classification based on GPs

$p(\text{disease} = c | \text{sources}) = \text{unknown function}$

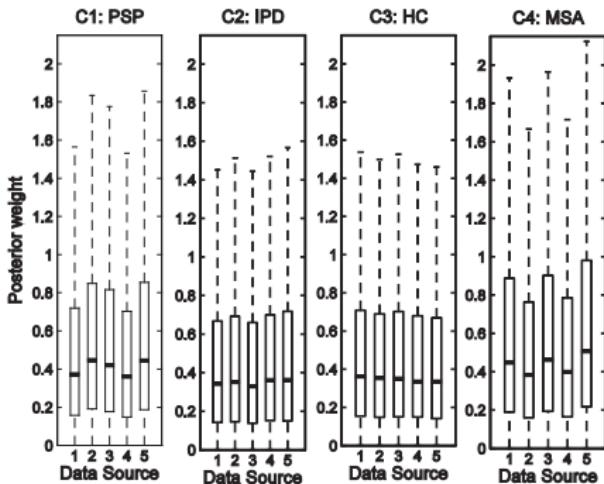
- unknown function modeled using GPs
- Covariance based on source-dependent covariances S_k

$$\sum_{k=1}^K w_{ck} S_k(\text{subject}_i, \text{subject}_j)$$

Filippone, Marquand et al., AoAS, 2012

Parkinsonian disorders data - multiple sources classification

Method	Accuracy
GP classifier	0.598
SimpleMKL	0.418



Filippone, Marquand et al., AoAS, 2012

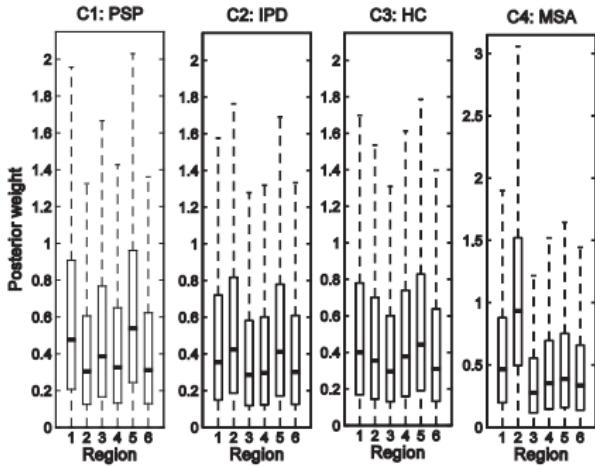
Multiclass classification with multiple regions

Analysis of brain regions

- ① brainstem
 - ② bilateral cerebellum
 - ③ bilateral caudate
 - ④ bilateral middle occipital gyrus
 - ⑤ bilateral putamen
 - ⑥ all other regions

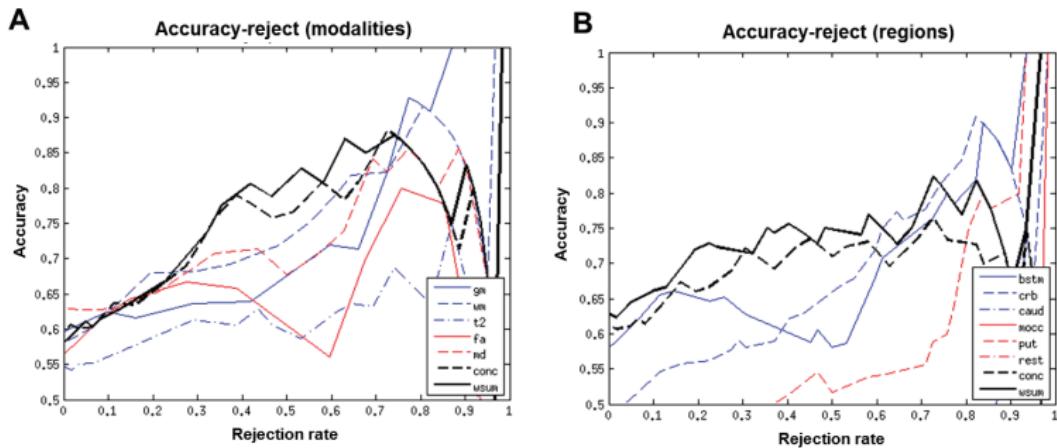
Parkinsonian disorders data - multiple regions classification

