Well-known shortcomings, advantages and computational challenges in Bayesian modelling: a few case stories

Ole Winther

The Bioinformatics Center
University of Copenhagen

Informatics and Mathematical Modelling
Technical University of Denmark
DK-2800 Lyngby, Denmark
ole.winther@gmail.com
owi@imm.dtu.dk

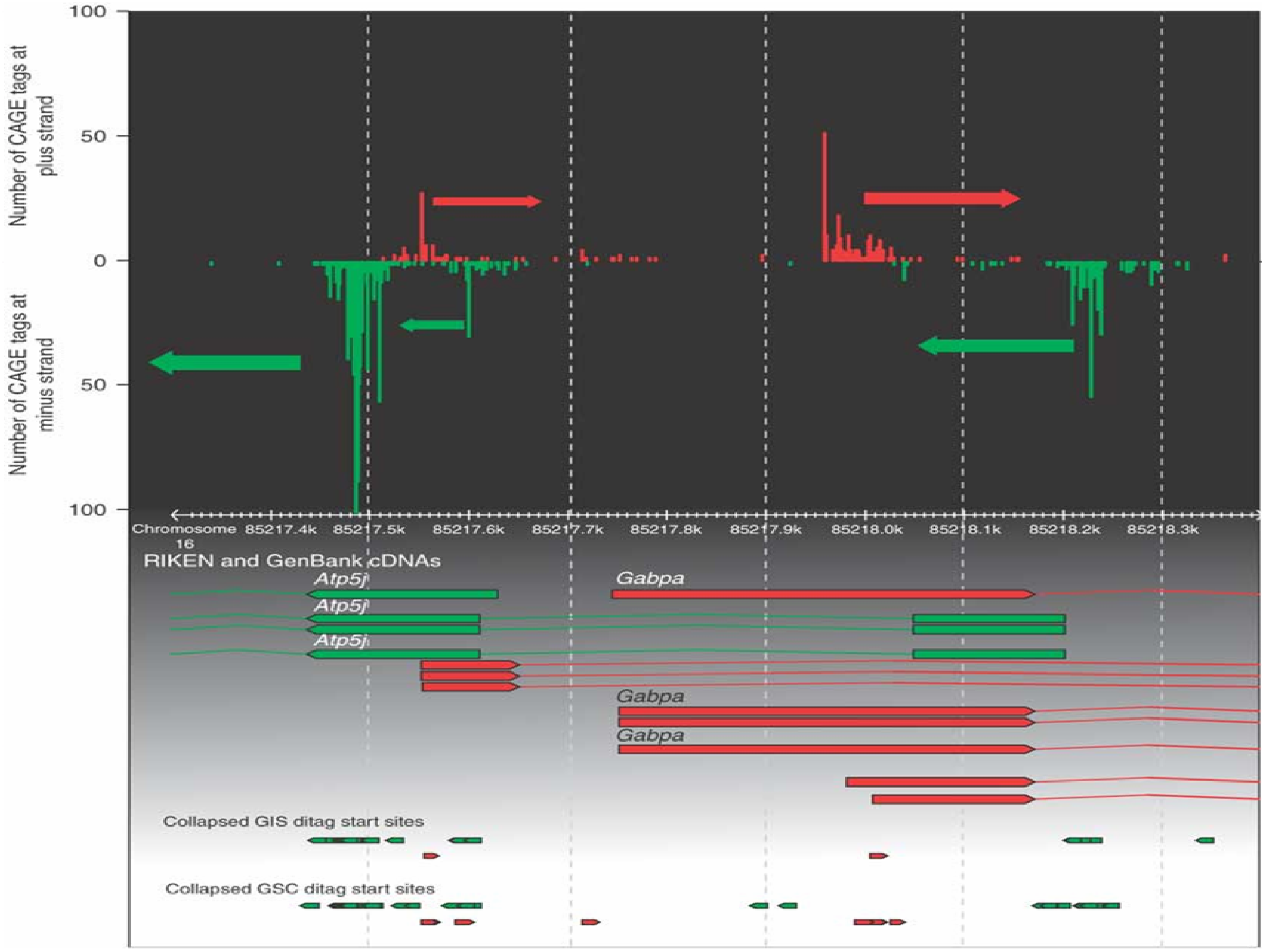
Overview

1. How many species? predicting sequence tags.
   - Non-parametric Bayes
   - Averaging beats maximum likelihood
   - The model is always wrong (and Bayes can't tell)

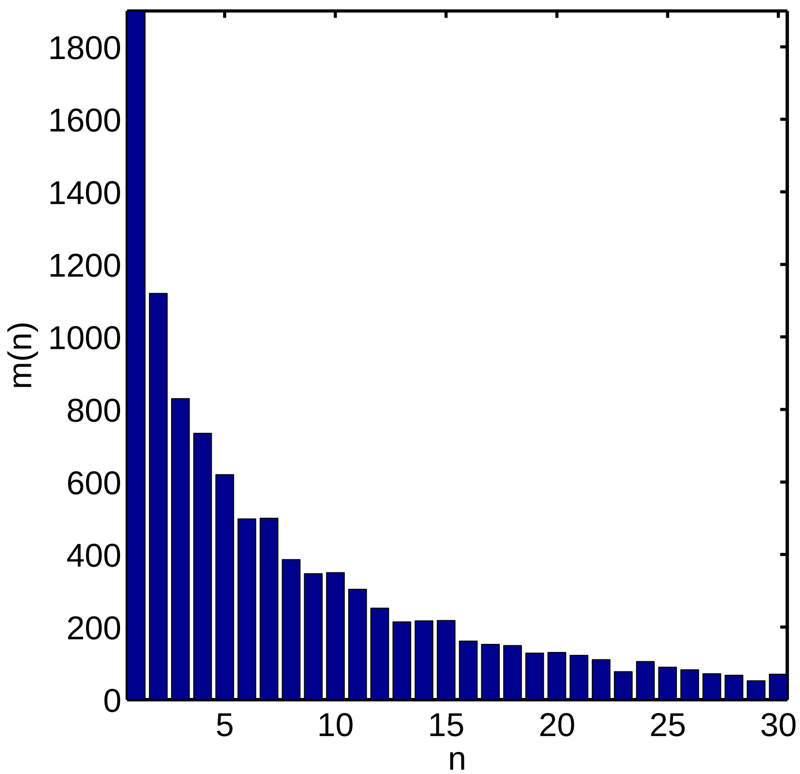
2. Computing the marginal likelihood with MCMC
   - Motivation: computing corrections to EP/C
   - The trouble with Gibbs sampling
   - Gaussian process classification

