Leave Pima Indians alone

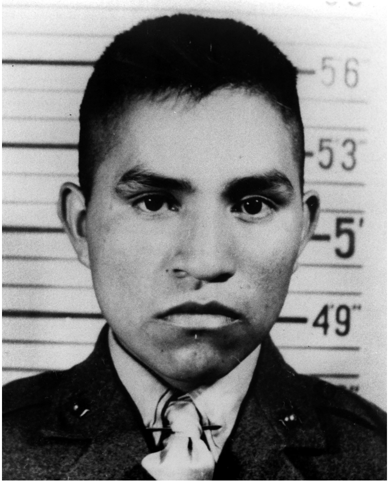
Nicolas Chopin
(joint work with James Ridgway)

ENSAE-CREST

Outline

Pima maze

Ira Hayes



Binary regression models

Models wih data $y_i \in \{-1, 1\}$, predictors $\boldsymbol{x}_i \in \mathbb{R}^p$, and likelihood

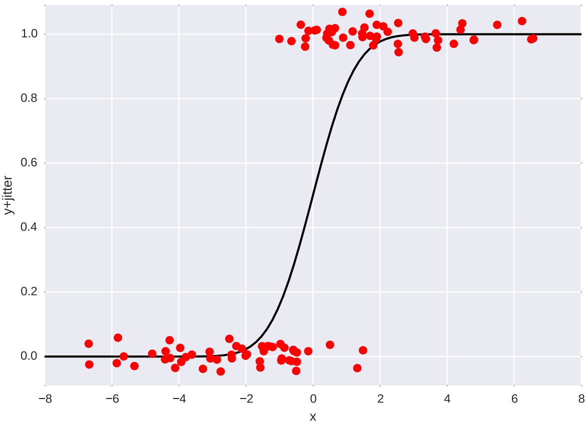
$$p(\mathcal{D}|\boldsymbol{\beta}) = \prod_{i=1}^{n_{\mathcal{D}}} F(y_i \boldsymbol{\beta}^T \boldsymbol{x}_i)$$

where $F : \mathbb{R} \to [0, 1]$ is a CDF.

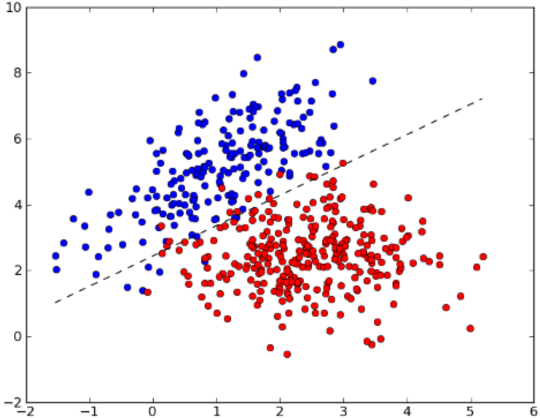
Common examples:

- $F = \Phi$ (probit),
- $F = L$ (logit), where $L(z) = 1/(1 + e^{-z})$.

When $p = 1$



Connection with classification



Properties

- Unless there is **complete separation** in the data, the log-likelihood is concave: MLE is uniquely defined.
- One nice way to deal with complete seperation is to add a proper prior, e.g. Gaussian or Cauchy. (Under Gaussian prior, log-post is concave.)
- Good practice is to standardise the predictors before eliciting the prior (Gelman et al, 2008).

Binary regression in Bayesian Computation papers

- a long chain of papers on Gibbs sampling for different variants of binary regression models (Albert & Chib, 1993; Holmes & Held, 2006; Fruwirth-Schnatter (2009); Gramacy and Polson, 2012; Polson et al, 2013)
- nearly any paper introducing any new **generic** way to compute a posterior includes a binary regression example:
  - SMC: C (2002), Del Moral et al (2006)
  - HMC and variants: Neal (2010), Shahbaba & Neal (2011), Girolami & Calderhead (2011)
  - NUTS: Hoffman and Gelman (2013)
  - nested sampling: C & Robert (2007)

Questions

1. Does it make sense to promote binary regression as a **benchmark** for Bayesian computation? (see similar practice in optimisation)
2. In practice, which method one should use???

Plan

1. review of fast approximation schemes:
    - Laplace (and variants)
    - EP
    - Variational Bayes? (see Consonni & Marin, 2007)
2. review of sampling-based approaches:
    - importance sampling
    - MCMC (Gibbs, RWHM)
    - HMC (and variants)
    - SMC
3. Discussion and comparison

Considered scenarios

- Model: probit and logit.
- prior: Gaussian and Cauchy (predictors are standardised).