Method	Accuracy
GP classifier	0.614
SimpleMKL	0.229



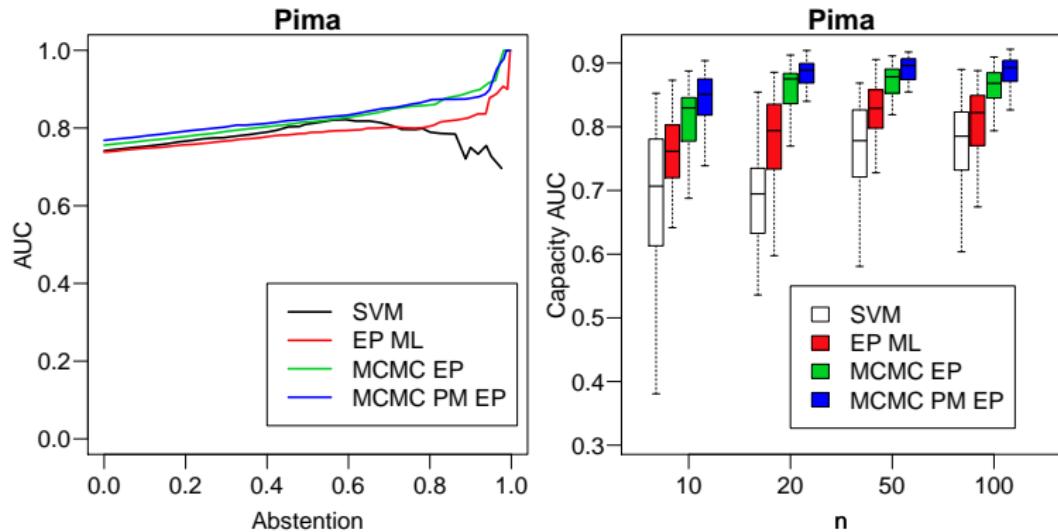
Filippone, Marquand et al., AoAS, 2012

Reject Option



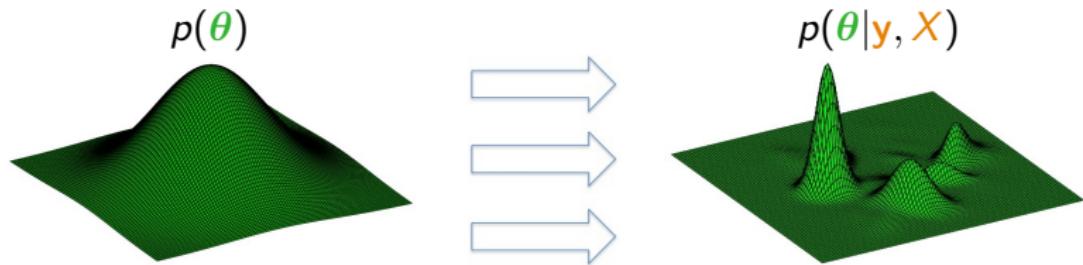
Filippone, Marquand et al., AoAS, 2012

Does being Bayesian buy you anything?



Bayesian Gaussian Processes

- Inputs = X Labels = y
- $K = K(X, \theta)$



$$p(\theta|y, X) = \frac{p(y|X, \theta)p(\theta)}{\int p(y|X, \theta)p(\theta)d\theta}$$

- Predictions for new data

$$p(\mathbf{y}_* | \mathbf{y}, \mathbf{X}_*, \mathbf{X}) = \int p(\mathbf{y}_* | \mathbf{X}_*, \boldsymbol{\theta}) p(\boldsymbol{\theta} | \mathbf{y}, \mathbf{X}) d\boldsymbol{\theta}$$

- Monte Carlo integration:

$$\int p(\mathbf{y}_* | \mathbf{X}_*, \boldsymbol{\theta}) p(\boldsymbol{\theta} | \mathbf{y}, \mathbf{X}) d\boldsymbol{\theta} \simeq \frac{1}{N_{\text{MC}}} \sum_{i=1}^{N_{\text{MC}}} p(\mathbf{y}_* | \boldsymbol{\theta}^{(i)})$$

with $\boldsymbol{\theta}^{(i)}$ drawn from $p(\boldsymbol{\theta} | \mathbf{y}, \mathbf{X})$

- Bayesian inference

$$p(\theta | \mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y} | \mathbf{X}, \theta) p(\theta)}{\int p(\mathbf{y} | \mathbf{X}, \theta) p(\theta) d\theta}$$

- Random walk sampler - accept a proposal with probability

$$\min \left(1, \frac{p(\theta' | \mathbf{y}, \mathbf{X})}{p(\theta | \mathbf{y}, \mathbf{X})} \right)$$

Markov chain Monte Carlo - Random walk example

Acceptance probability : $\min \left(1, \frac{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}') p(\boldsymbol{\theta}')}{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) p(\boldsymbol{\theta})} \right)$

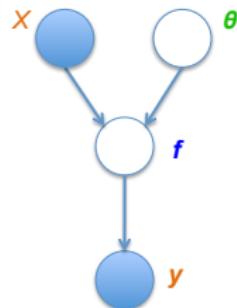
Metropolis et al., *JoCP*, 1953 - Hastings, *Biometrika*, 1970

Challenges and Limitations

- θ can be large dimensional (ARD/factor covariances)
- No exact Gibbs steps
- $p(\mathbf{y}|\mathbf{X}, \theta)$ can be multi-modal
- How to check convergence?

Challenges and Limitations

- θ can be large dimensional (ARD/factor covariances)
- No exact Gibbs steps
- $p(y|X, \theta)$ can be multi-modal
- How to check convergence?
- $p(y|X, \theta)$ might be expensive to compute
- $p(y|X, \theta)$ might not even be computable!



Marginal likelihood of GP models

- Marginal likelihood

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta})d\mathbf{f}$$

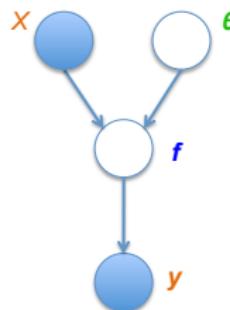
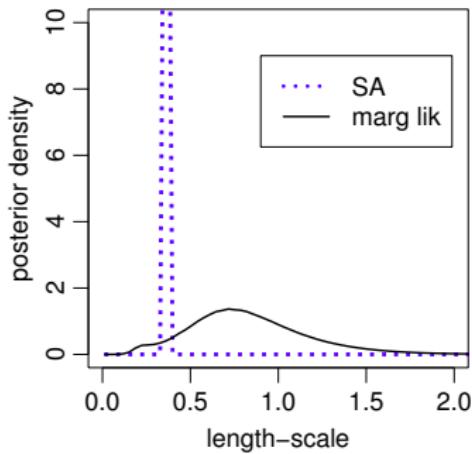
can only be computed if $p(\mathbf{y}|\mathbf{f})$ is Gaussian

- What if $p(\mathbf{y}|\mathbf{f})$ is **not** Gaussian?