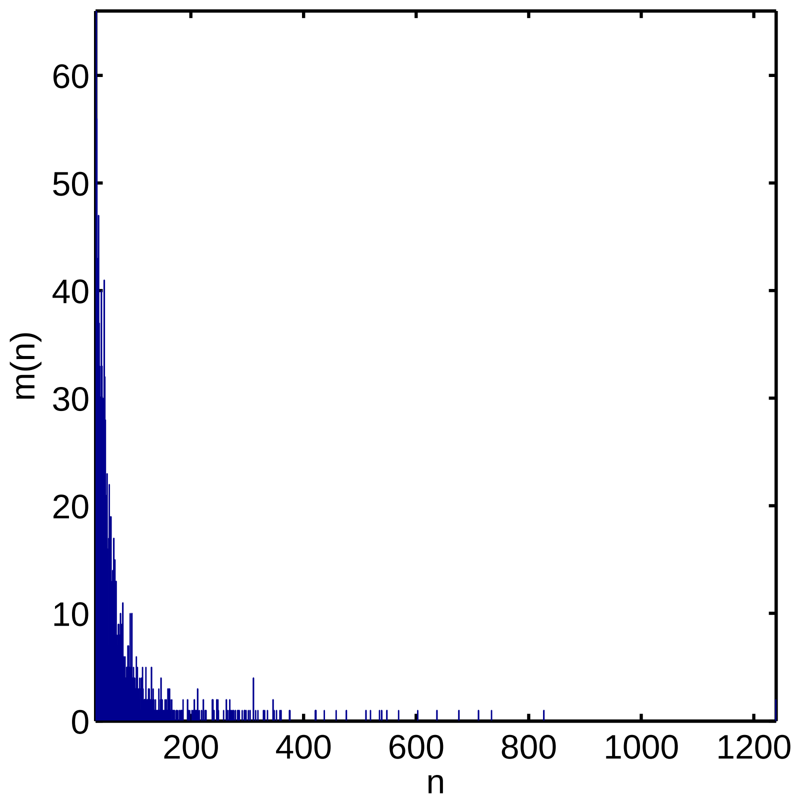
How many species?

DNA sequence tags - CAGE

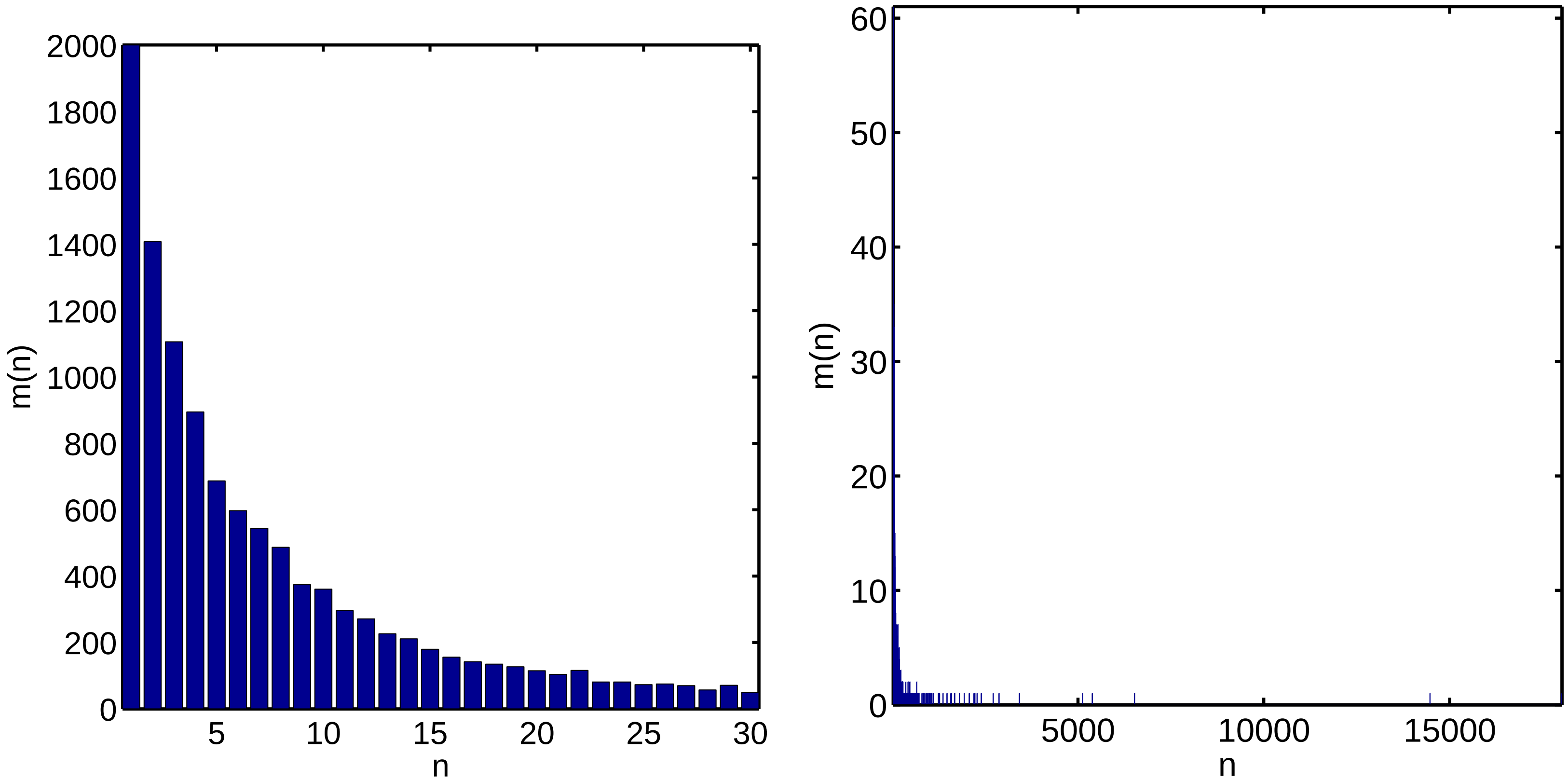


Look at the data - cerebellum





Look at the data - embryo



Chinese restaurant process - Yor-Pitman sampling formula

Observing new species given counts $\mathbf{n} = n_1, \ldots, n_k$ in k bins:

$$p(n_1, \ldots, n_k, 1|\mathbf{n}, \sigma, \theta) = \frac{\theta + k\sigma}{n + \theta} \quad \text{with} \quad \sum_{i=1}^{k} n_i = n$$

Re-observing j:

$$p(n_1, \ldots, n_{j-1}, n_j + 1, n_{j+1}, \ldots, n_k|\mathbf{n}, \sigma, \theta) = \frac{n_j - \sigma}{n + \theta}$$

Exchangeability – invariant to re-ordering

$$E, E, M, T, T: \quad p_1 = \frac{\theta}{\theta} \frac{1-\sigma}{1+\theta} \frac{\theta+\sigma}{2+\theta} \frac{\theta+2\sigma}{3+\theta} \frac{1-\sigma}{4+\theta}$$

$$M, T, E, T, E: \quad p_2 = \frac{\theta}{\theta} \frac{\theta+\sigma}{1+\theta} \frac{\theta+2\sigma}{2+\theta} \frac{1-\sigma}{3+\theta} \frac{1-\sigma}{4+\theta} = \ldots = p1$$

Inference and prediction

Likelihood function, e.g. E, E, M, T, T

$$\begin{aligned} p(\mathbf{n}|\sigma,\theta) &= \frac{\theta}{\theta}\,\frac{1-\sigma}{1+\theta}\,\frac{\theta+\sigma}{2+\theta}\,\frac{\theta+2\sigma}{3+\theta}\,\frac{1-\sigma}{4+\theta} \\ &= \frac{1}{\prod_{i=1}^{n-1}(i+\theta)} \prod_{j=1}^{k-1}(\theta+j\sigma) \prod_{i'=1}^{k} \prod_{j'=1}^{n_{i'}-1} (j'-\sigma) \end{aligned}$$

