Unbiased computations for tractable and scalable learning of Gaussian processes

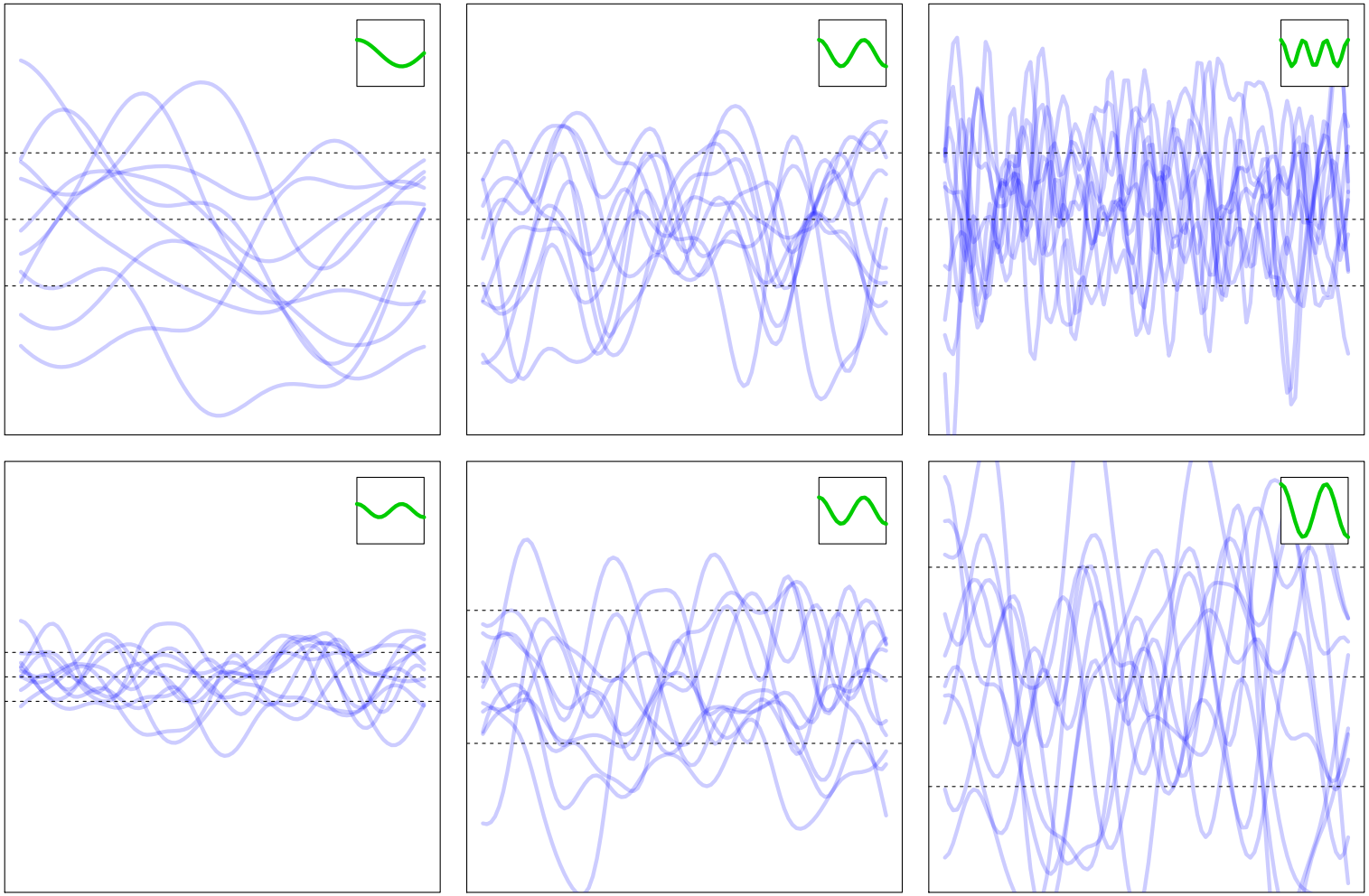
Maurizio Filippone

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

March 29th, 2017

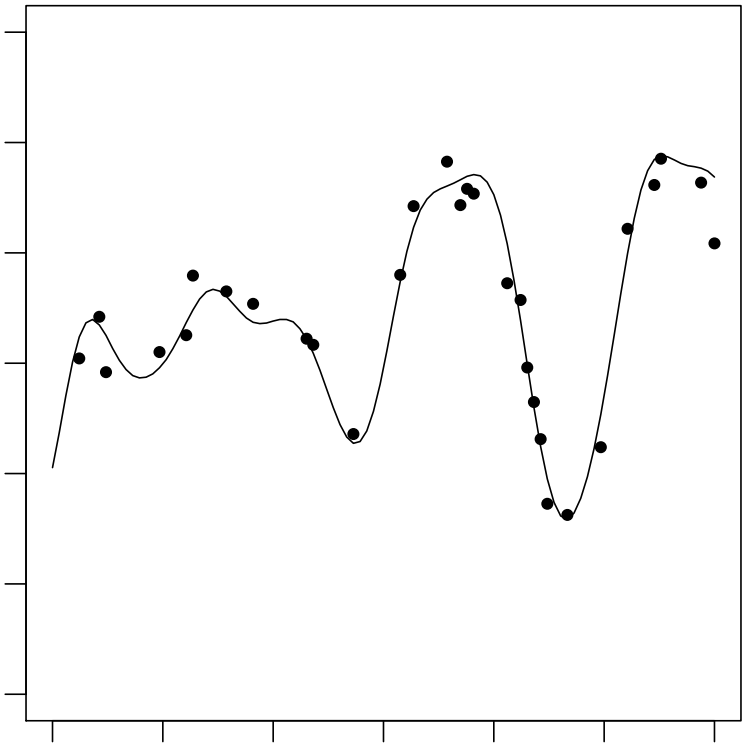
Gaussian Processes - Priors over Functions

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



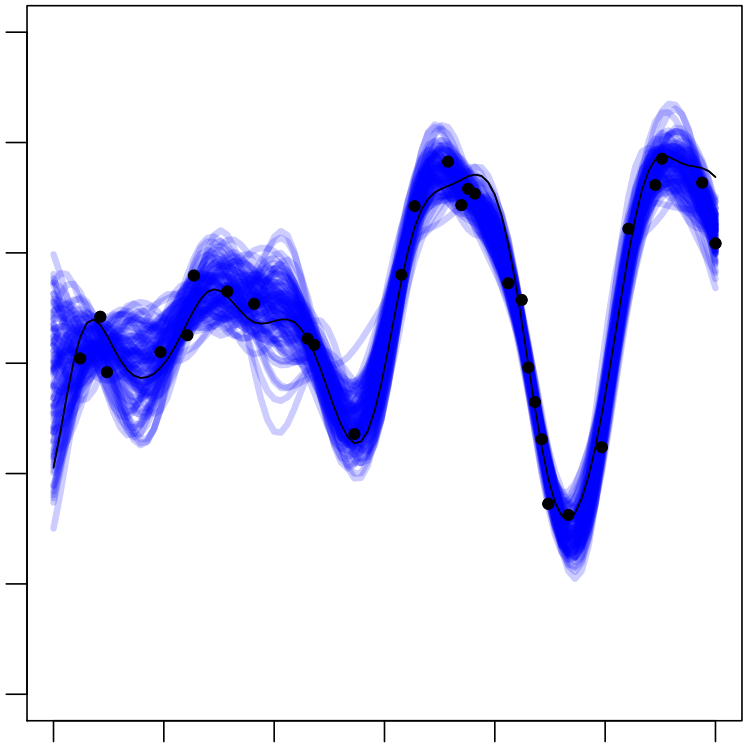
Gaussian Processes

- Regression example

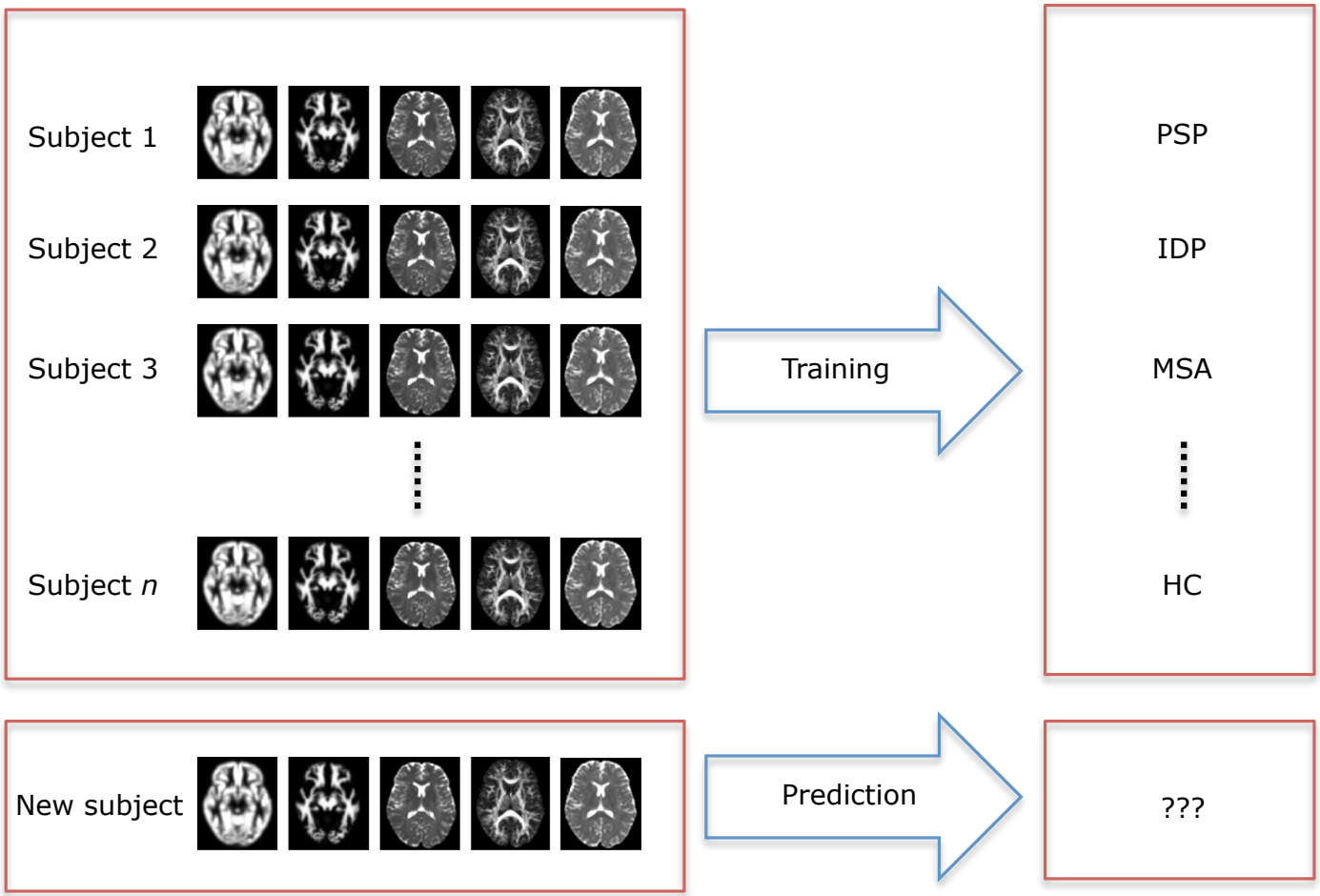


Gaussian Processes

- Regression example



Motivating Application



HC - Healthy control

MSA - Multiple system atrophy

PSP - Progressive Supranuclear Palsy

IDP - Idiopathic Parkinson's disease

Filippone, Marquand et al., *AoAS*, 2012

Multiclass classification with multiple sources

- Multiclass classification based on GPs

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

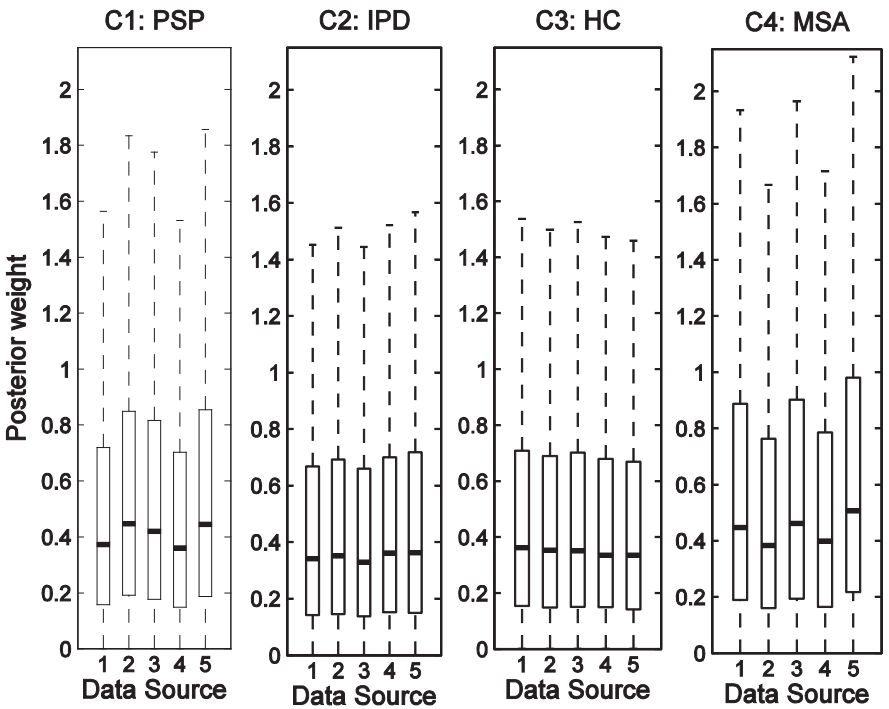
- unknown function modeled using GPs
- Covariance based on source-dependent covariances S_k

$$\sum_{k=1}^{K} w_{ck} S_k(subject_i, subject_j)$$

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

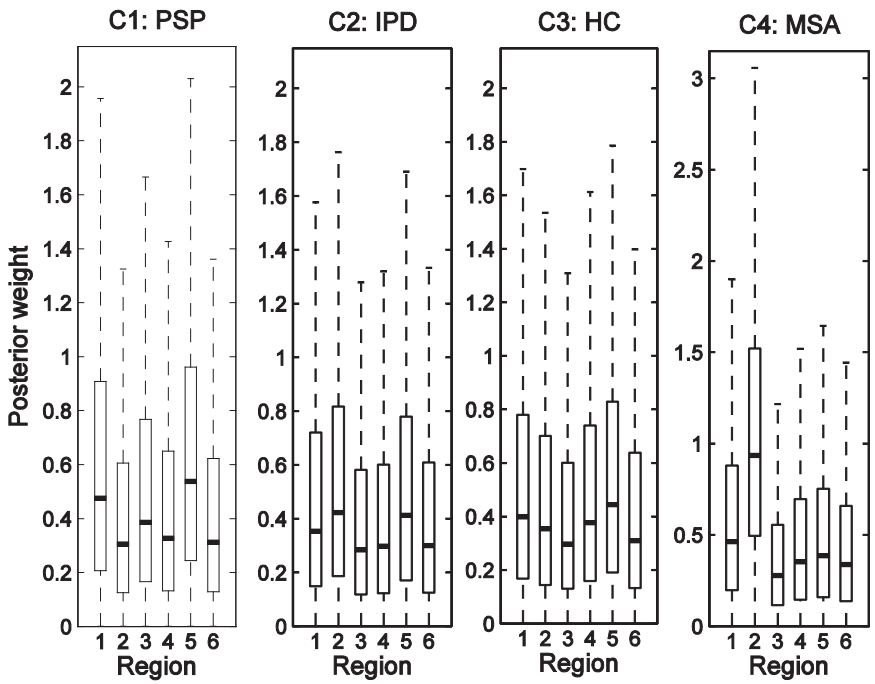
Multiclass classification with multiple regions