Challenges in MCMC for GPMs - Structure

Sufficient Augmentation (SA) scheme - alternate between:

- Drawing from $p(\mathbf{f}|\boldsymbol{\theta}, \mathbf{y}, \mathbf{X})$
- Drawing from $p(\boldsymbol{\theta}|\mathbf{f})$ - bad idea



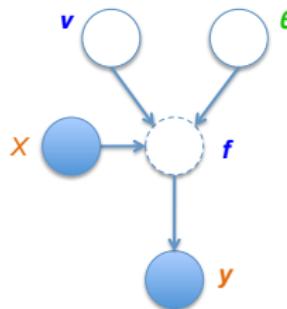
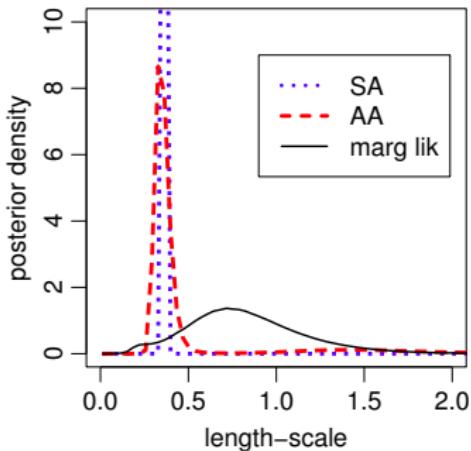
Murray and Adams, NIPS, 2010 - Filippone et al., Mach. Learn., 2013.

Mitigating coupling effect through reparameterization

Ancillary Augmentation (AA) scheme - reparameterization:

$$K = L L^\top \quad \boldsymbol{\nu} = L^{-1} \mathbf{f}$$

- Sampling of $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta}|\boldsymbol{\nu}, \mathbf{y}, \mathbf{X})$

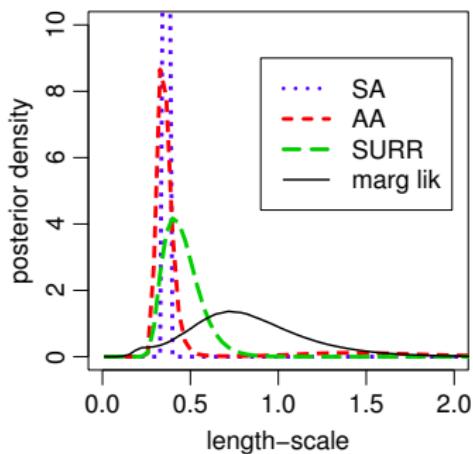


Murray and Adams, NIPS, 2010 - Filippone et al., Mach. Learn., 2013.

Mitigating coupling effect through reparameterization

Surrogate data model (SURR):

- Introduce auxiliary vars informed by the posterior over \mathbf{f}

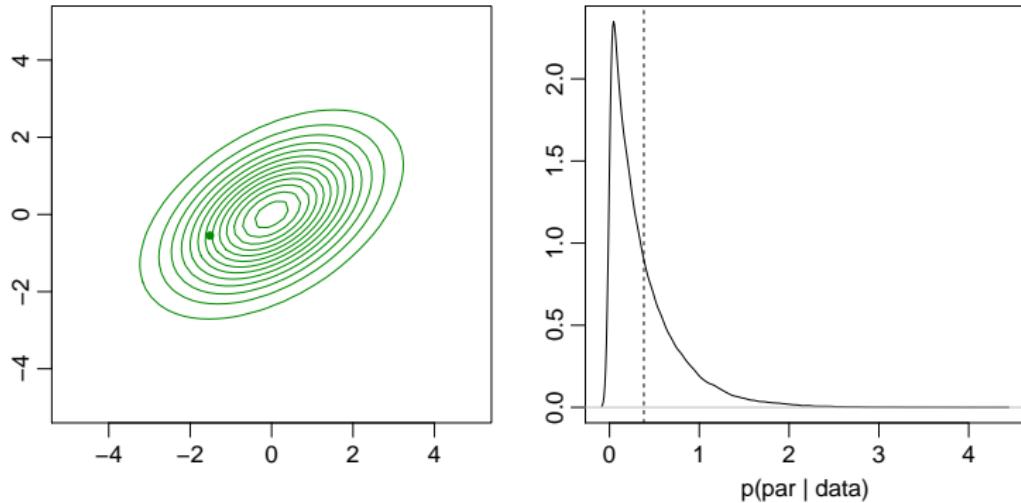


$$\mathbf{g} = \varphi(\mathbf{f}, \boldsymbol{\theta})$$

Murray and Adams, *NIPS*, 2010 - Filippone et al., *Mach. Learn.*, 2013.