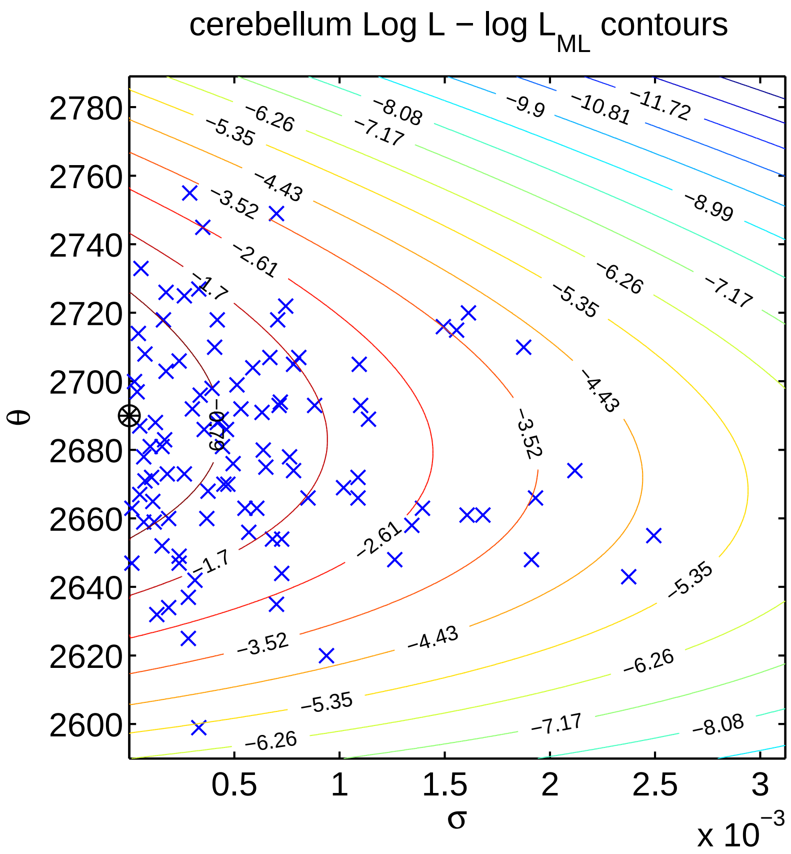
Flat prior distribution for $\sigma \in [0,1]$ and θ pseudo-count parameter.

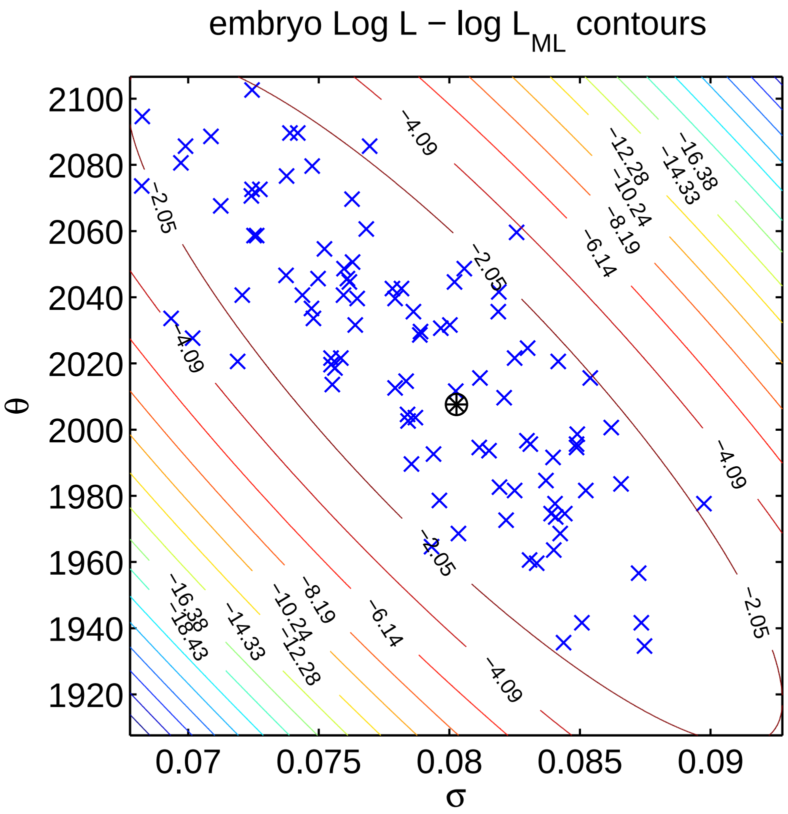
Predictions for new count $\mathbf{m}$:

$$p(\mathbf{m}|\mathbf{n}) = \int p(\mathbf{m}|\mathbf{n},\sigma,\theta)\, p(\sigma,\theta)\, d\sigma d\theta$$

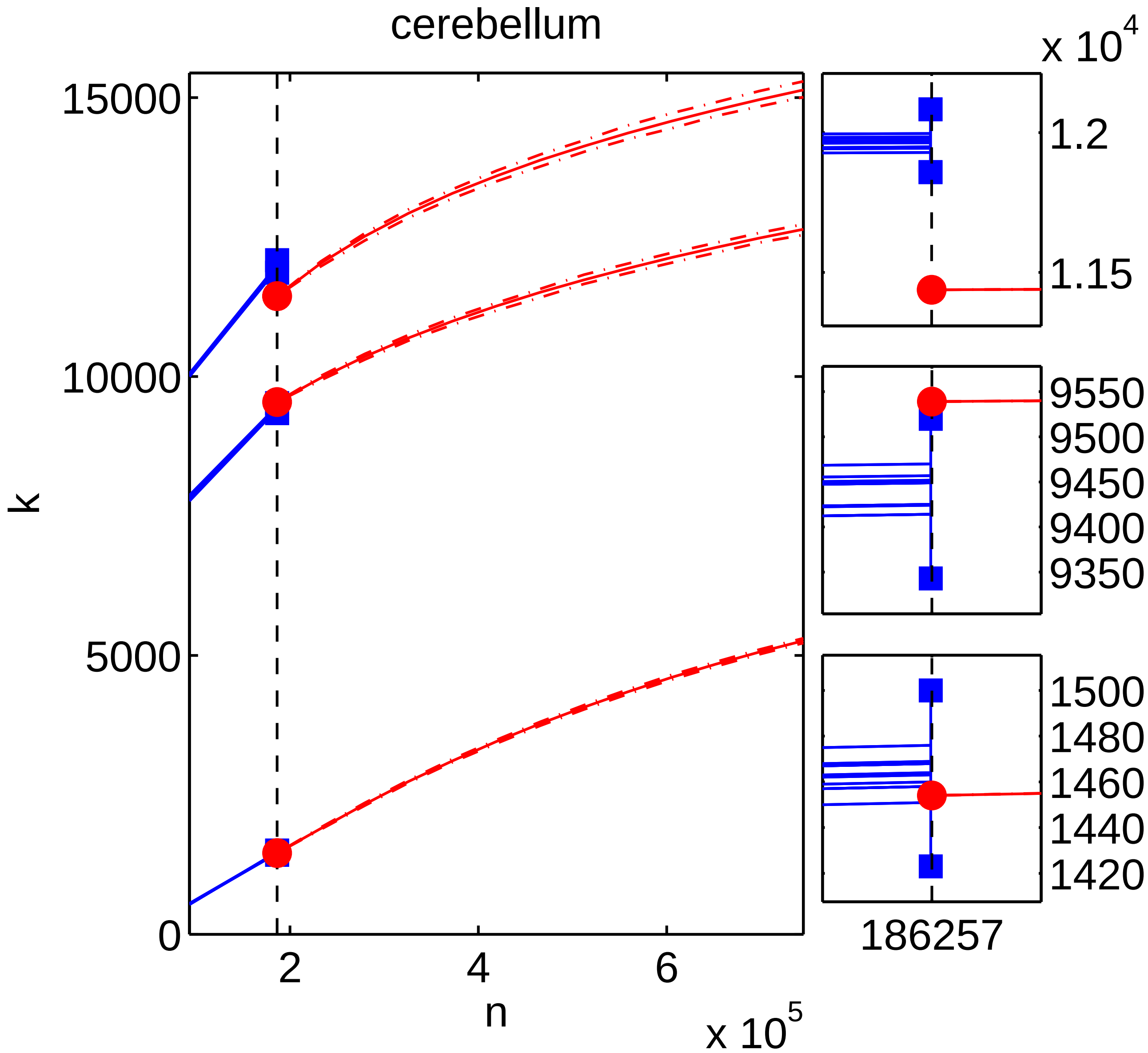
with Gibbs sampling (σ, θ) and Yor-Pitman sampling for $\mathbf{m}$.