Analysis of brain regions

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

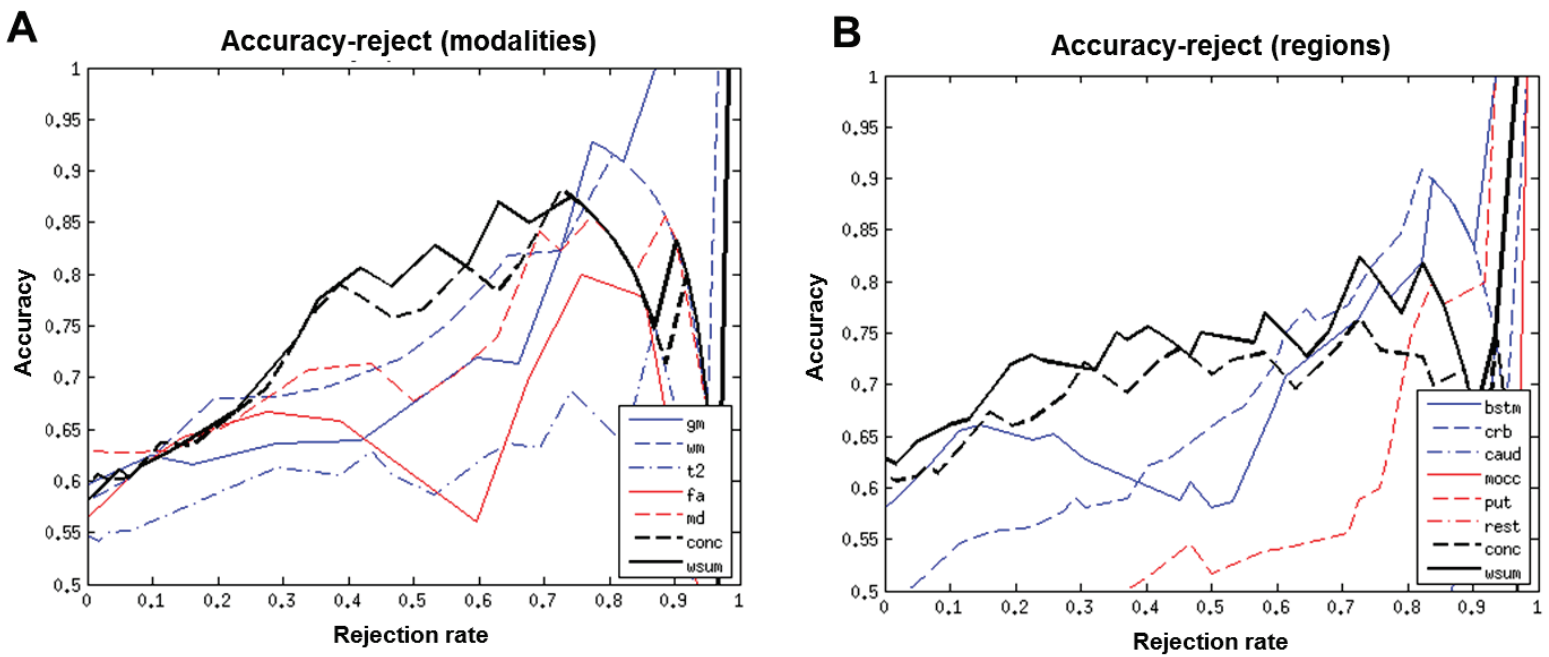
Parkinsonian disorders data - multiple regions classification

Method	Accuracy
GP classifier	0.614
SimpleMKL	0.229



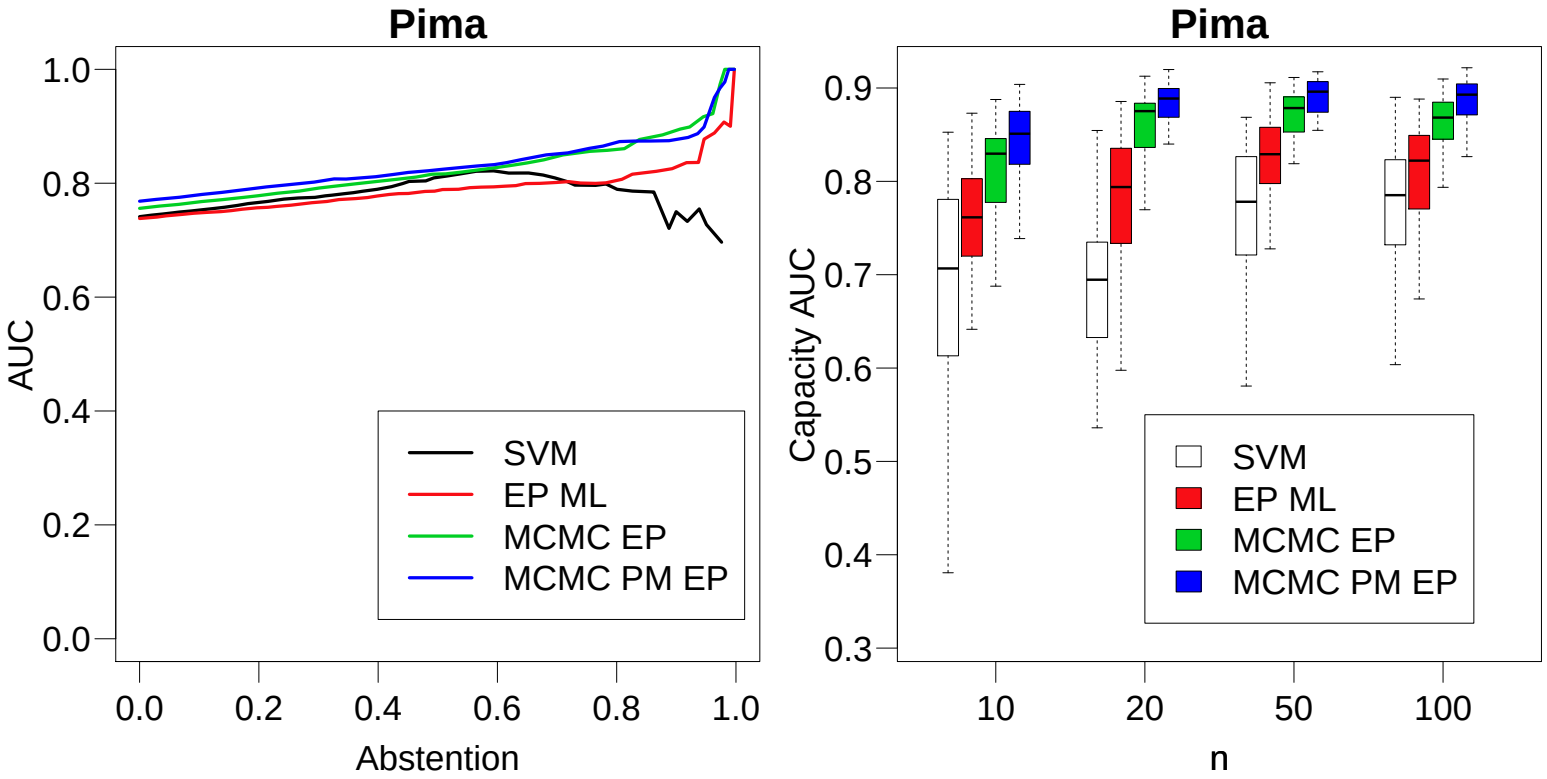
Filippone, Marquand et al., *AoAS*, 2012

Reject Option



Filippone, Marquand et al., *AoAS*, 2012

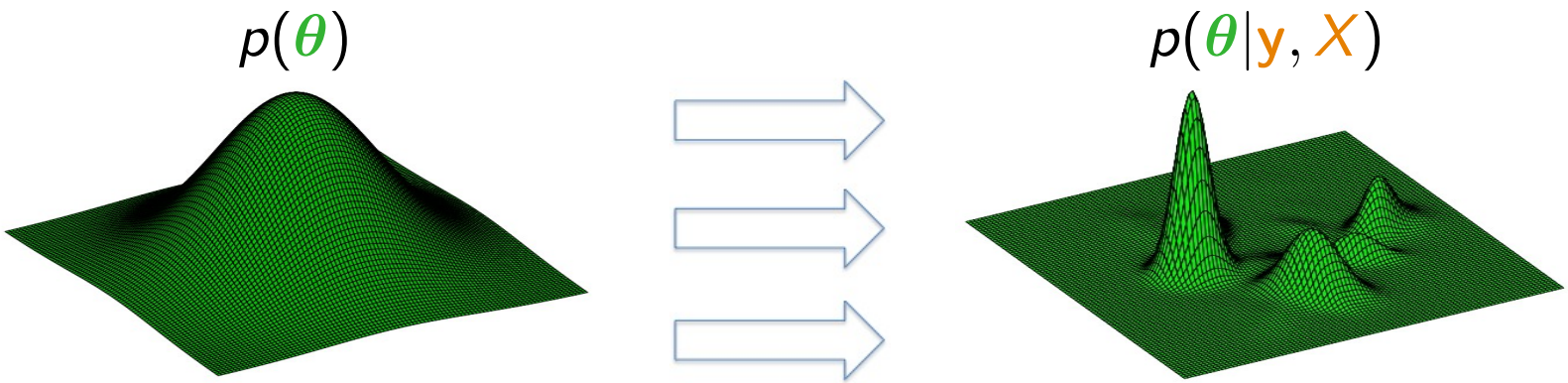
Does being Bayesian buy you anything?