“Noisy” Markov chain Monte Carlo

$$E \{ \tilde{p}(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}) \} = p(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta})$$



Andrieu and Roberts, AoS, 2009

“Noisy” Markov chain Monte Carlo

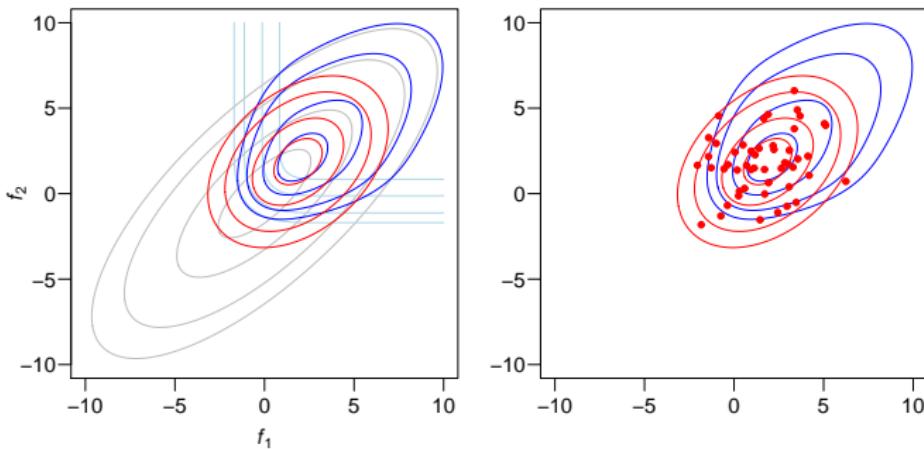
Acceptance probability : $\min \left(1, \frac{\tilde{p}(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}') p(\boldsymbol{\theta}')}{\tilde{p}(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) p(\boldsymbol{\theta})} \right)$

Andrieu and Roberts, AoS, 2009

Importance Sampling estimator

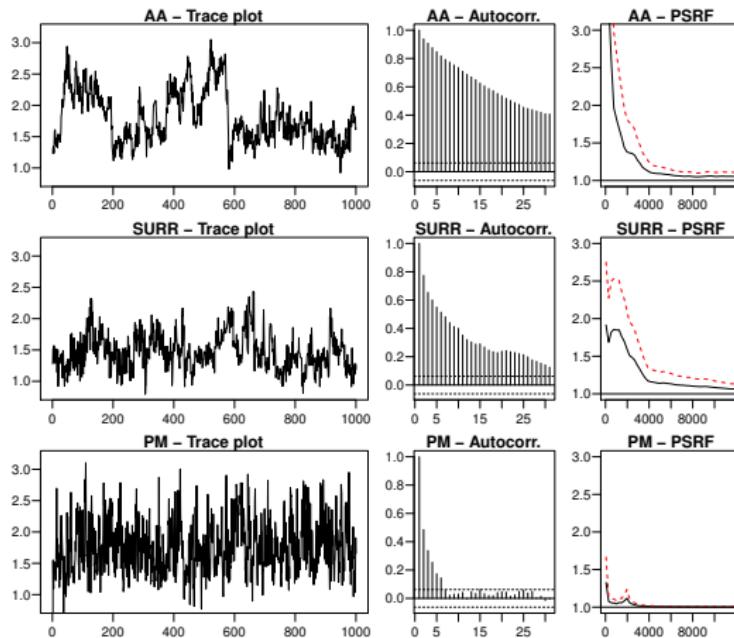
- Approximate posterior over latent variables using $q(\mathbf{f})$
- Then

$$\tilde{p}(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \frac{1}{N_{\text{IS}}} \sum_{i=1}^{N_{\text{IS}}} \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{f}^{(i)}) p(\mathbf{f}^{(i)}|\boldsymbol{\theta})}{q(\mathbf{f}^{(i)})}$$