Averaging versus max. likelihood

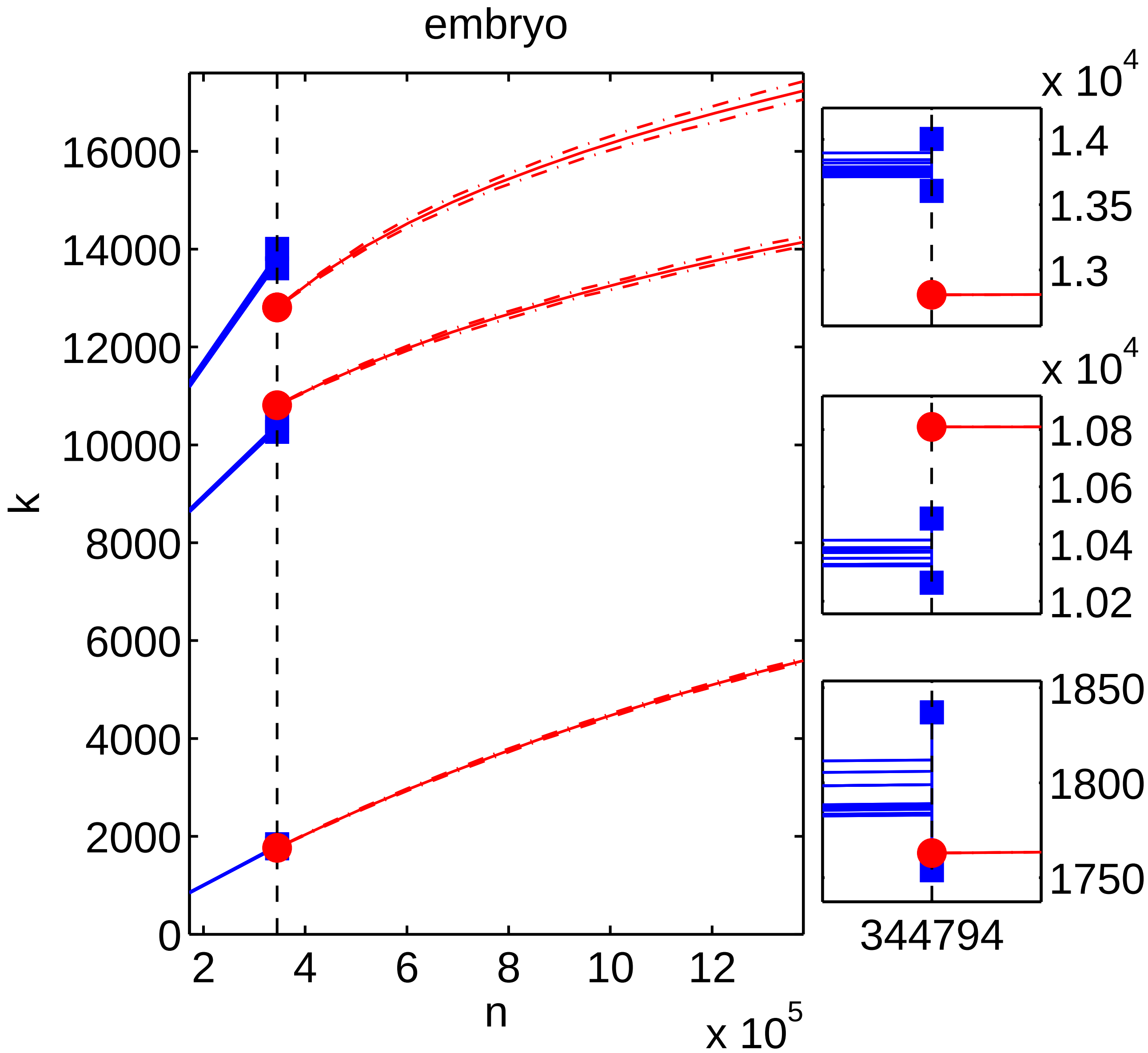




Notice anything funny?



Notice anything funny? Example 2



(Well-known) take home messages

- Parameter averaging works!

- The model is always wrong! (as revealed with sufficient data)

- Only one model considered!

- What happened to being Bayesian about model selection?



Calculating the marginal likelihood

The marginal likelihood:

$$p(\mathcal{D}|\mathcal{H}) = \int p(\mathcal{D}|\mathbf{f}, \mathcal{H})\, p(\mathbf{f}|\mathcal{H})\, d\mathbf{f}$$

Approximate inference needed!

Monte Carlo: slow mixing, non-trivial to get marginal likelihood estimates

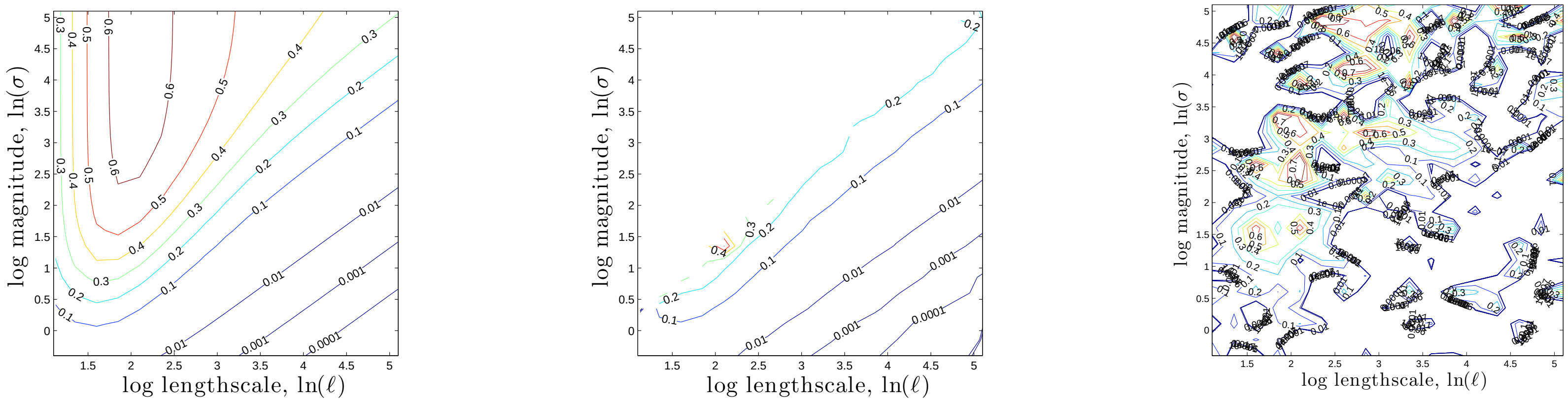
Expectation propagation+, variational Bayes, loopy BP+: sometimes not precise, approximation errors not controllable, not applicable to all models.