Filippone and Girolami, *IEEE-TPAMI*, 2014

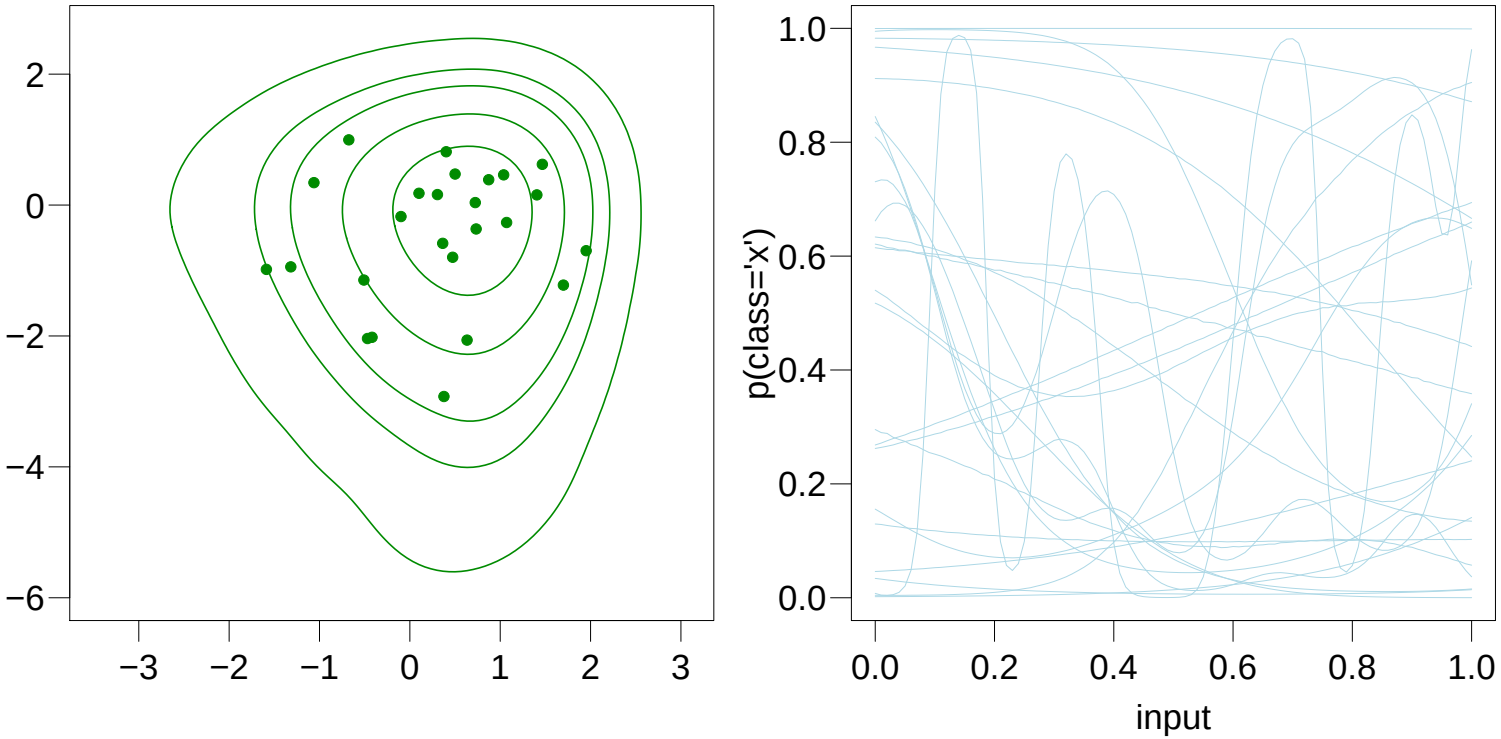
Bayesian Gaussian Processes

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

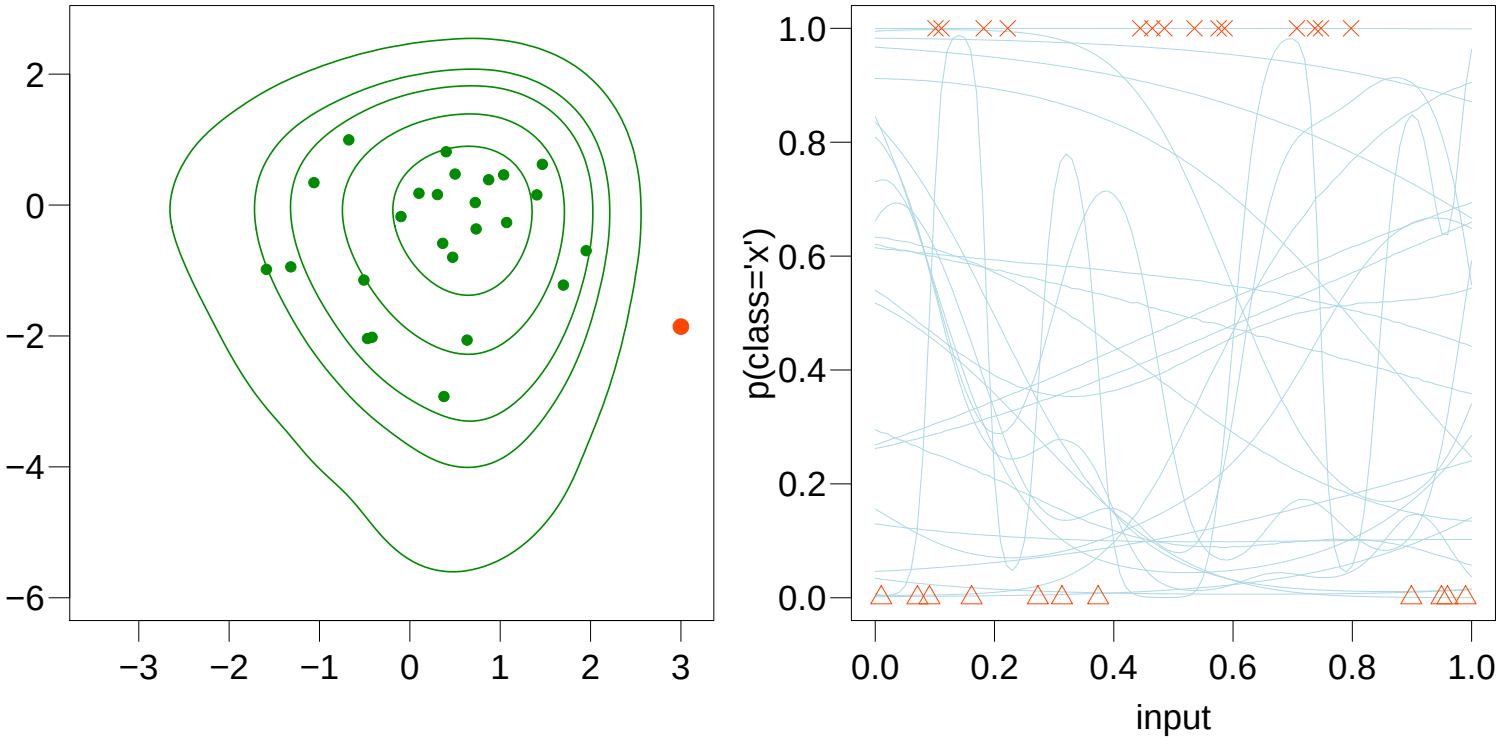


$$p(\theta|\mathbf{y}, X) = \frac{p(\mathbf{y}|X, \theta)p(\theta)}{\int p(\mathbf{y}|X, \theta)p(\theta)d\theta}$$

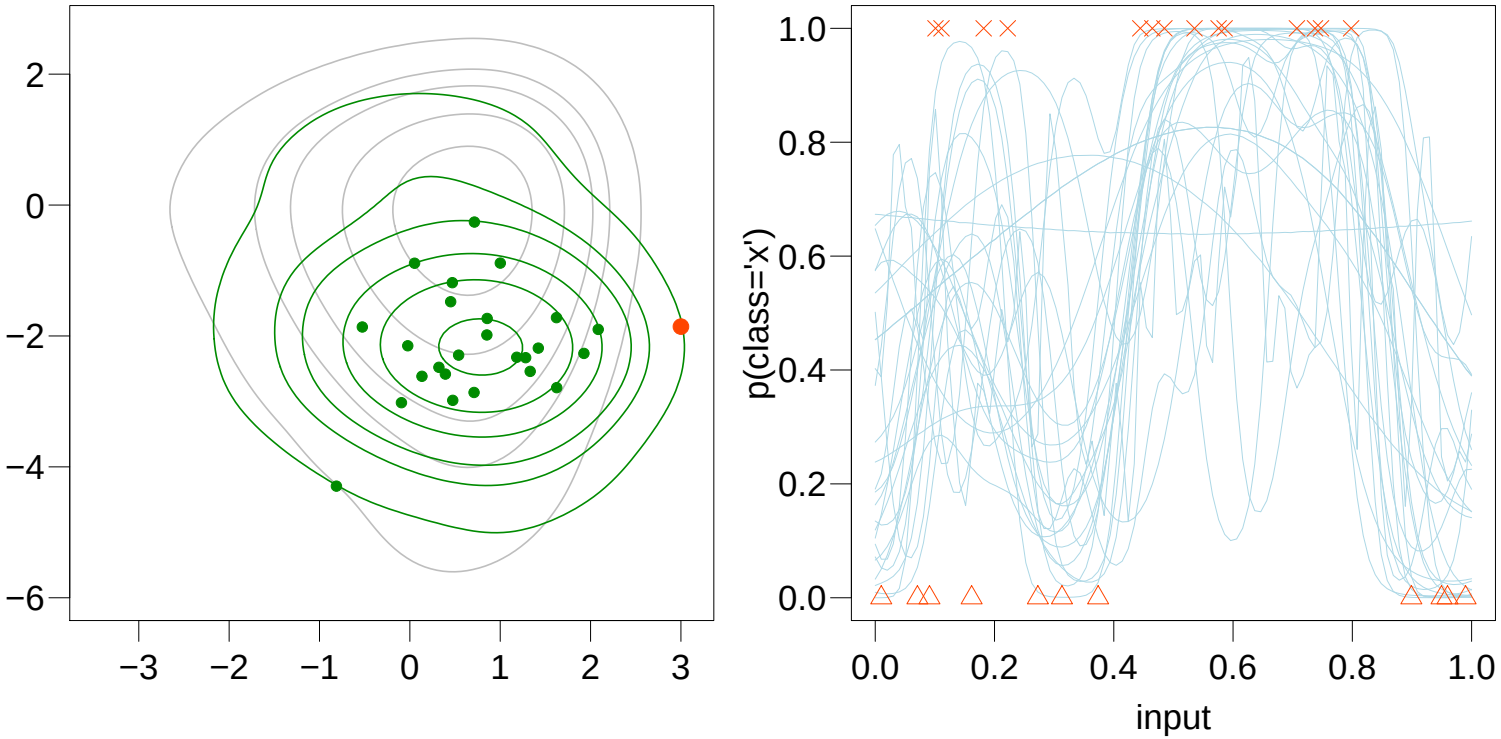
Bayesian Gaussian Processes - Prior



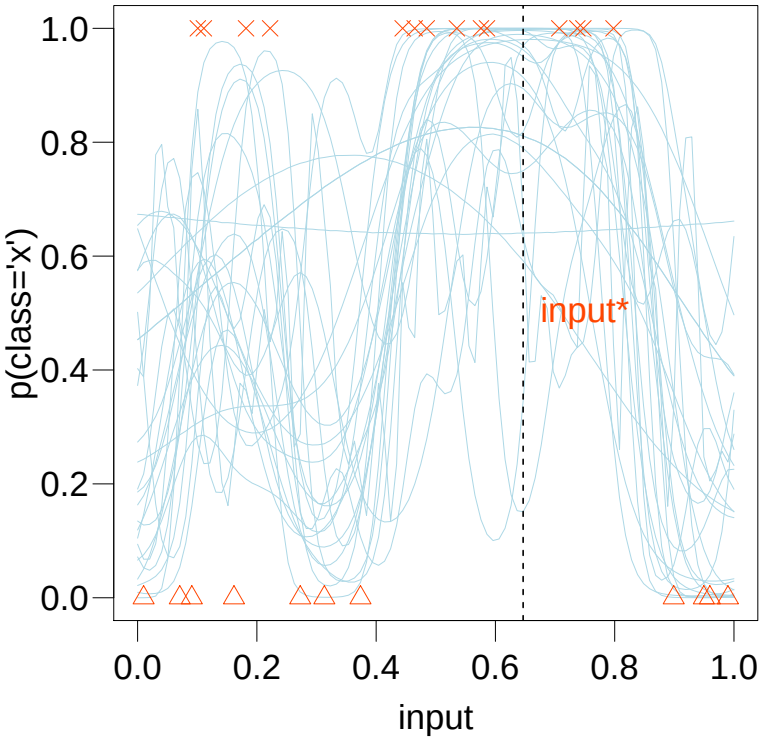
Bayesian Gaussian Processes - Data

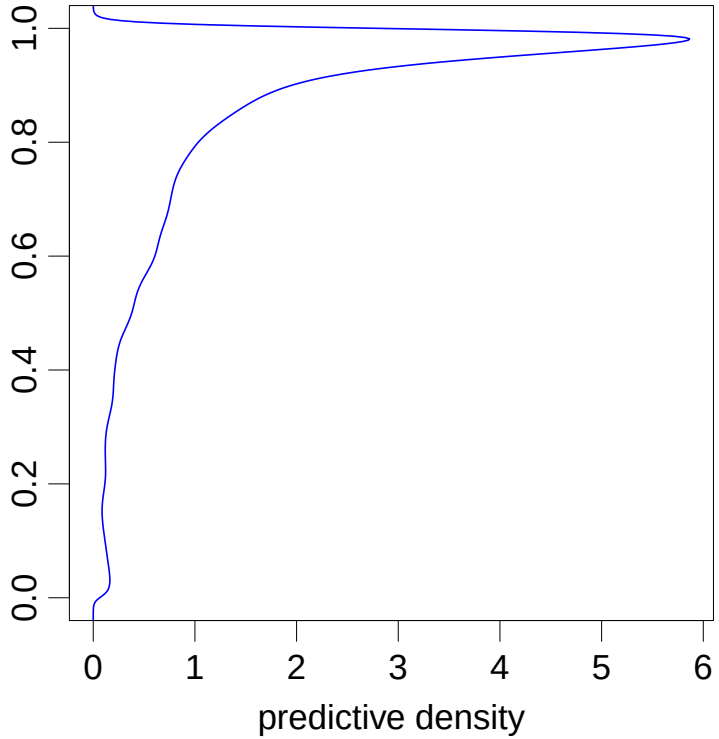


Bayesian Gaussian Processes - Posterior



Bayesian Gaussian Processes - Predictions





Bayesian Gaussian Processes - Predictions

- Predictions for new data

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

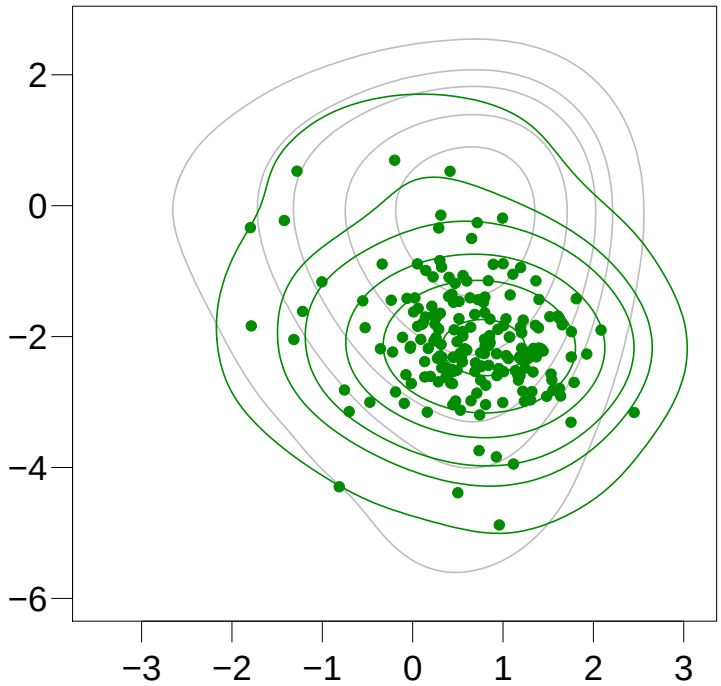
- Monte Carlo integration:

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

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

Bayesian Gaussian Processes - Predictions

- Draw samples according to the posterior density



Markov chain Monte Carlo

- Bayesian inference

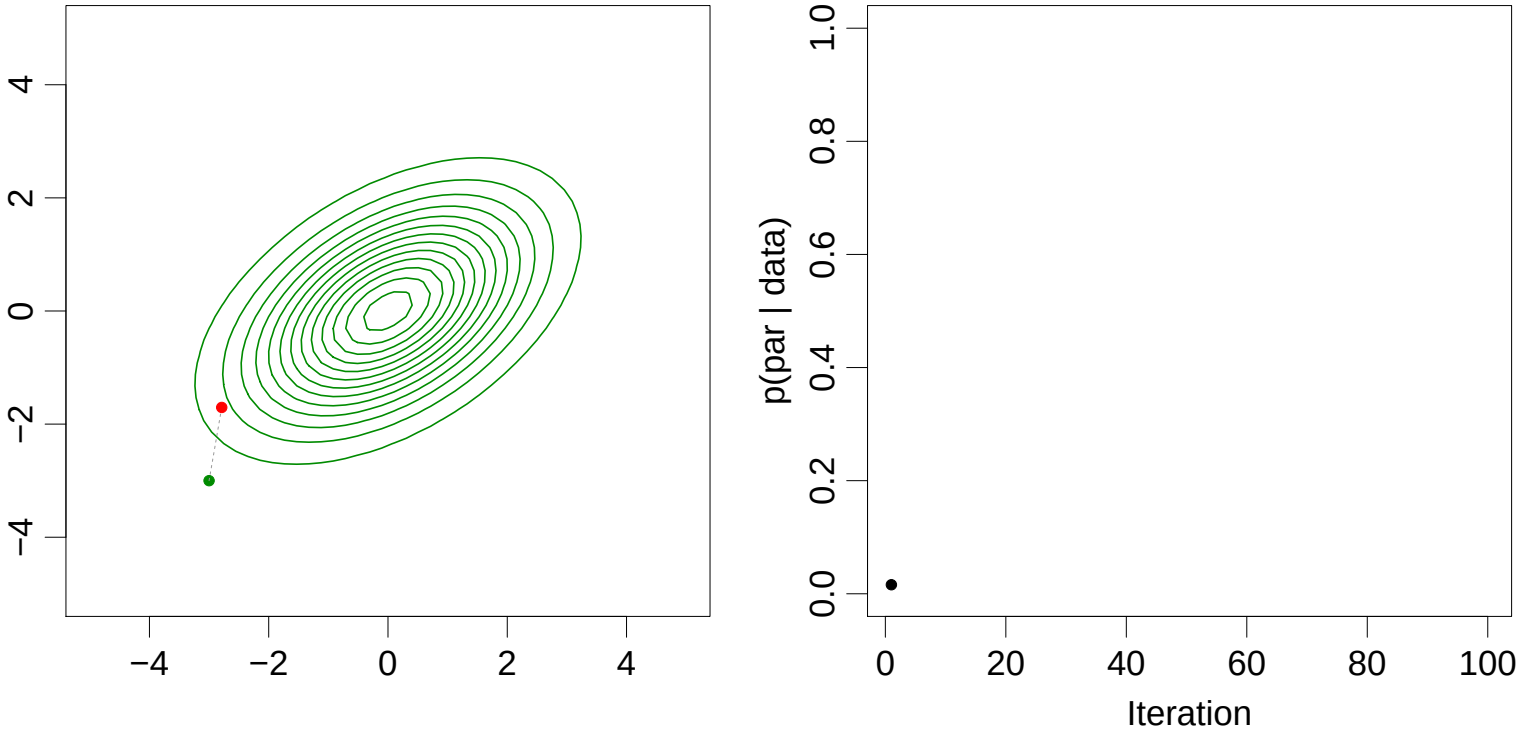
$$p(\theta|\mathbf{y}, X) = \frac{p(\mathbf{y}|X, \theta)p(\theta)}{\int p(\mathbf{y}|X, \theta)p(\theta)d\theta}$$

- Random walk sampler - accept a proposal with probability

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

Markov chain Monte Carlo - Random walk example

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



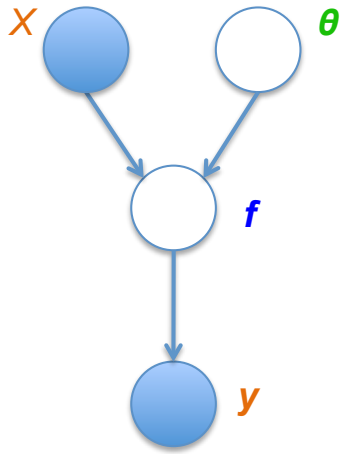
Metropolis et al., *JoCP*, 1953 - Hastings, *Biometrika*, 1970

Challenges and Limitations

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

Challenges and Limitations

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



Marginal likelihood of GP models

- Marginal likelihood

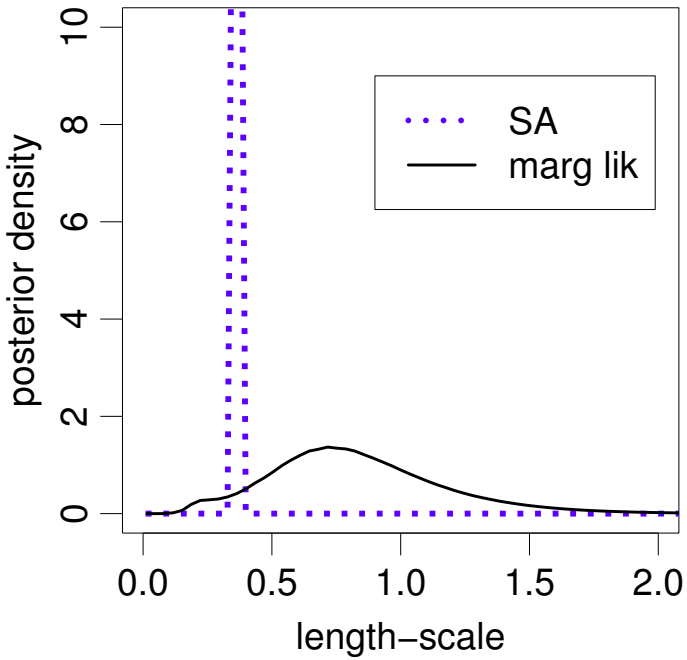
$$p(\mathbf{y}|X,\theta) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X,\theta)d\mathbf{f}$$

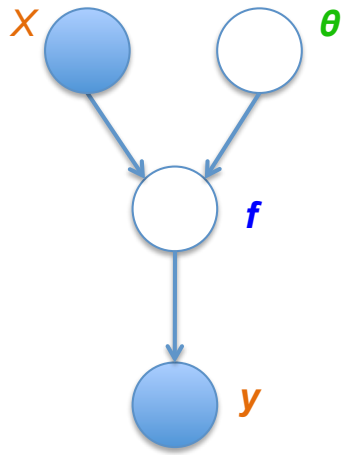
  can only be computed if $p(\mathbf{y}|\mathbf{f})$ is Gaussian
- What if $p(\mathbf{y}|\mathbf{f})$ is **not** Gaussian?

Challenges in MCMC for GPMs - Structure

Sufficient Augmentation (SA) scheme - alternate between:

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





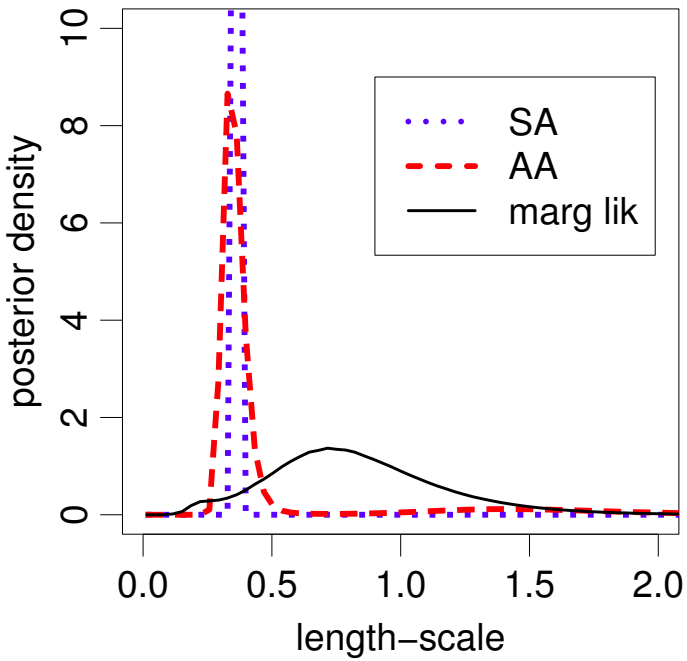
Murray and Adams, *NIPS*, 2010 - Filippone et al., *Mach. Learn.*, 2013.

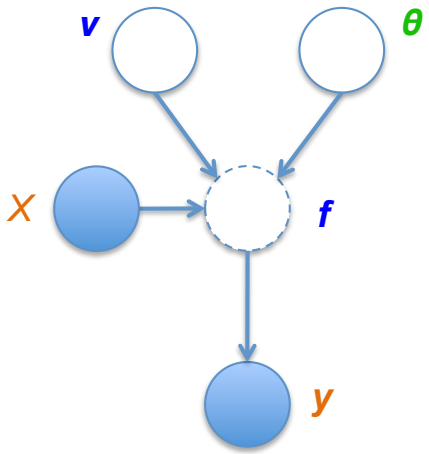
Mitigating coupling effect through reparameterization

Ancillary Augmentation (AA) scheme - reparametrization:

$$K = L\,L^{\top} \qquad \nu = L^{-1}\,\mathbf{f}$$

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



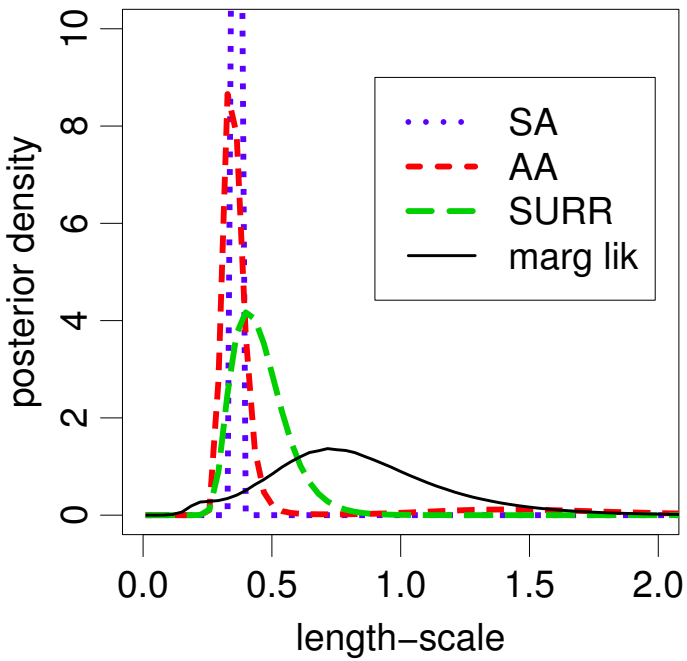


Murray and Adams, *NIPS*, 2010 - Filippone et al., *Mach. Learn.*, 2013.

Mitigating coupling effect through reparameterization

Surrogate data model (SURR):