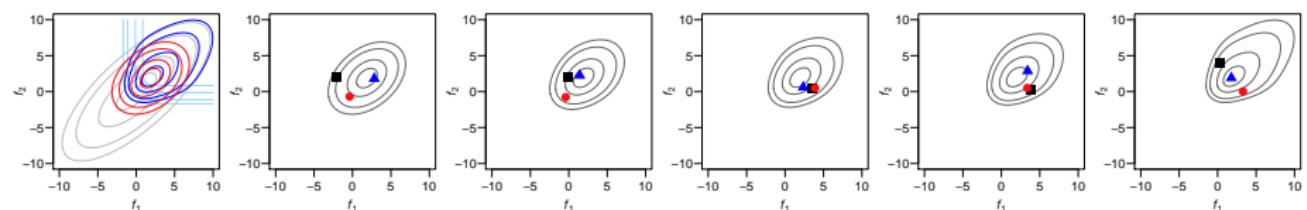
Convergence speed and efficiency

Abalone data set (two classes) $n = 2835$ - inference of length-scale



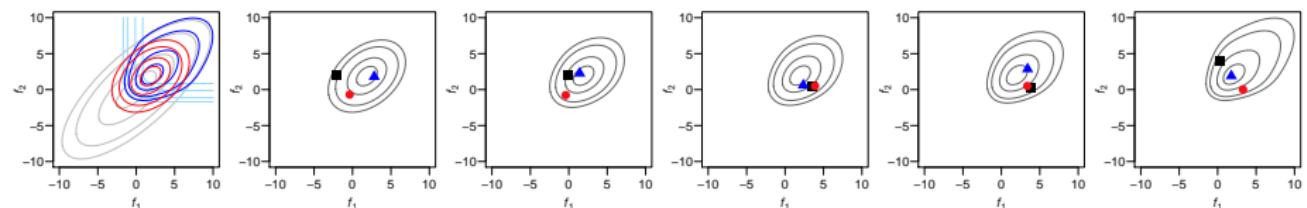
Annealed Importance Sampling estimator

- Annealing from an approximation

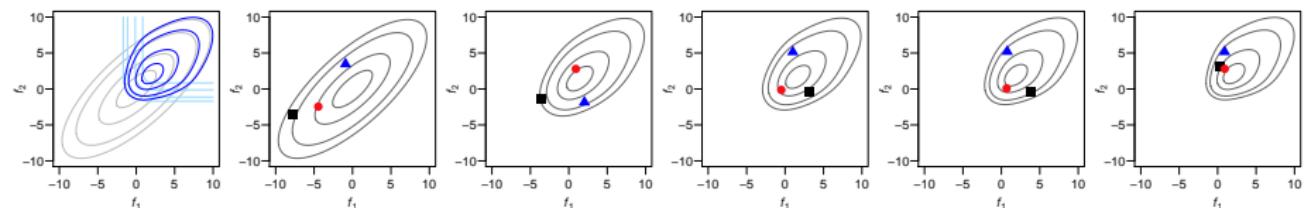


Annealed Importance Sampling estimator

- Annealing from an approximation



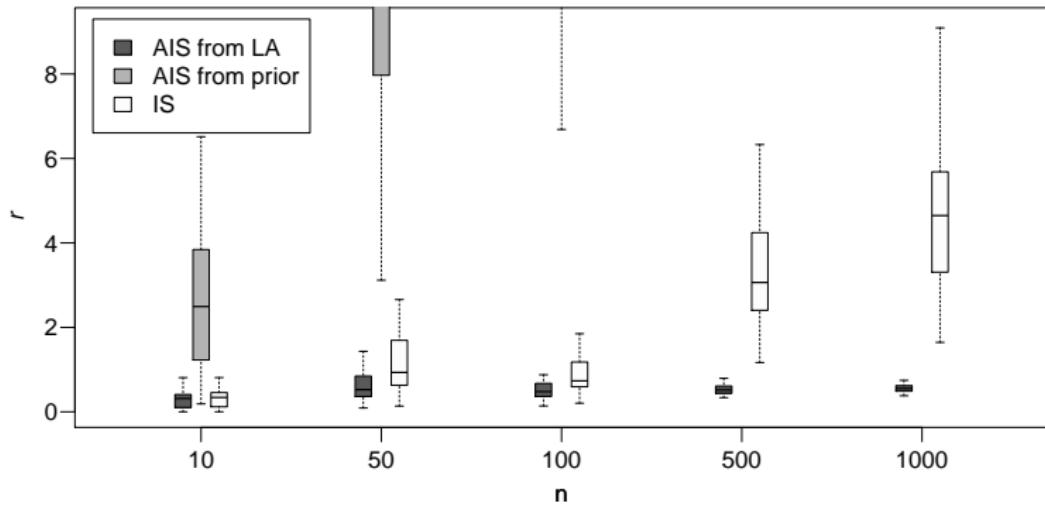
- Annealing from the prior



Comparison between AIS with IS

Analysis of the variance of the AIS and IS estimators

$$r = \text{var} \{ \log_{10} [p(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta})] \}$$



- Marginal likelihood

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta})d\mathbf{f}$$

can only be computed if $p(\mathbf{y}|\mathbf{X}, \mathbf{f})$ is Gaussian

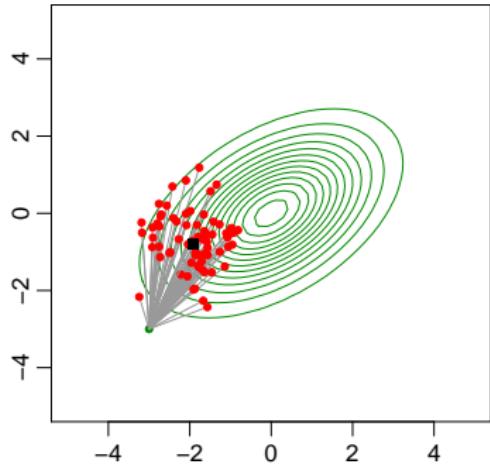
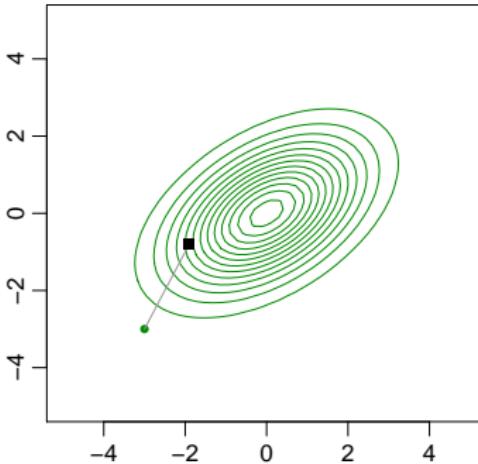
- ... even then

$$\log[p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})] = -\frac{1}{2} \log |K| - \frac{1}{2} \mathbf{y}^T K^{-1} \mathbf{y} + \text{const.}$$

where $K = K(\mathbf{X}, \boldsymbol{\theta})$ is a $n \times n$ dense matrix!

Stochastic Gradient ascent

$$\mathbb{E} \left\{ \widetilde{\nabla_{\theta}} \log[p(\mathbf{y}|\mathbf{X}, \theta)] \right\} = \nabla_{\theta} \log[p(\mathbf{y}|\mathbf{X}, \theta)]$$



Robbins and Monro, AoMS, 1951

Stochastic Gradient ascent

$$\theta' = \theta + \frac{\alpha_t}{2} \widetilde{\nabla_{\theta}} \log[p(\mathbf{y}|\mathbf{X}, \theta)p(\theta)] \quad \alpha_t \rightarrow 0$$

Robbins and Monro, AoMS, 1951

Stochastic Gradient Langevin Dynamics (SGLD) algorithm

$$\boldsymbol{\theta}' = \boldsymbol{\theta} + \frac{\alpha_t}{2} \widetilde{\nabla_{\boldsymbol{\theta}}} \log[p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})p(\boldsymbol{\theta})] + \boldsymbol{\eta}_t \quad \boldsymbol{\eta}_t \sim \mathcal{N}(0, \alpha_t)$$