Motivation: validating EP corrections

Kuss-Rasmussen (JMLR 2006) $N = 767$ 3-vs-5 GP USPS digit classification with

$$K(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2\ell^2}\right)$$

I: $\log R = \log Z_{\mathrm{EPc}} - \log Z_{\mathrm{EP}}$ and II+III: $\log Z_{\mathrm{MCMC}} - \log Z_{\mathrm{EP}}$.



Thanks to Malte Kuss for making III available.

Marginal likelihood from importance sampling

Importance sampling

$$p(\mathcal{D}|\mathcal{H}) = \int \frac{p(\mathcal{D}|\mathbf{f},\mathcal{H})\,p(\mathbf{f}|\mathcal{H})}{q(\mathbf{f})}\,q(\mathbf{f})\,d\mathbf{f}$$

Draw samples $\mathbf{f}_1, \ldots, f_R$ from $q(\mathbf{f})$ and set

$$p(\mathcal{D}|\mathcal{H}) \approx \frac{1}{R}\sum_{r=1}^{R} \frac{p(\mathcal{D}|\mathbf{f}_r,\mathcal{H})\,p(\mathbf{f}_r|\mathcal{H})}{q(\mathbf{f}_r)}$$

This will usually not work because ratio varies too much.

Marginal likelihood from thermodynamic integration

Variants: parallel tempering, simulated tempering and annealed importance sampling

$$\begin{aligned} h(\mathbf{f}) &= p(\mathcal{D}|\mathbf{f},\mathcal{H})\, p(\mathbf{f}|\mathcal{H}) \\ p(\mathbf{f}|\beta) &= \frac{1}{Z(\beta)} h^{\beta}(\mathbf{f})\, q^{1-\beta}(\mathbf{f}) \\ \log Z(\beta_2) - \log Z(\beta_1) &= \int_{\beta_1}^{\beta_2} \frac{d \log Z(\beta)}{d\beta}\, d\beta \\ &= \int_{\beta_1}^{\beta_2} \int \log \frac{h(\mathbf{f})}{q(\mathbf{f})}\, p(\mathbf{f}|\beta)\, d\mathbf{f}\, d\beta \end{aligned}$$

Run N_β chains and interpolate

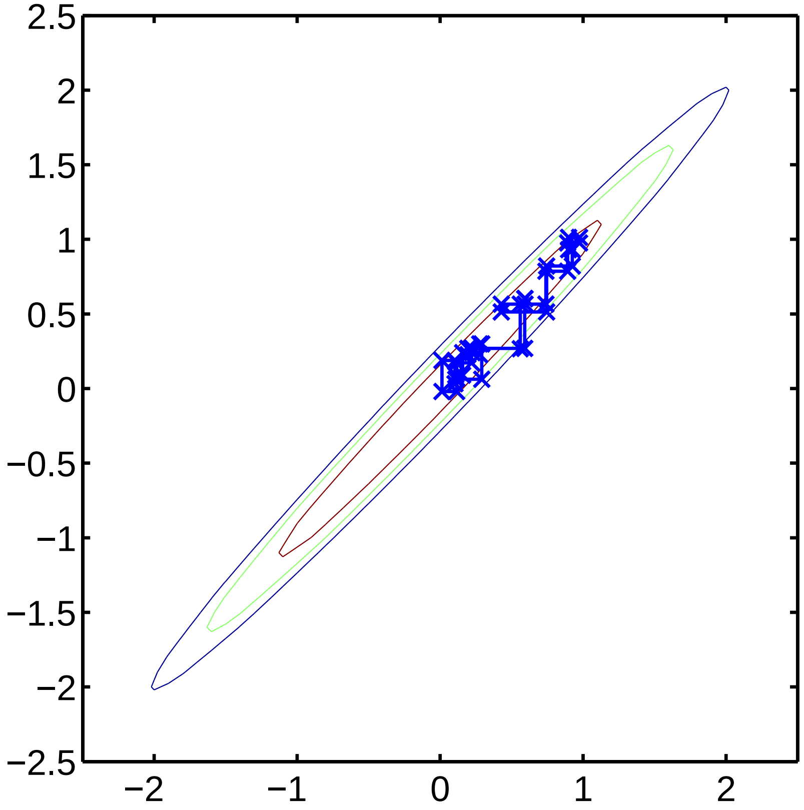
$$\log Z(\beta_2) - \log Z(\beta_1) \approx \frac{\Delta\beta}{R} \sum_{b=1}^{N_\beta} \sum_{r=1}^{R} \log \frac{h(\mathbf{f}_{rb})}{q(\mathbf{f}_{rb})}$$

Other things that might work even better: multi-canonical.

The trouble with Gibbs sampling

Cycle over variables f_i, $i = 1, \ldots, N$ and sample conditionals

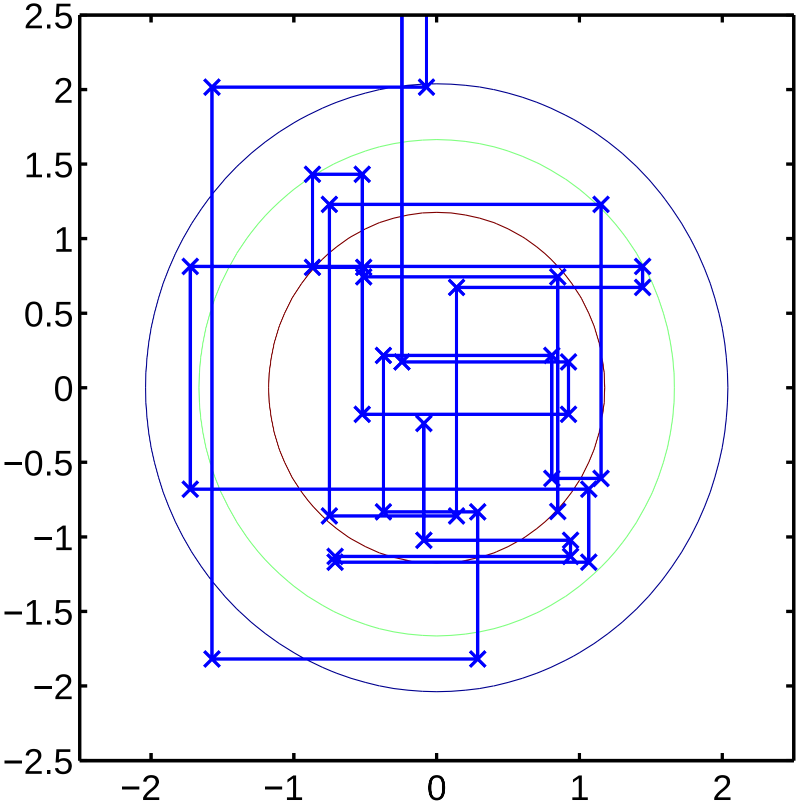
$$p(f_i|\mathbf{f}_{\setminus i}) = \frac{p(\mathbf{f})}{p(\mathbf{f}_{\setminus i})} \propto p(f)$$