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

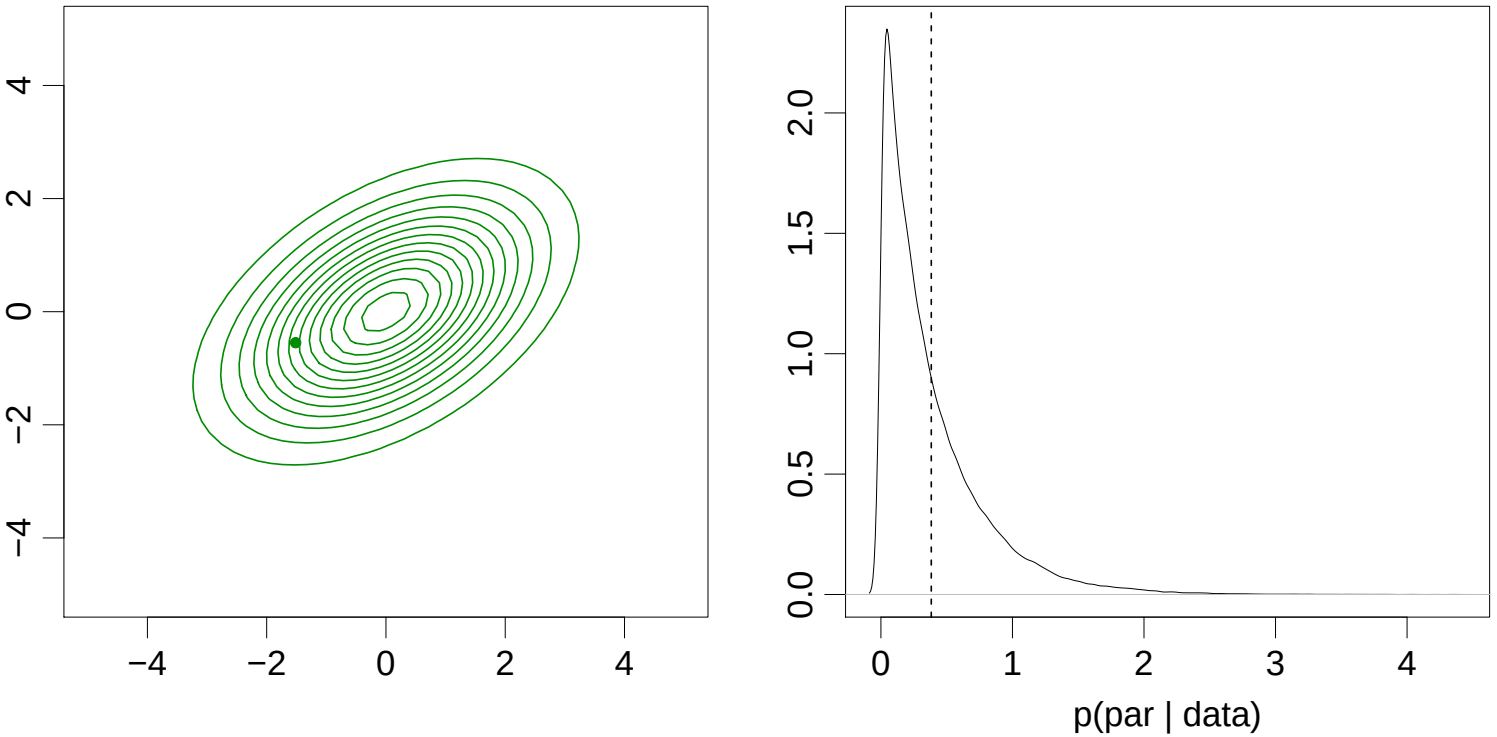


$$\mathbf{g} = \varphi(\mathbf{f}, \boldsymbol{\theta})$$

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

"Noisy" Markov chain Monte Carlo

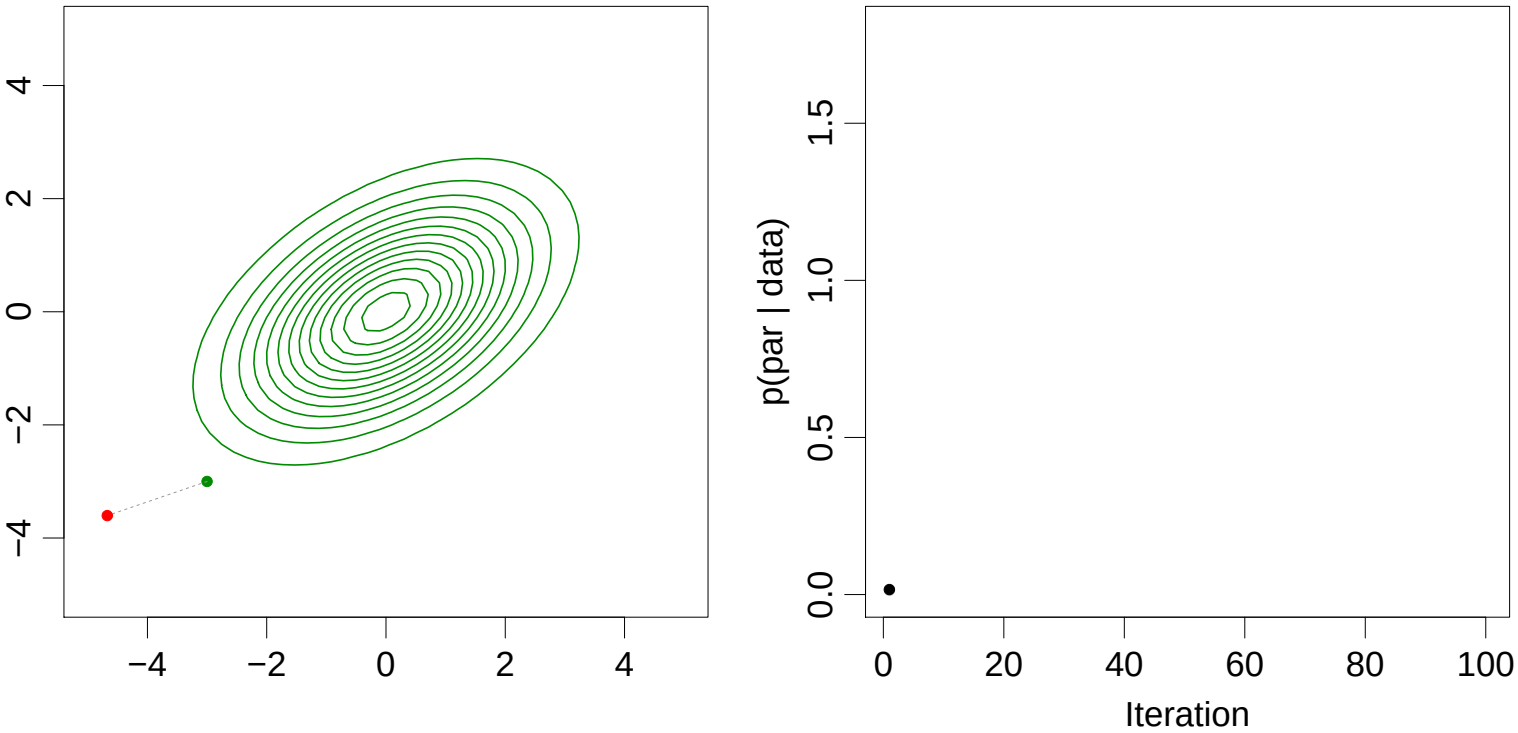
$$\mathrm{E}\left\{\tilde{p}(\mathbf{y}|X,\theta)\right\} = p(\mathbf{y}|X,\theta)$$



Andrieu and Roberts, *AoS*, 2009

"Noisy" Markov chain Monte Carlo

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

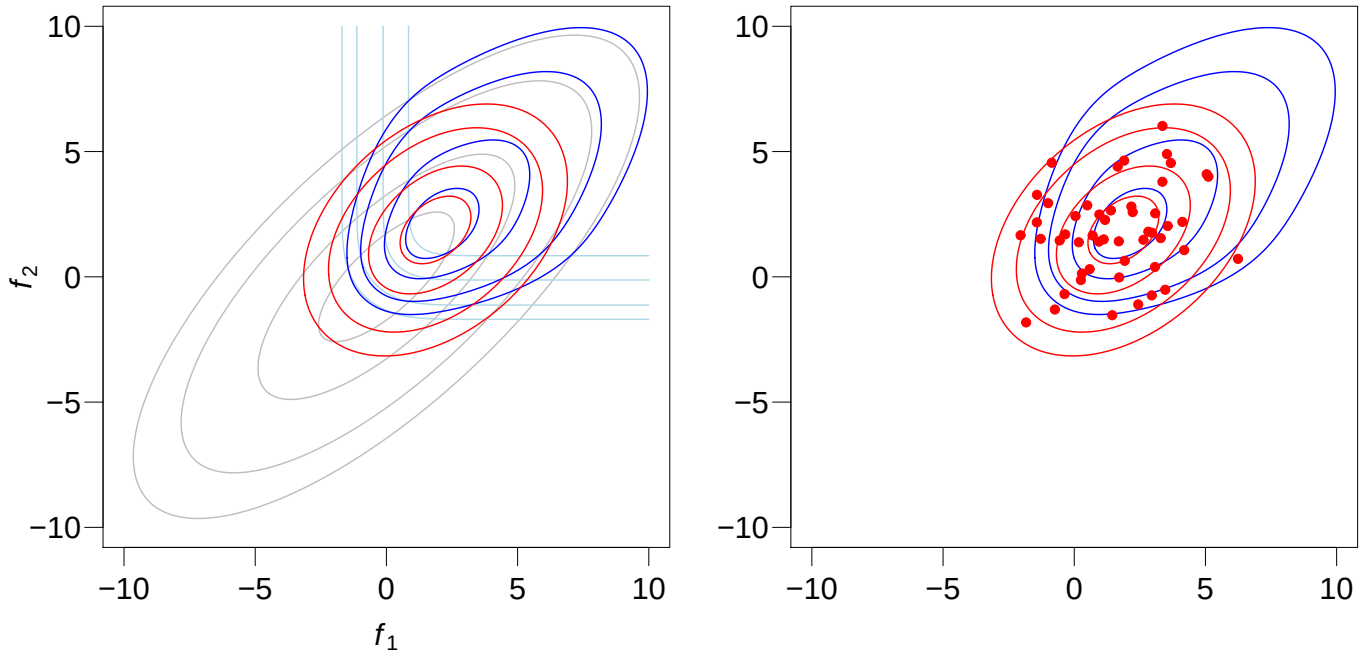


Andrieu and Roberts, *AoS*, 2009

Importance Sampling estimator

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

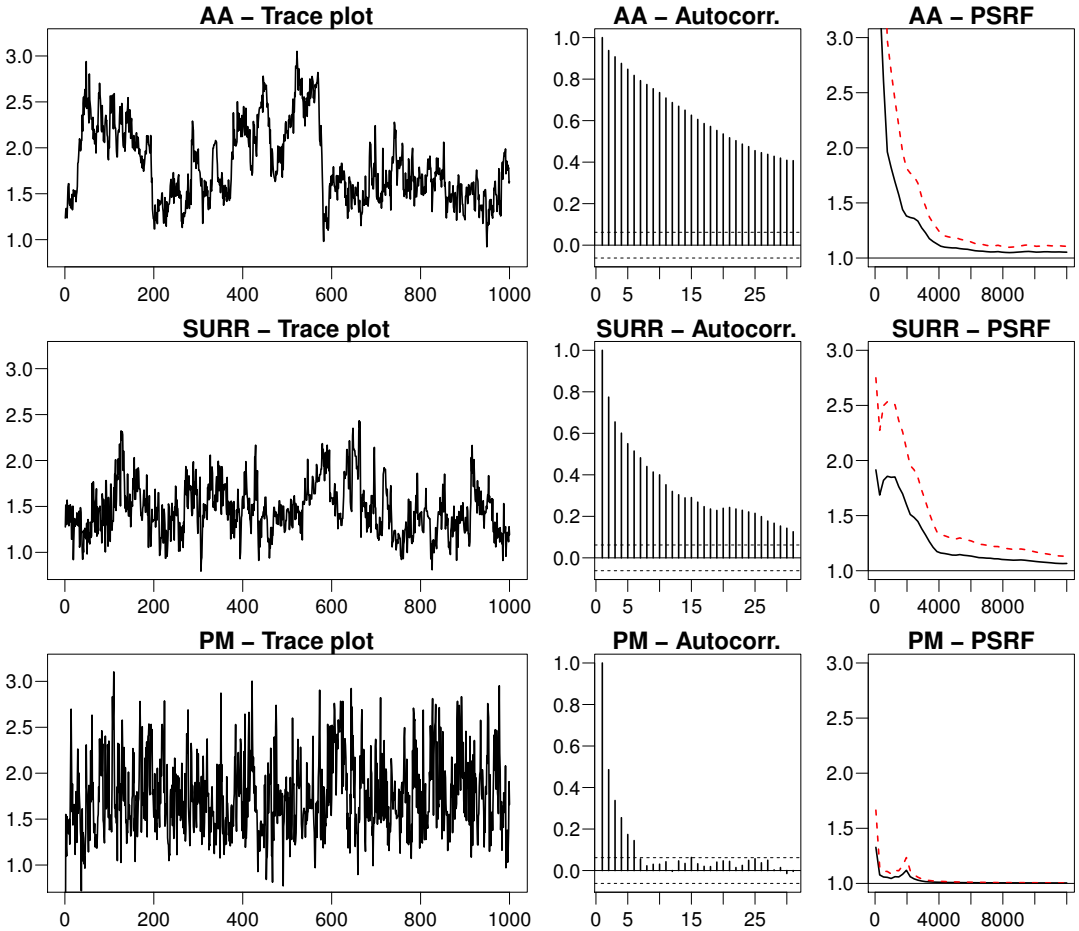
$$\tilde{p}(\mathbf{y}|X,\boldsymbol{\theta}) = \frac{1}{N_{\text{IS}}} \sum_{i=1}^{N_{\text{IS}}} \frac{p(\mathbf{y}|X,\mathbf{f}^{(i)})p(\mathbf{f}^{(i)}|\boldsymbol{\theta})}{q(\mathbf{f}^{(i)})}$$



Filippone and Girolami, *IEEE-TPAMI*, 2014 - Filippone, *ICPR*, 2014

Convergence speed and efficiency

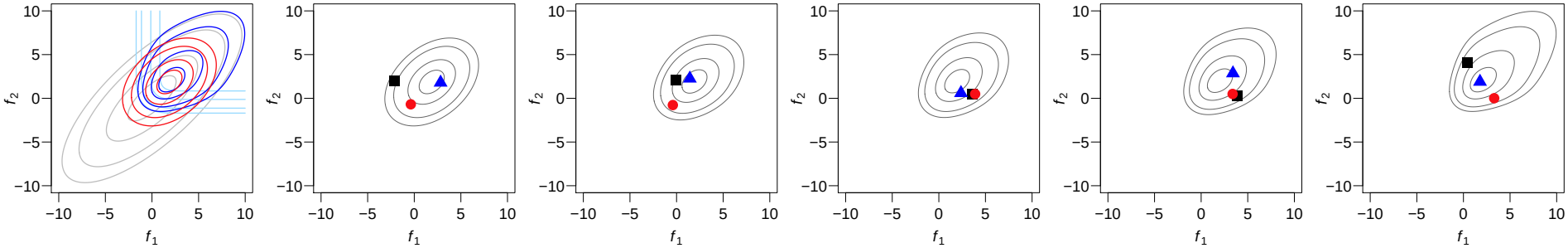
Abalone data set (two classes) $n = 2835$ - inference of length-scale