Stochastic Gradients in GP regression

- Marginal likelihood

$$\log[p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})] = -\frac{1}{2} \log |K| - \frac{1}{2} \mathbf{y}^T K^{-1} \mathbf{y} + \text{const.}$$

- Derivatives wrt $\boldsymbol{\theta}$

$$\frac{\partial \log[p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})]}{\partial \boldsymbol{\theta}_i} = -\frac{1}{2} \text{Tr} \left(K^{-1} \frac{\partial K}{\partial \boldsymbol{\theta}_i} \right) + \frac{1}{2} \mathbf{y}^T K^{-1} \frac{\partial K}{\partial \boldsymbol{\theta}_i} K^{-1} \mathbf{y}$$

Stochastic Gradients in GP regression

- Stochastic estimate of the trace

$$\text{Tr} \left(K^{-1} \frac{\partial K}{\partial \theta_i} \right) = \text{Tr} \left(K^{-1} \frac{\partial K}{\partial \theta_i} \mathbb{E}[\mathbf{r}\mathbf{r}^\top] \right) = \mathbb{E} \left[\mathbf{r}^\top K^{-1} \frac{\partial K}{\partial \theta_i} \mathbf{r} \right]$$

with $\mathbb{E}[\mathbf{r}\mathbf{r}^\top] = I$

Stochastic Gradients in GP regression

- Stochastic estimate of the trace

$$\text{Tr} \left(K^{-1} \frac{\partial K}{\partial \theta_i} \right) = \text{Tr} \left(K^{-1} \frac{\partial K}{\partial \theta_i} \mathbb{E}[\mathbf{r}\mathbf{r}^\top] \right) = \mathbb{E} \left[\mathbf{r}^\top K^{-1} \frac{\partial K}{\partial \theta_i} \mathbf{r} \right]$$

with $\mathbb{E}[\mathbf{r}\mathbf{r}^\top] = I$

- Stochastic gradient

$$-\frac{1}{2N_r} \sum_{i=1}^{N_r} \mathbf{r}^{(i)\top} K^{-1} \frac{\partial K}{\partial \theta_i} \mathbf{r}^{(i)} + \frac{1}{2} \mathbf{y}^\top K^{-1} \frac{\partial K}{\partial \theta_i} K^{-1} \mathbf{y}$$

- Linear systems only!

Solving linear systems

- Linear systems:

$$Ks = b$$

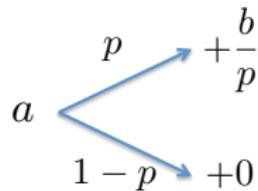
- Can be solved using conjugate gradient:

$$s = \arg \min_x \left(\frac{1}{2} x^\top K x - x^\top b \right)$$

- Iterative update $s = s_0 + \delta_1 + \dots + \delta_T$
- Requires only Kv multiplications! $O(n^2)$ time
- No need to store K ! $O(n)$ space

- Accelerate the solution of dense linear systems
- ... returning an unbiased estimate of the solution

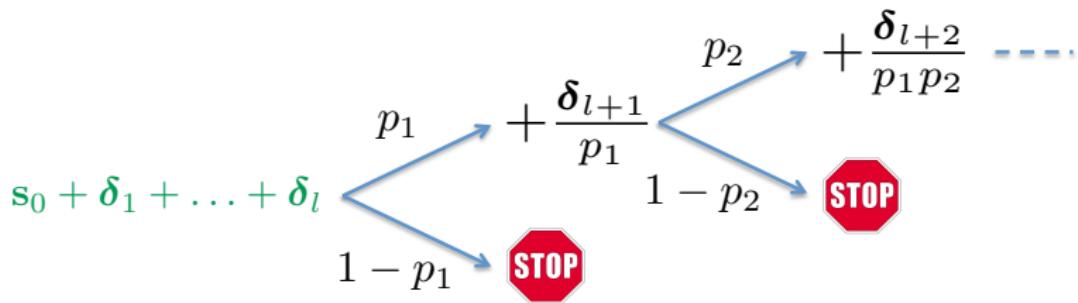
- Accelerate the solution of dense linear systems
- ... returning an unbiased estimate of the solution
- Basic idea - unbiased estimator for generic sums $a + b$:



- Full CG solution:

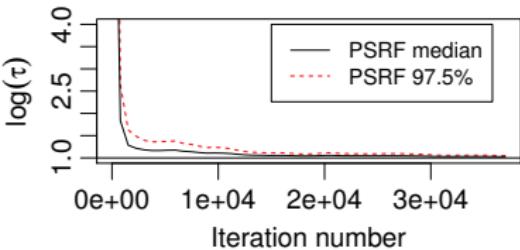
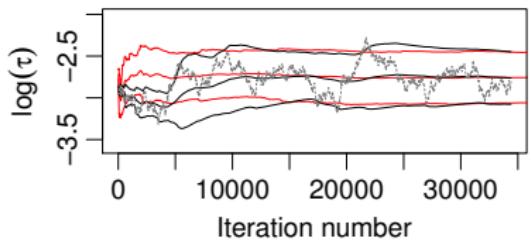
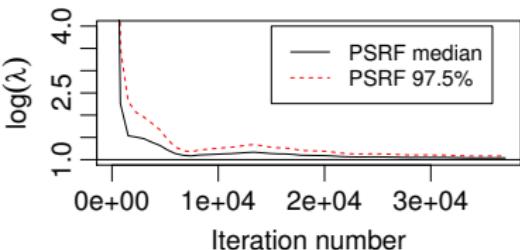
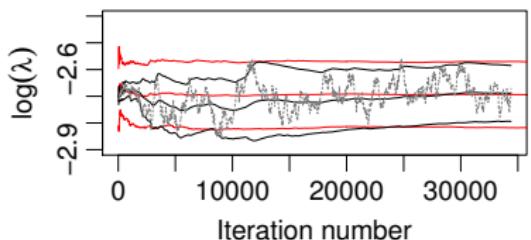
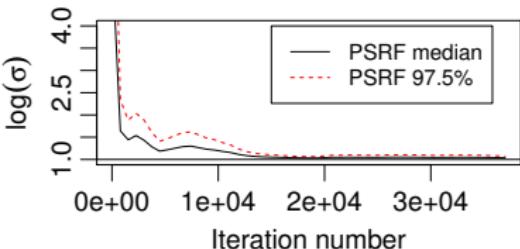
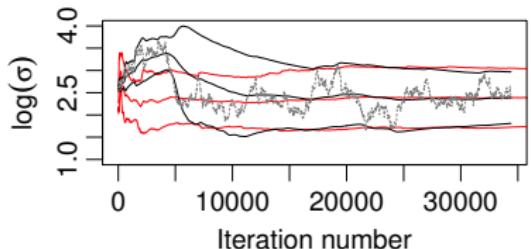
$$\mathbf{s} = \mathbf{s}_0 + \boldsymbol{\delta}_1 + \dots + \boldsymbol{\delta}_l + \boldsymbol{\delta}_{l+1} \dots + \boldsymbol{\delta}_T$$

- ULISSE:

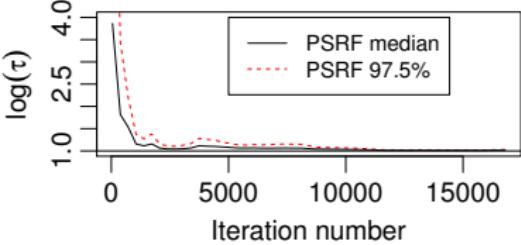
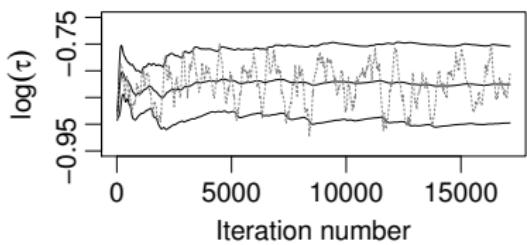
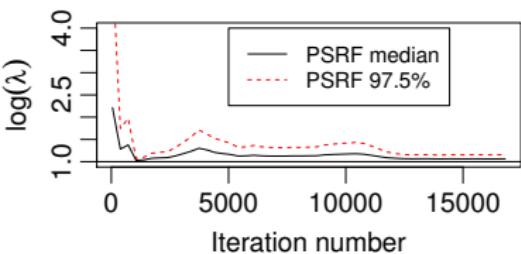
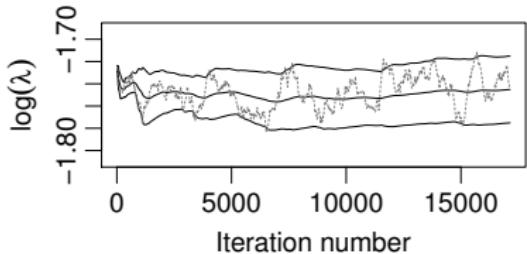
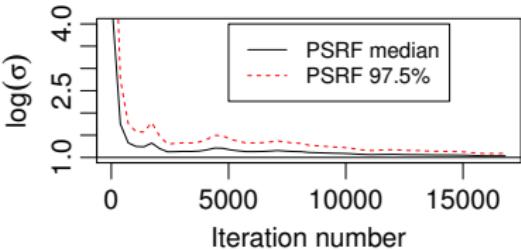
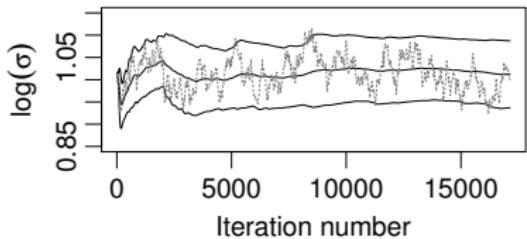


- Final solution is an unbiased estimate of \mathbf{s} !