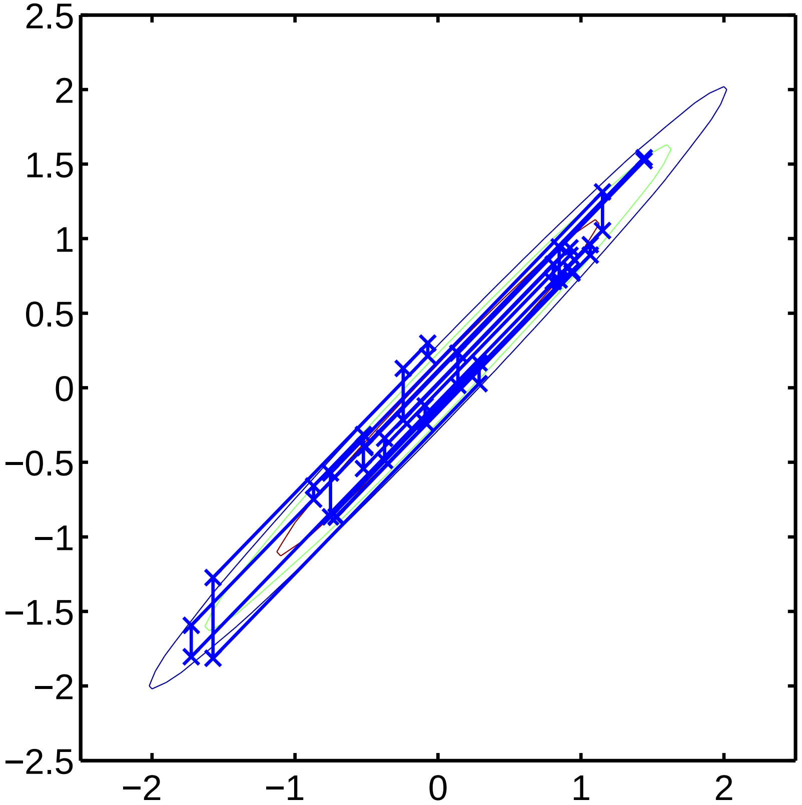


A trivial cure for $\mathcal{N}(\mathbf{f}|0, \mathbf{C})$

Gibbs sample z_i, $i = 1, \ldots, N$ with $\mathcal{N}(\mathbf{z}|0, \mathbf{I})$

$$\mathbf{f} = \mathbf{L}\mathbf{z} \quad \text{with} \quad \mathbf{C} = \mathbf{L}\mathbf{L}^T$$





Gaussian process classification (GPC)

$$p(\mathbf{f}|\mathbf{y}, \mathbf{K}, \beta) = \frac{1}{Z(\beta)} \prod_n \phi(y_n f_n) \exp\left(-\frac{\beta}{2}\mathbf{f}^T \mathbf{K}^{-1} \mathbf{f}\right)$$

Noise-free formulation $\mathbf{f}_{\mathsf{nf}}$

$$\phi(yf) = \int \theta(y f_{\mathsf{nf}}) \mathcal{N}(f_{\mathsf{nf}}|\mathbf{f}, \mathbf{I})$$

Joint distribution

$$p(\mathbf{f}, \mathbf{f}_{\mathsf{nf}}, \mathbf{y}|\mathbf{K}, \beta) = p(\mathbf{y}|\mathbf{f}_{\mathsf{nf}}) p(\mathbf{f}_{\mathsf{nf}}|\mathbf{f}) p(\mathbf{f}|\mathbf{K}, \beta)$$

Marginalize out f

$$p(\mathbf{f}_{\mathsf{nf}}|\mathbf{y}, \mathbf{K}, \beta) \propto \prod_n \theta(y_n f_n) \mathcal{N}(f_{\mathsf{nf}}|0, \mathbf{I} + \mathbf{K}/\beta)$$

Samples of $\mathbf{f}$ can be recovered from $p(\mathbf{f}|\mathbf{f}_{\mathsf{nf}})$ (Gaussian)

Efficient sampler of truncated Gaussian needed!

MCMC for GPC - related work

G. Rodriguez-Yam, R. Davis, and L. Scharf: "Efficient Gibbs Sampling of Truncated Multivariate Normal with Application to Constrained Linear Regression" (preprint 2004)

R. Neal, U. Paquet: Sample joint $\mathbf{f}, \mathbf{f}_{\mathsf{nf}}$ and use Adler's over-relaxation on $\mathbf{f}$.

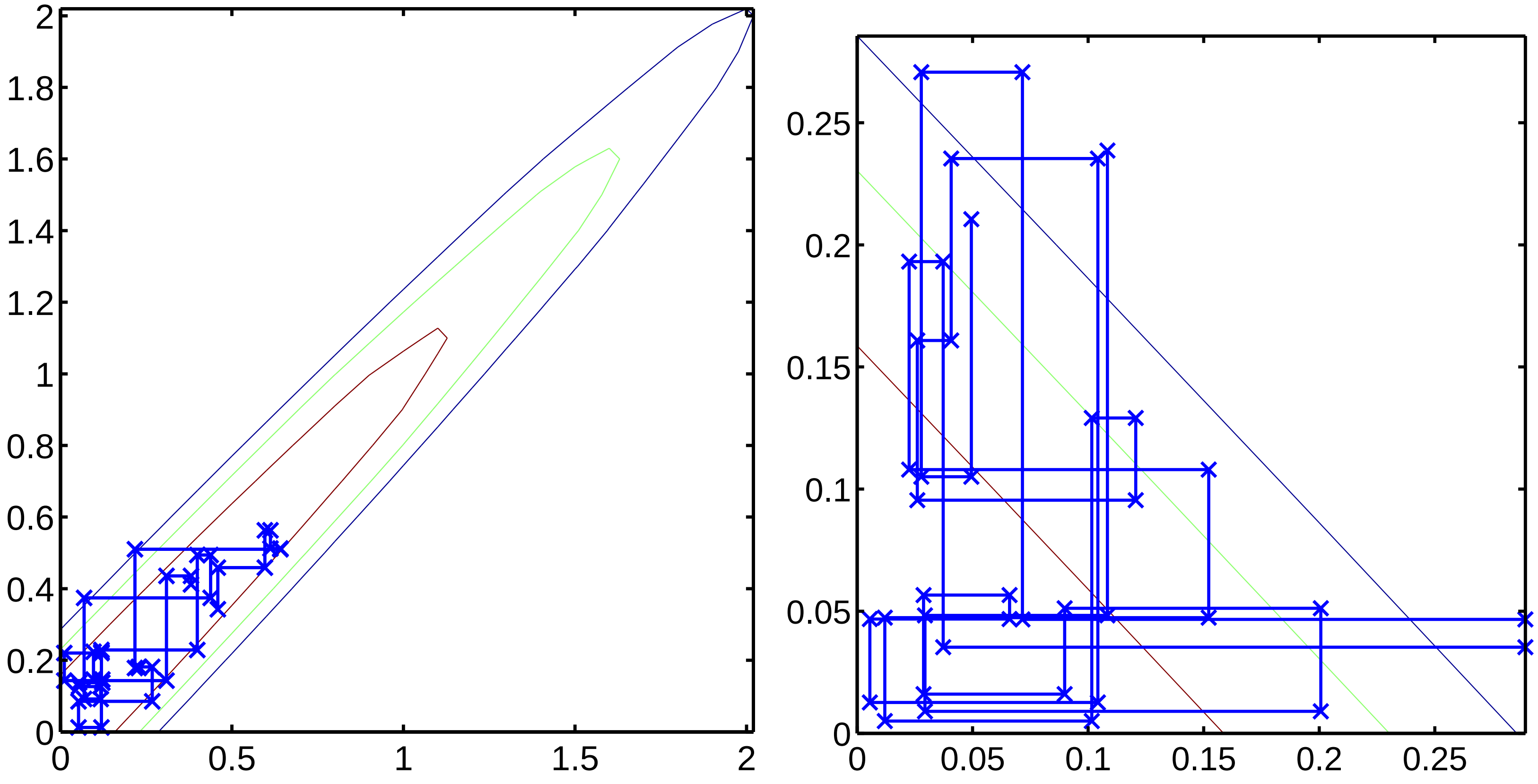
P. Rujan, R. Herbrich: Playing billiards in version space, the Bayes point machine.