Filippone and Girolami, *IEEE-TPAMI*, 2014 - Filippone, *ICPR*, 2014

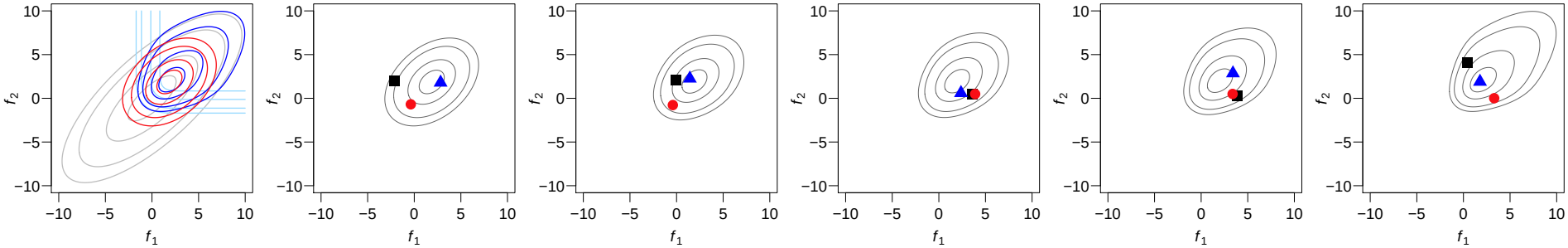
Annealed Importance Sampling estimator

- Annealing from an approximation

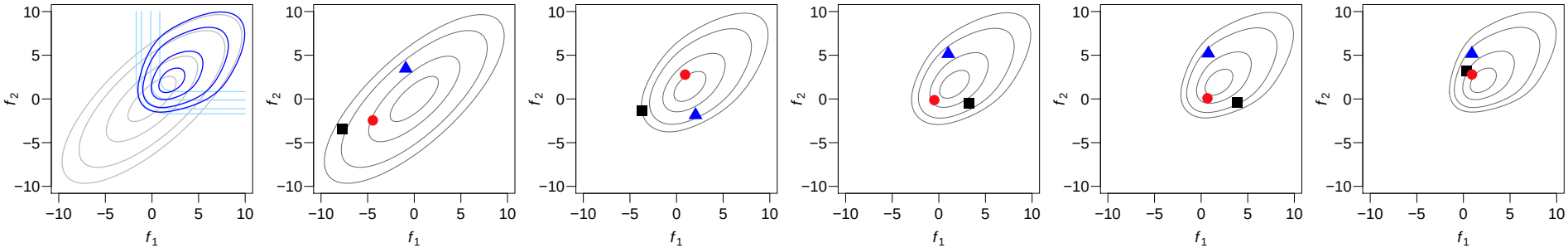


Annealed Importance Sampling estimator

- Annealing from an approximation



- Annealing from the prior

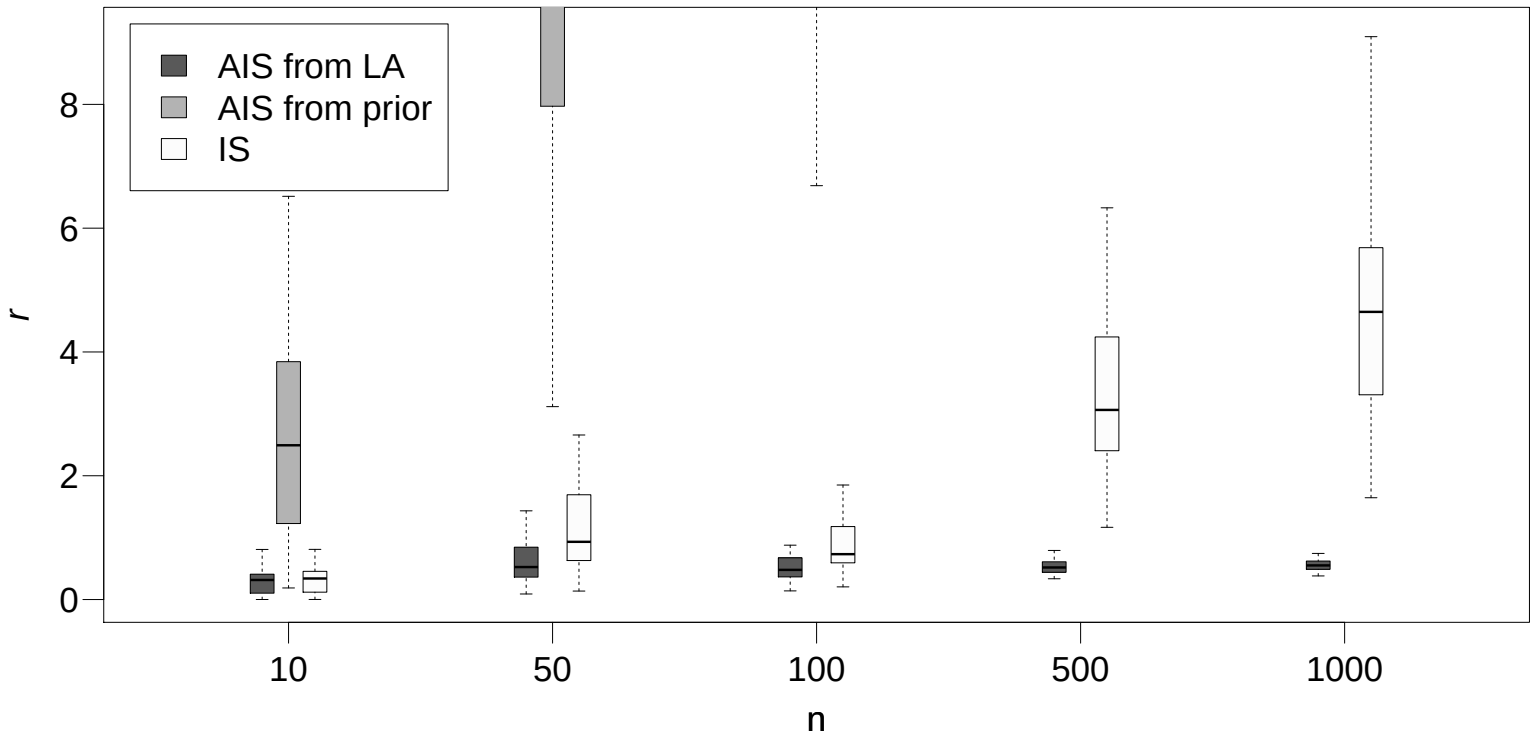


Filippone, *ICPR*, 2014

Comparison between AIS with IS

Analysis of the variance of the AIS and IS estimators

- r is the variance of the $\log_{10}$ marginal likelihood



Filippone, *ICPR*, 2014

Marginal likelihood in GP models

- Marginal likelihood

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

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

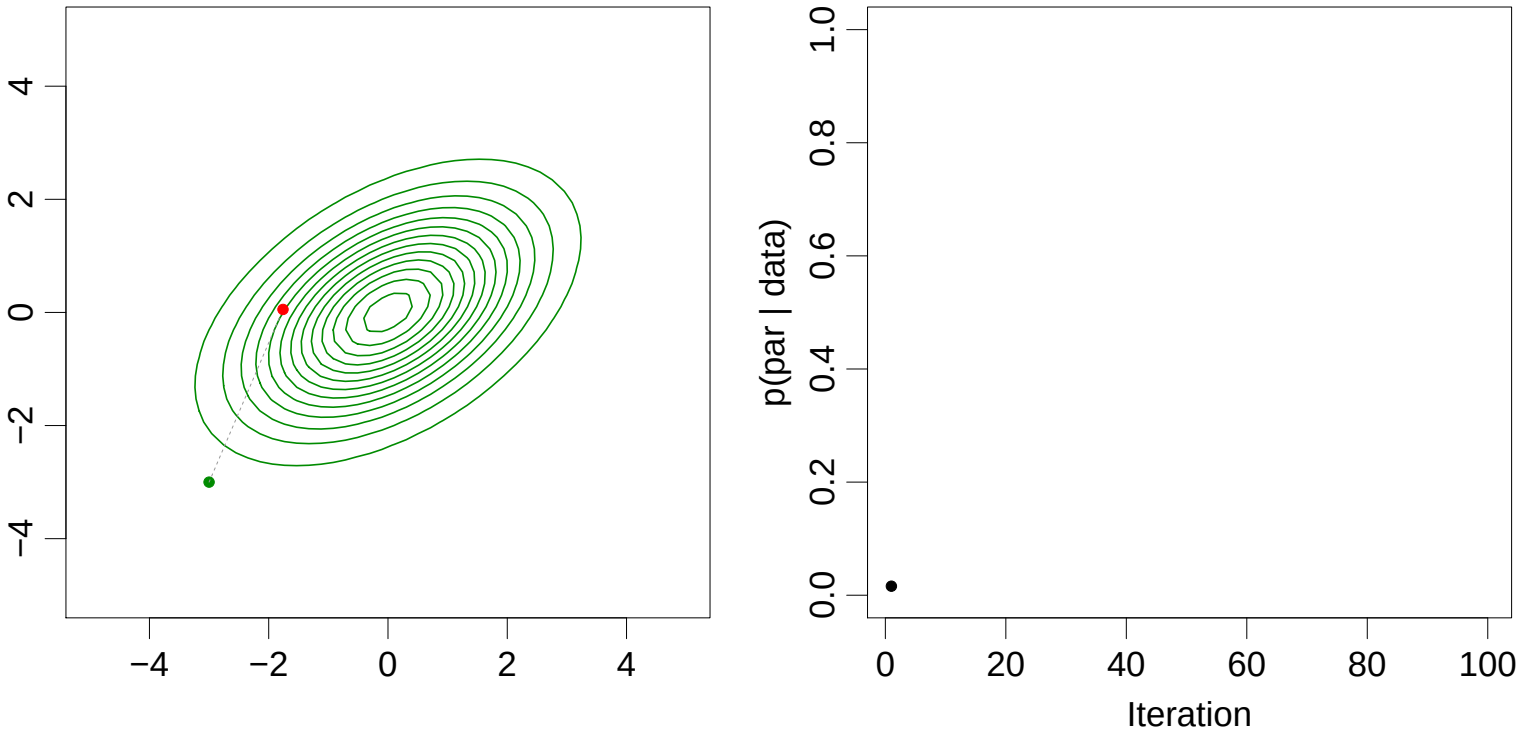
- ... even then

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

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

Stochastic Gradient ascent

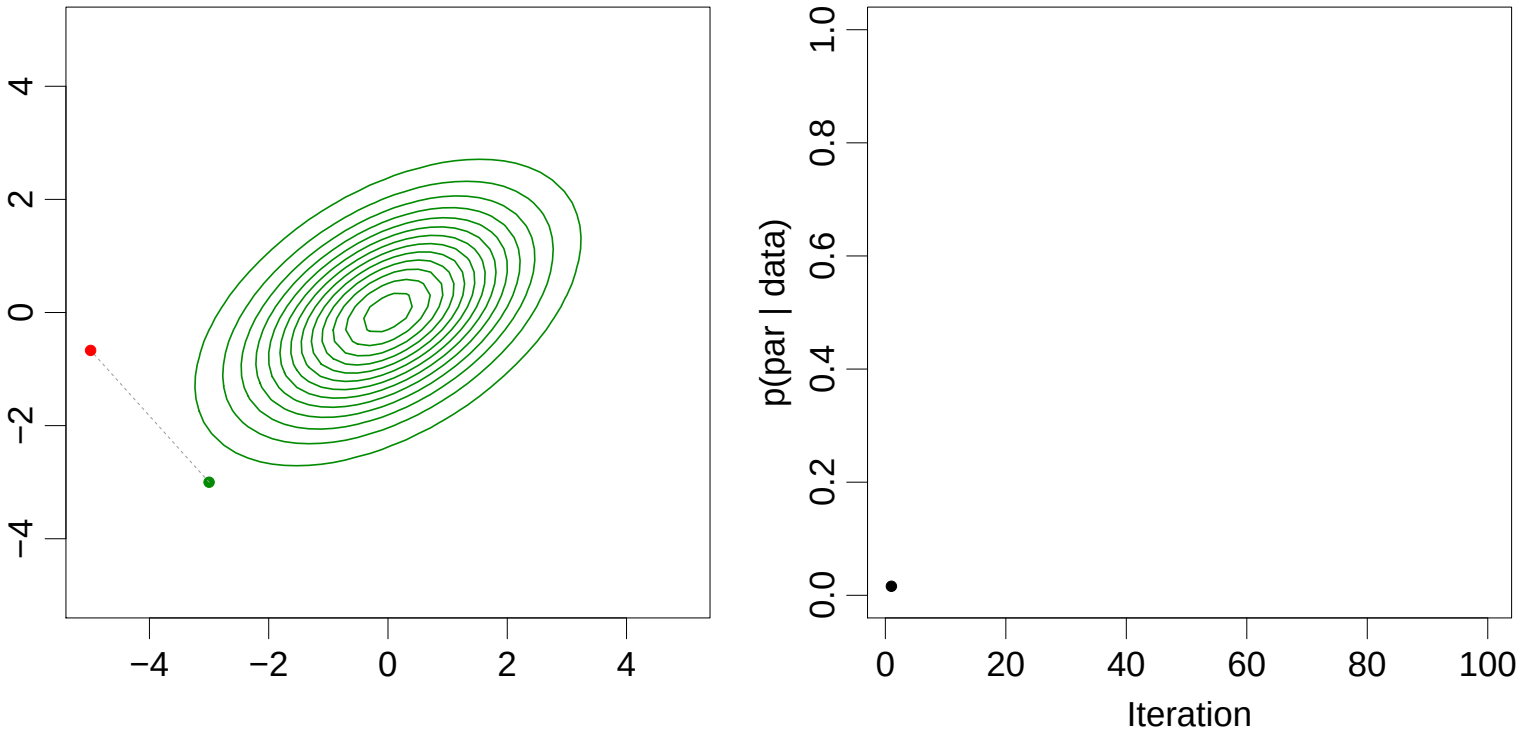
$$\theta' = \theta + \frac{\alpha_t}{2}\widetilde{\nabla}_\theta \log[p(\mathbf{y}|X,\theta)p(\theta)] \qquad \alpha_t \to 0$$



Robbins and Monro, *AoMS*, 1951

Stochastic Gradient Langevin Dynamics (SGLD) algorithm

$$\theta' = \theta + \frac{\alpha_t}{2}\widetilde{\nabla}_\theta \log[p(y|X,\theta)p(\theta)] + \eta_t \qquad \eta_t \sim \mathcal{N}(0,\alpha_t)$$



Welling and Teh, *ICML*, 2011 - Filippone and Engler, *ICML*, 2015

Stochastic Gradients in GP regression

- Marginal likelihood

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

- Derivatives wrt $\boldsymbol{\theta}$

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

Filippone and Engler, *ICML*, 2015 - Cutajar, Osborne, Cunningham, Filippone, *ICML*, 2016

Stochastic Gradients in GP regression

- Stochastic estimate of the trace

$$\mathrm{Tr}\left(K^{-1}\frac{\partial K}{\partial \boldsymbol{\theta}_i}\right) = \mathrm{Tr}\left(K^{-1}\frac{\partial K}{\partial \boldsymbol{\theta}_i}\mathrm{E}[\mathbf{r}\mathbf{r}^\top]\right) = \mathrm{E}\left[\mathbf{r}^\top K^{-1}\frac{\partial K}{\partial \boldsymbol{\theta}_i}\mathbf{r}\right]$$

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

Stochastic Gradients in GP regression

- Stochastic estimate of the trace

$$\mathrm{Tr}\left(K^{-1}\frac{\partial K}{\partial \boldsymbol{\theta}_i}\right) = \mathrm{Tr}\left(K^{-1}\frac{\partial K}{\partial \boldsymbol{\theta}_i}\mathrm{E}[\mathbf{r}\mathbf{r}^\top]\right) = \mathrm{E}\left[\mathbf{r}^\top K^{-1}\frac{\partial K}{\partial \boldsymbol{\theta}_i}\mathbf{r}\right]$$

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

- Stochastic gradient

$$-\frac{1}{2N_\mathbf{r}}\sum_{i=1}^{N_\mathbf{r}} \mathbf{r}^{(i)\top} K^{-1}\frac{\partial K}{\partial \boldsymbol{\theta}_i}\mathbf{r}^{(i)} + \frac{1}{2}\mathbf{y}^\top K^{-1}\frac{\partial K}{\partial \boldsymbol{\theta}_i}K^{-1}\mathbf{y}$$

- Linear systems only!