Comparison with MCMC - Concrete dataset - $n \approx 1K$

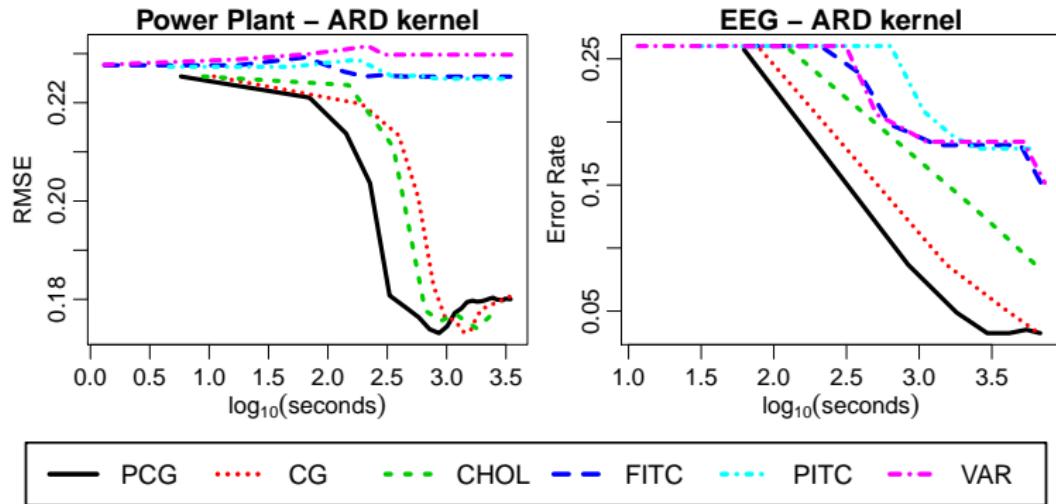


Larger n - Census dataset - $n \approx 23K$



Preconditioning Kernel Matrices

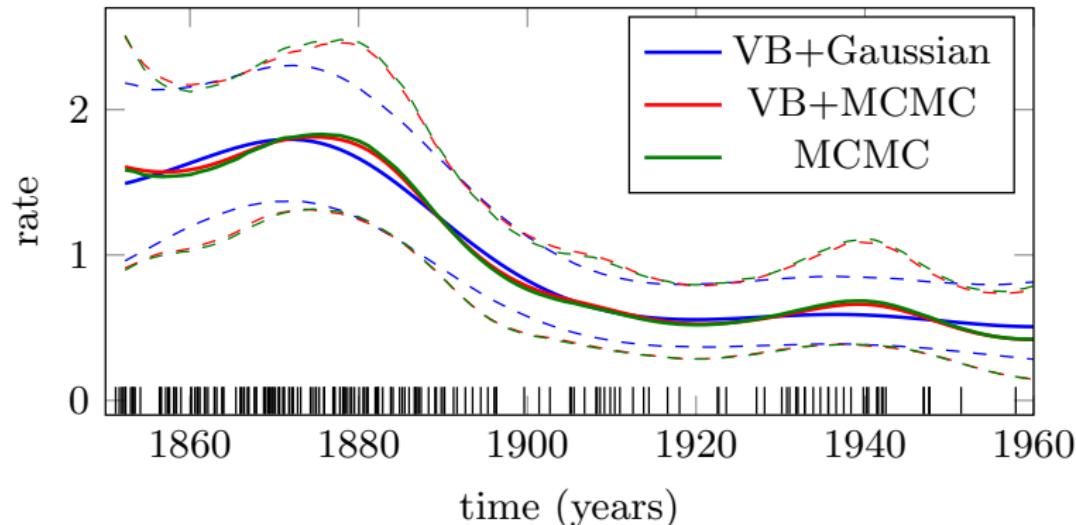
- Stochastic Gradient Optimization



Cutajar, Osborne, Cunningham, Filippone, ICML, 2016

Marrying Sparse Variational and MCMC learning for GPs

- Coal mining disaster data
- Analysis with 30 inducing points

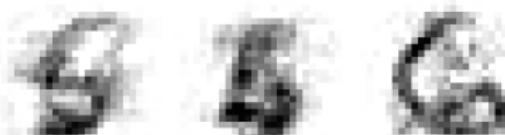


Marrying Sparse Variational and MCMC learning for GPs

- MNIST data
- Three inducing points before optimizing their position



- ... and after

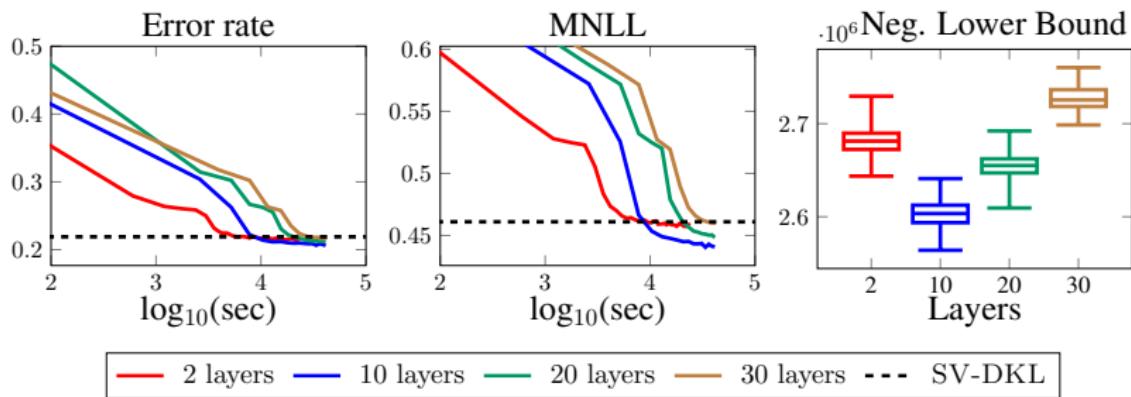


Hensman, Matthews, Filippone, Ghahramani, *NIPS*, 2015

Deep Gaussian Processes

- Layer-wise random feature expansion
- Stochastic variational inference

Airline dataset
($n = 5M+$, $d = 8$)



Cutajar, Bonilla, Michiardi, Filippone, *arXiv*, 2016

Conclusions and ongoing work

- GPs form the basis of flexible and interpretable nonparametric statistical models

Conclusions and ongoing work

- GPs form the basis of flexible and interpretable nonparametric statistical models
- GP learning is generally computationally intractable

Conclusions and ongoing work

- GPs form the basis of flexible and interpretable nonparametric statistical models
- GP learning is generally computationally intractable
- Introducing approximations can severely affect quantification of uncertainty and reduce accuracy

Conclusions and ongoing work

- GPs form the basis of flexible and interpretable nonparametric statistical models
- GP learning is generally computationally intractable
- Introducing approximations can severely affect quantification of uncertainty and reduce accuracy
- Approximate **unbiased** computations as a way to construct practical and scalable inference of GP models

References and Acknowledgments

Thank you!