M. Kuss+C. Rasmussen: Hybrid Monte Carlo in $\mathbf{z}$-space, $\mathbf{f} = \mathbf{L}\mathbf{z}$ and annealed importance sampling

Y. Qi+T. Minka: Hessian based Metropolis-Hastings - local Gaussian proposals.

Gibbs sampling - pos/neg covariance

Conditionals are truncated Gaussians! so sampling is easy.



Efficient Gibbs sampling I

Gibbs sample in whitened space: z_i, $i = 1, \ldots, N$ with $\mathcal{N}(\mathbf{z}|0, \mathbf{I})$

$$\mathbf{f} = \mathbf{L}\mathbf{z} \quad \text{with} \quad \mathbf{C} = \mathbf{L}\mathbf{L}^T$$

What happens to the constraints, say $\mathbf{f} \geq 0$

$$\mathbf{f} = \mathbf{L}\mathbf{z} \geq 0$$

Just a linear transformation so region in $\mathbf{z}$-space is convex and conditional (double) truncated Gaussian.

Determine limits of conditionals

jth conditional constraints $i = 1, \ldots, N$

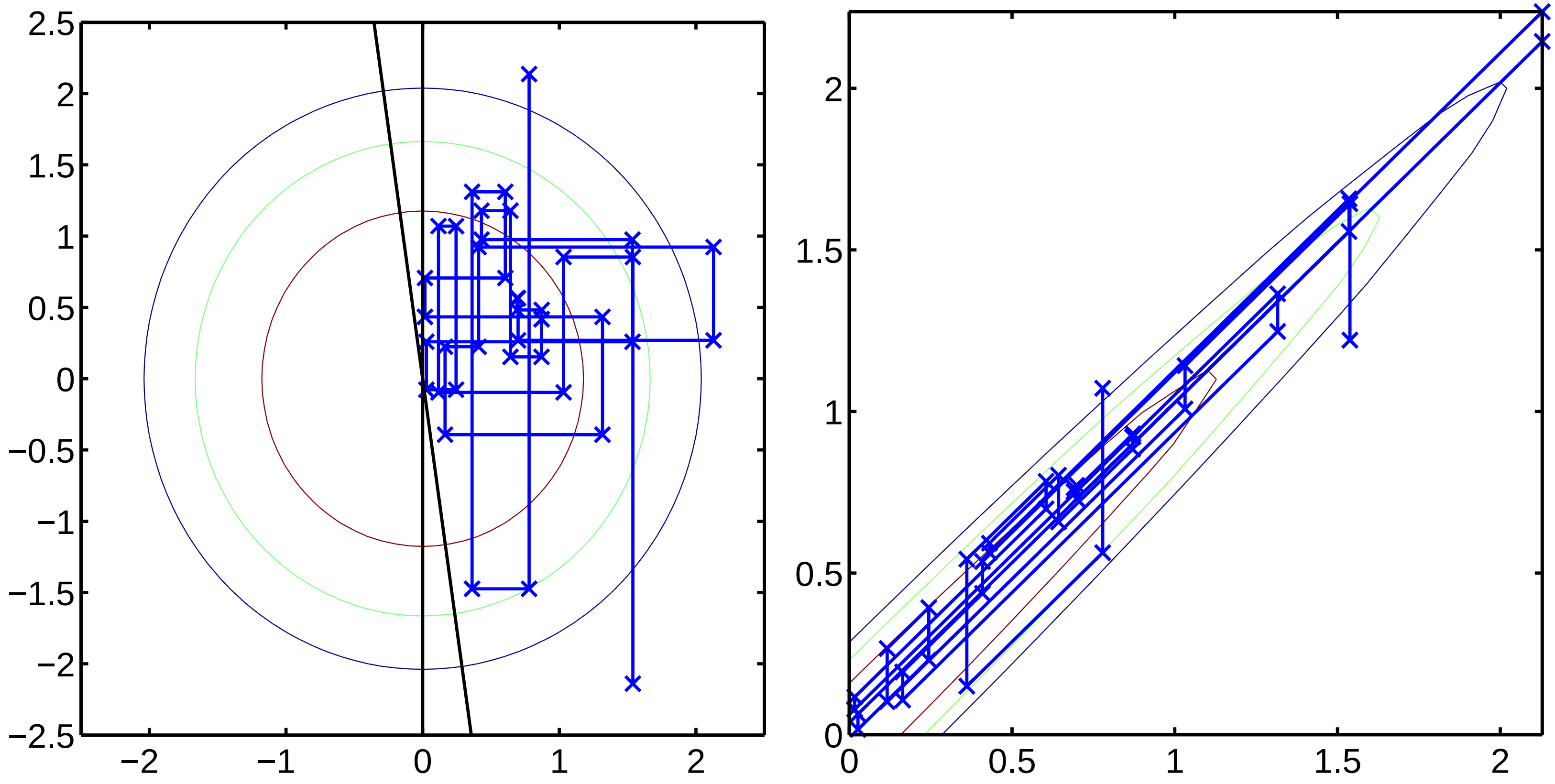
$$L_{ij} z_j \geq - \sum_{k \neq i} L_{ik} z_k$$