Filippone and Engler, *ICML*, 2015 - Cutajar, Osborne, Cunningham, Filippone, *ICML*, 2016

Solving linear systems

- Linear systems:

$$K\mathbf{s} = \mathbf{b}$$

- Can be solved using conjugate gradient:

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

- Iterative update $\mathbf{s} = \mathbf{s}_0 + \boldsymbol{\delta}_1 + \ldots + \boldsymbol{\delta}_T$
- Requires only $K\mathbf{v}$ multiplications! $O(n^2)$ time
- No need to store K! $O(n)$ space

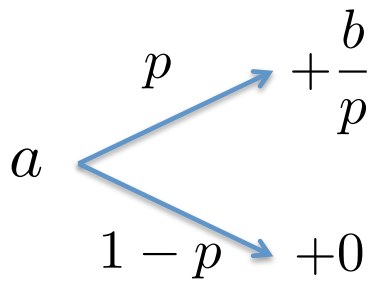
Filippone and Engler, *ICML*, 2015 - Cutajar, Osborne, Cunningham, Filippone, *ICML*, 2016

The Unbiased LInear System SolvEr - ULISSE

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

The Unbiased LInear System SolvEr - ULISSE

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

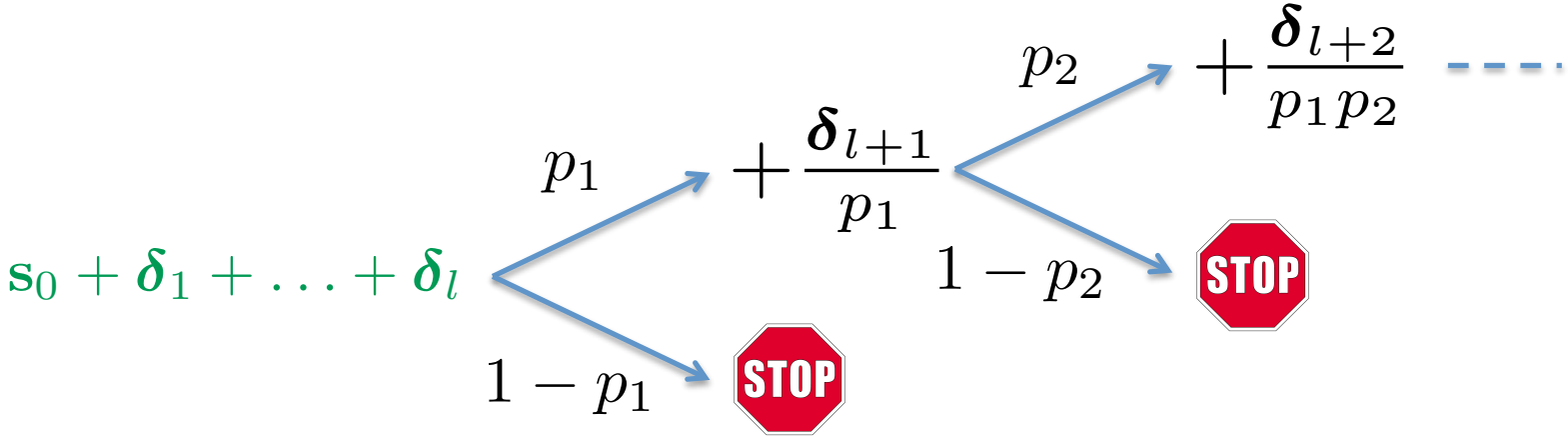


The Unbiased LInear System SolvEr - ULISSE

- Full CG solution:

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

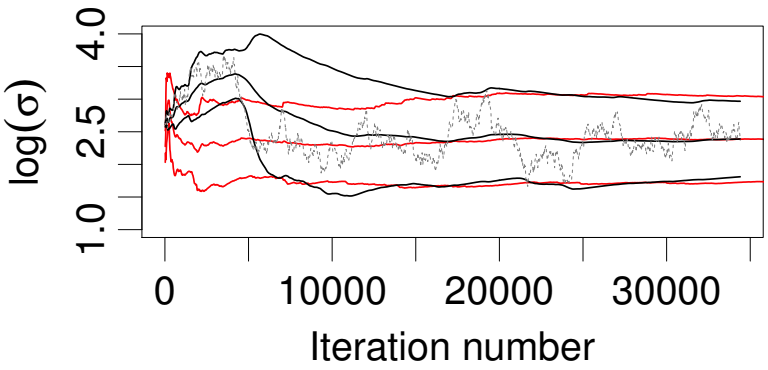
- ULISSE:

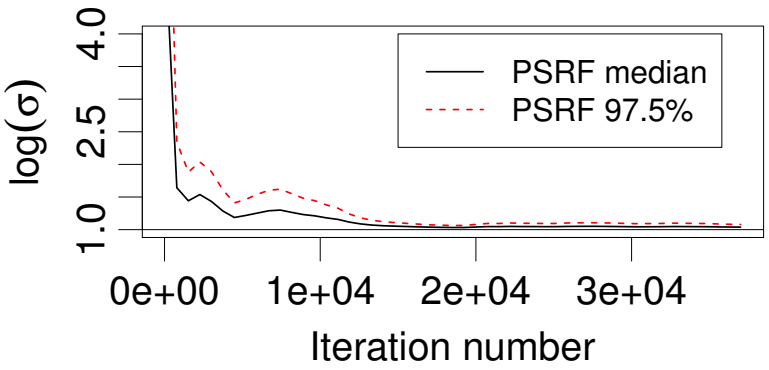


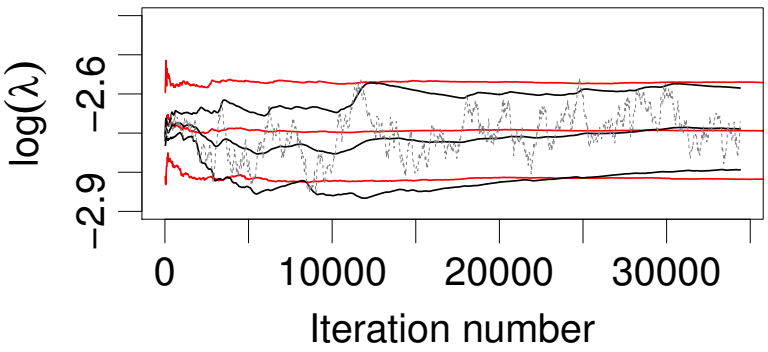
- Final solution is an unbiased estimate of $\mathbf{s}$!

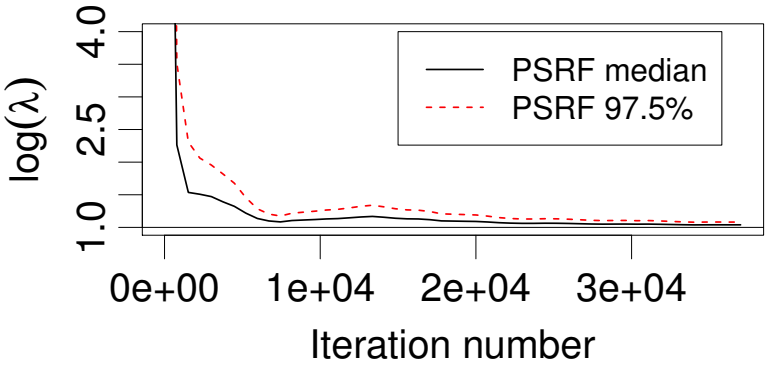
Filippone and Engler, *ICML*, 2015

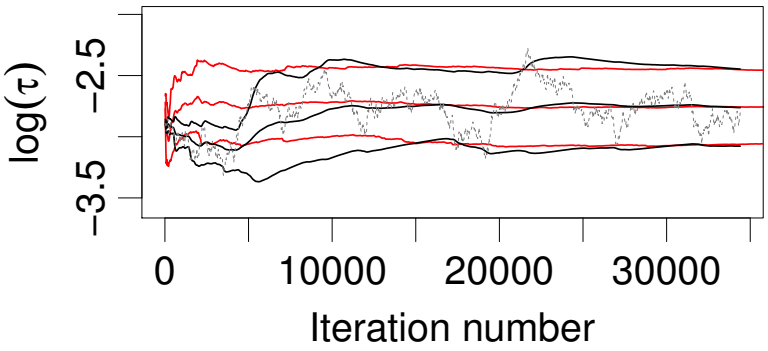
Comparison with MCMC - Concrete dataset - $n \approx 1K$

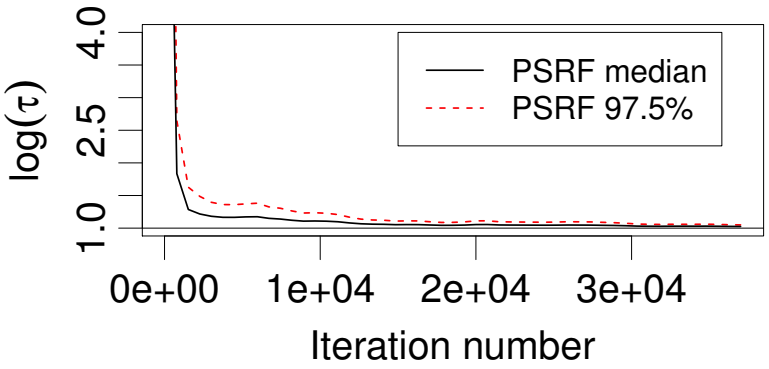




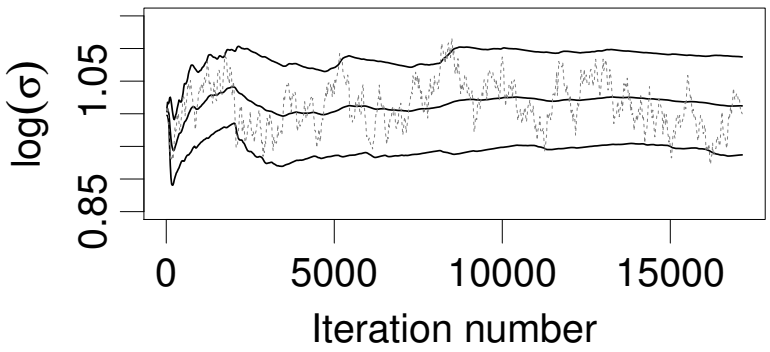


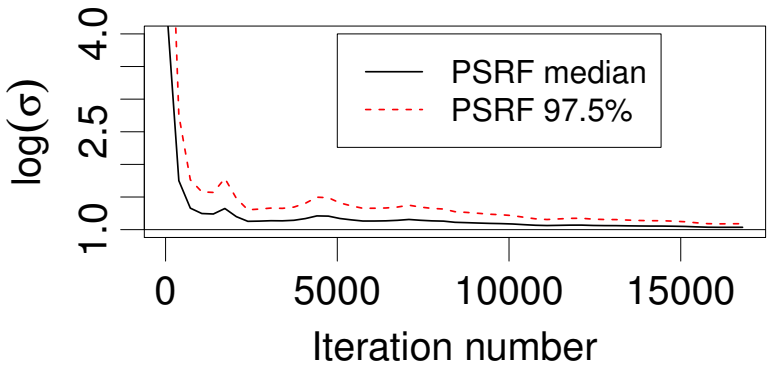


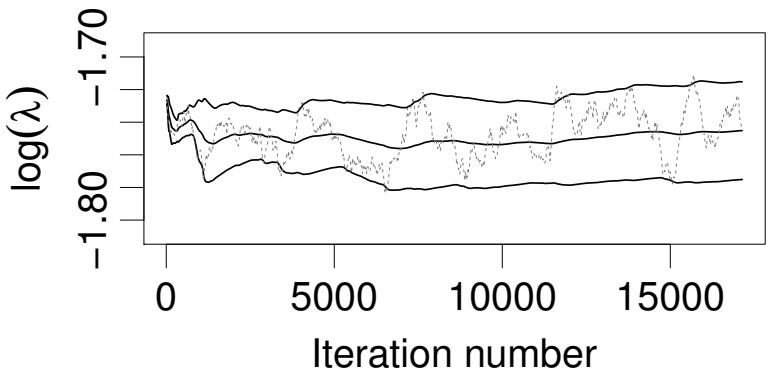


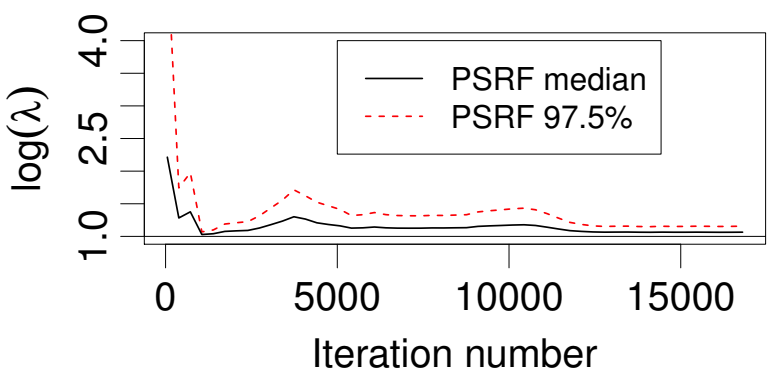


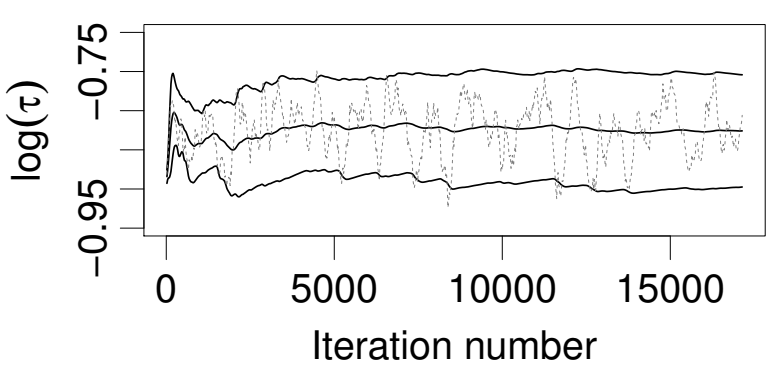
Larger n - Census dataset - $n \approx 23K$