divide in sets

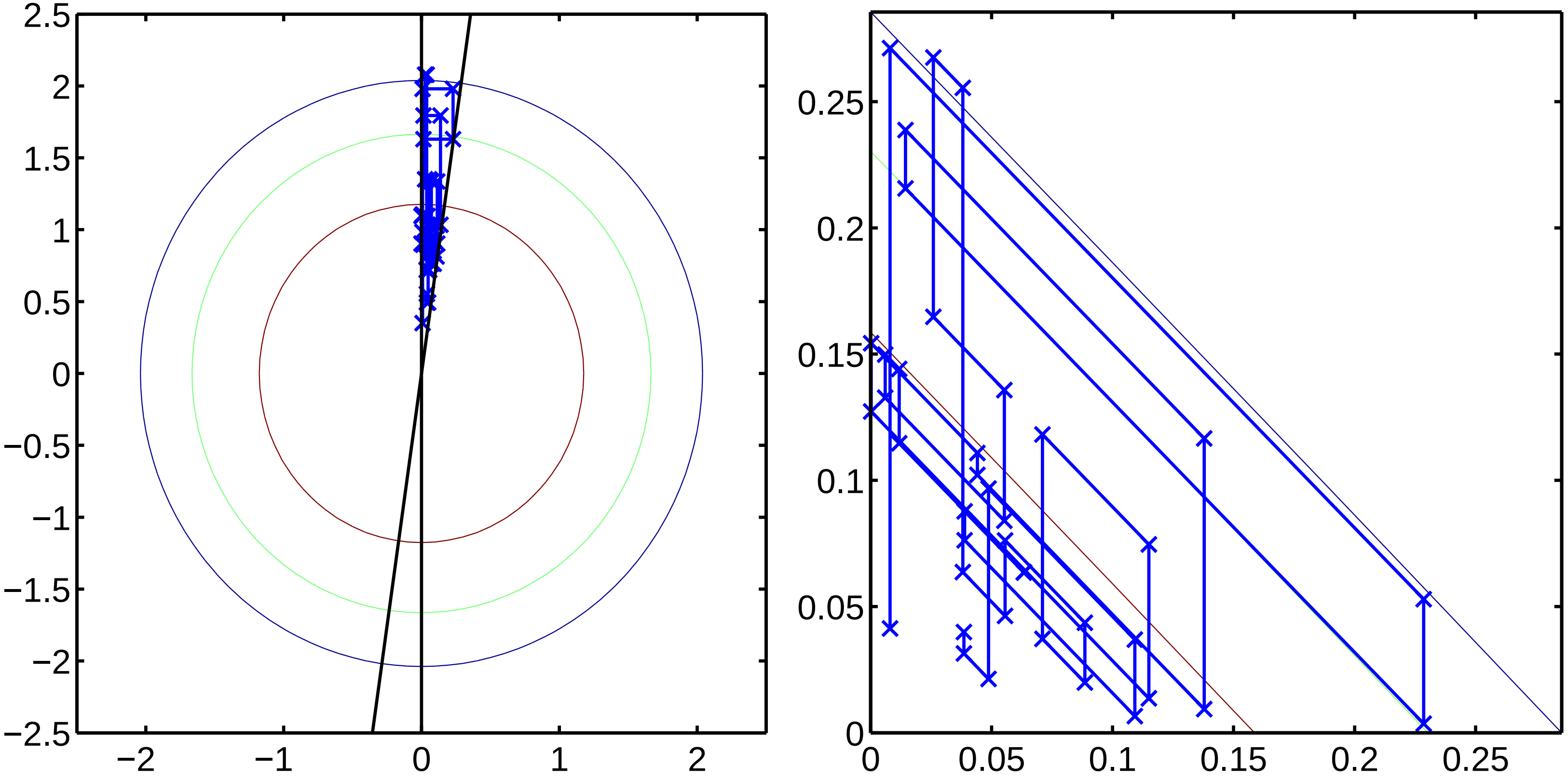
$$\begin{aligned} S_{+,j} &= \{i | L_{ij} > 0\} \\ S_{-,j} &= \{i | L_{ij} < 0\} \\ S_{0,j} &= \{i | L_{ij} = 0\} \end{aligned}$$

$$\begin{aligned} z_{j,\text{lower}} &= \max_{i \in S_{+,j}} \frac{-\sum_{k \neq i} L_{ik} z_k}{L_{ij}} \\ z_{j,\text{upper}} &= \min_{i \in S_{-,j}} \frac{-\sum_{k \neq i} L_{ik} z_k}{L_{ij}} \\ z_j &= \phi^{-1} \left\{ \phi(z_{j,\text{lower}}) + \text{rand} \left(\phi(z_{j,\text{upper}}) - \phi(z_{j,\text{lower}}) \right) \right\} \end{aligned}$$

Gibbs sampling positive covariance

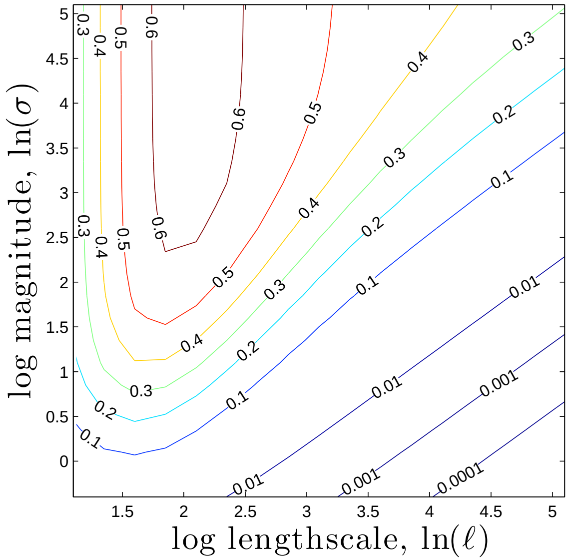


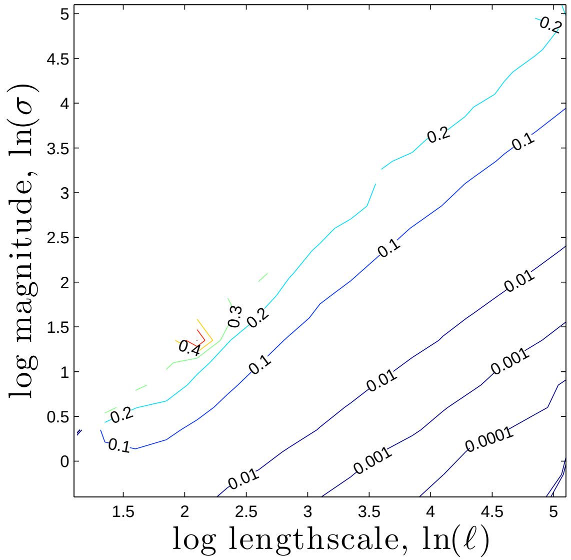
Gibbs sampling negative covariance



Kuss+Rasmussen set-up

EP, EP+corrections and MCMC are all very precise!





Details of the EP corrections will (hopefully) come to a conference near you soon!

Summary

- Averaging works!
- (X-)validation points to model miss-specification!
- How to find better (noise) models?
- Marginal likelihood from sampling
- The trouble with Gibbs sampling and a cure
- Is machine learning becoming (Bayesian) statistics with big data set?

Acknowledgments

Bioinformatics:
Albin Sandelin
Eivind Valen

Machine learning:
Ulrich Paquet
Manfred Opper

Students: (many shared)
Ricardo Henao (machine learning and bioinformatics)
Morten Hansen (communication)
Carsten Stahlhut (source location in EEG)
Jóan Petur Petersen (machine learning for ship propulsion)
Man-Hung Tang (bioinformatics)
Eivind Valen (bioinformatics)
Troels Marstrand (bioinformatics)