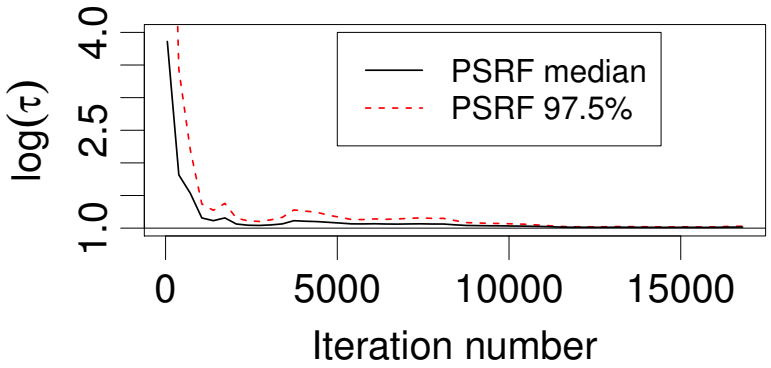






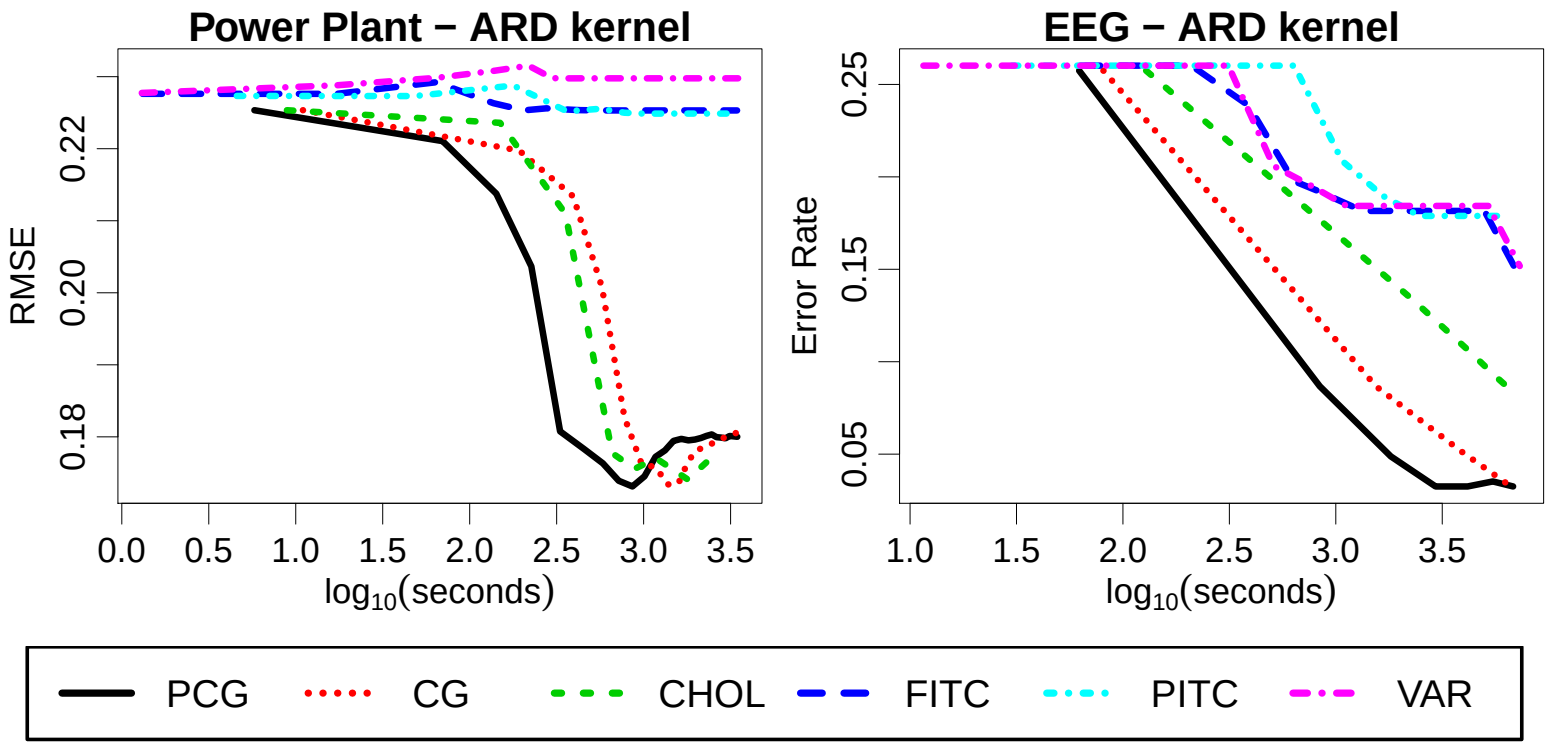






Preconditioning Kernel Matrices

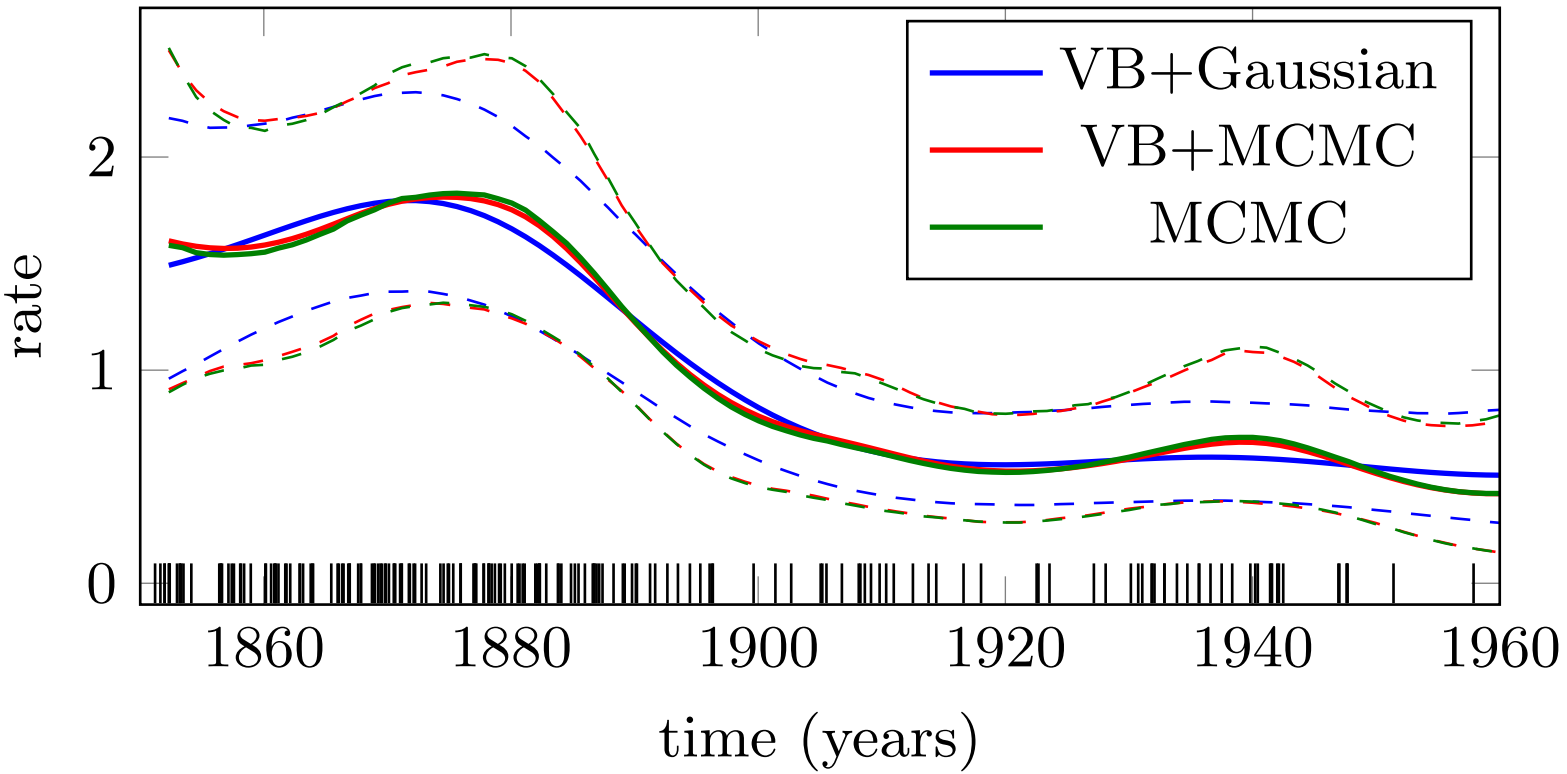
- Stochastic Gradient Optimization



Cutajar, Osborne, Cunningham, Filippone, *ICML*, 2016

Marrying Sparse Variational and MCMC learning for GPs

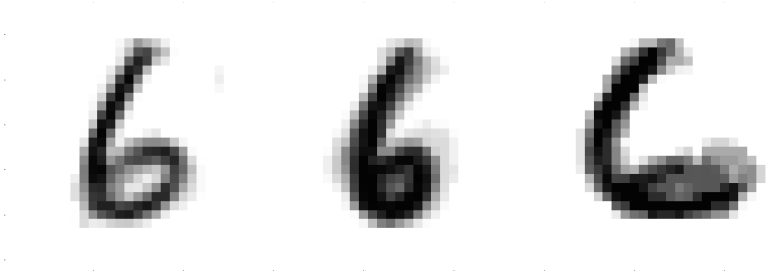
- Coal mining disaster data
- Analysis with 30 inducing points



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

Marrying Sparse Variational and MCMC learning for GPs

- MNIST data
- Three inducing points before optimizing their position



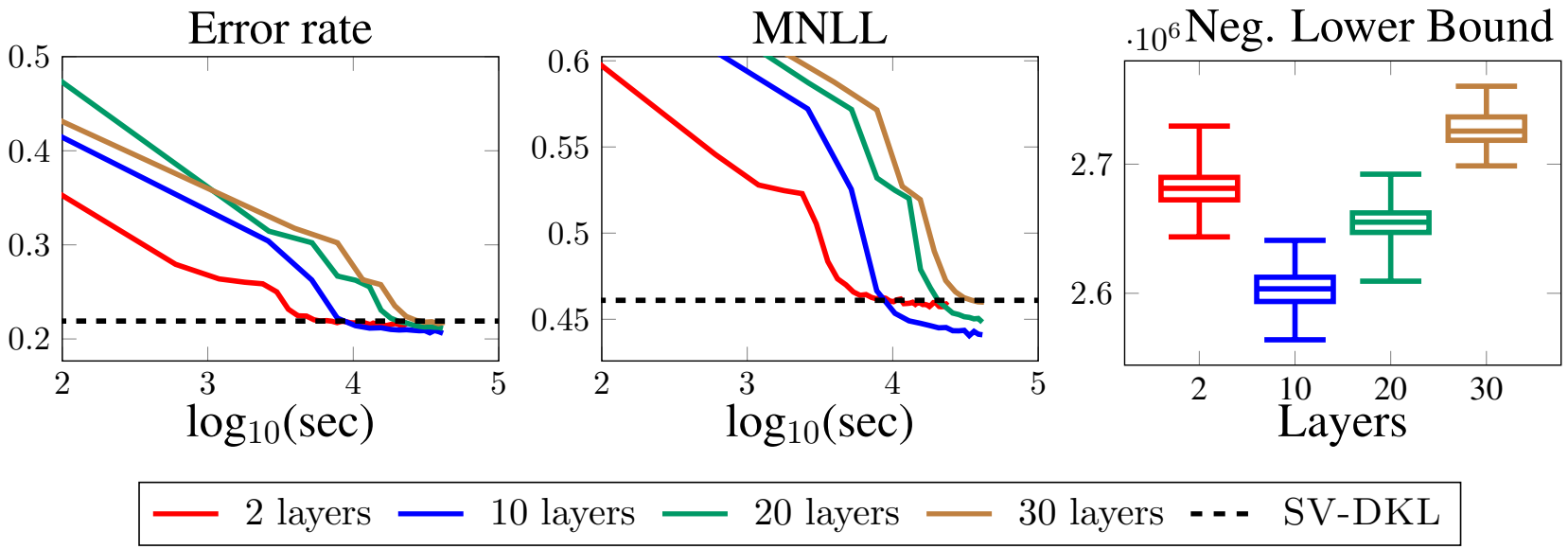
- . . . and after

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

Deep Gaussian Processes

- Layer-wise random feature expansion
- Stochastic variational inference

Airline dataset
$(n = 5\text{M}+,\ d = 8)$



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** as a way to improve practicality and scalability without affecting performance

Thank you!