Laplace

Based on a second order Taylor expansion of the log posterior:

$$\log p(\beta|\mathcal{D}) \approx \log p(\beta_{\mathrm{MAP}}|\mathcal{D}) - \frac{1}{2}(\beta - \beta_{\mathrm{MAP}})^T \boldsymbol{Q} (\beta - \beta_{\mathrm{MAP}})$$

where $\boldsymbol{Q}$ is minus the Hessian of $\log p(\beta|\mathcal{D})$ at $\beta = \beta_{\mathrm{MAP}}$.

Exponentiate to get a Gaussian approximation of the posterior. In practice, use Newton-Raphson to obtain β_{MAP} and $\boldsymbol{Q}$.

Very fast. May not converge if p is very large.

Impoved Laplace

For each marginal:

$$p(\beta_j|\mathcal{D}) \propto \frac{p(\boldsymbol{\beta})p(\mathcal{D}|\boldsymbol{\beta})}{p(\boldsymbol{\beta}_{-j}|\beta_j,\mathcal{D})}$$

Choose a fine grid of β_j values; for each β_j value, compute a Laplace approximation of $p(\boldsymbol{\beta}_{-j}|\beta_j,\mathcal{D})$.

Note: more expensive, connection with INLA.

EM-Laplace

For a Student prior, Gelman et al (2008) derive an approximate EM scheme based on

$$\beta_j|\sigma_j^2 \sim \mathrm{N}_1(0, \sigma_j^2), \quad \sigma_j^2 \sim \mathrm{Inv}-\mathrm{Gamma}(\nu/2, s_j\nu/2)$$

However, we will observe in our simulations that Laplace still works well for such a prior.

Expectation Propagation

Based on the following decomposition:

$$p(\boldsymbol{\beta}|\mathcal{D}) = \frac{1}{p(\mathcal{D})} \prod_{i=0}^{n_\mathcal{D}} l_i(\boldsymbol{\beta}), \quad l_i(\boldsymbol{\beta}) = F(y_i \boldsymbol{\beta}^T \boldsymbol{x}_i) \text{ for } i \geq 1,$$

$l_0 =$ prior, EP computes iteratively a parametric approx.:

$$q_{\mathrm{EP}}(\boldsymbol{\beta}) = \prod_{i=0}^{n_\mathcal{D}} \frac{1}{Z_i} q_i(\boldsymbol{\beta}).$$

Taking q_i to be an unnormalised Gaussian density

$$q_i(\boldsymbol{\beta}) = \exp\left\{-\frac{1}{2}\boldsymbol{\beta}^T \boldsymbol{Q}_i \boldsymbol{\beta} + \boldsymbol{\beta}^T \boldsymbol{r}_i\right\},$$

q_{EP} is a Gaussian with parameters $\boldsymbol{Q} = \sum_{i=0}^n \boldsymbol{Q}_i$, $\boldsymbol{r} = \sum_{i=0}^n \boldsymbol{r}_i$.

EP site update

Update each 'site' in turn: update q_i, while keeping q_j, $j \neq i$ fixed, by minimising the Kullback-Leibler divergence between

$$h(\boldsymbol{\beta}) \propto l_i(\boldsymbol{\beta}) \prod_{j \neq i} q_j(\boldsymbol{\beta})$$

and $q(\boldsymbol{\beta}) \propto \prod_j q_j$.

Thanks to nice properties of exponential families, this boils to match the moments of h and q.

In binary regression, these site updates lead to explicit expressions (probit) or one-dimensional integrals that are easy to approximate accurately (logit).

General remarks

- Since the approximation methods covered in the previous section are faster by orders of magnitude than sampling-based methods, we will assume that a Gaussian approximation $q(\boldsymbol{\beta})$ (from Laplace or EP) has been computed in a preliminary step.
- Complexity: Laplace is $O(n_{\mathcal{D}} + p^3)$, EP is $O(n_{\mathcal{D}} p^3)$.

Importance sampling

Proposal q set to some Gaussian approx of the posterior. Then to approximate $p(\mathcal{D})$, generate $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N \sim q$, compute

$$Z_N = \frac{1}{N} \sum_{n=1}^{N} w(\boldsymbol{\beta}_n), \quad w(\boldsymbol{\beta}) := \frac{p(\boldsymbol{\beta}) p(\mathcal{D}|\boldsymbol{\beta})}{q(\boldsymbol{\beta})}$$

and to approximate the posterior expectation of φ, compute

$$\varphi_N = \frac{\sum_{n=1}^{N} w(\boldsymbol{\beta}_n) \varphi(\boldsymbol{\beta}_n)}{\sum_{n=1}^{N} w(\boldsymbol{\beta}_n)}.$$

IS pros and cons

Pros:

- simple, generic
- embarassingly parallel
- approximates the marginal likelihood at no extra cost
- IID sampling: MC error is easy to assess
- can plug in QMC points

Cons:

- ESS may collapse when p is large.

MCMC general remarks

The following points

- choice of starting point
- MCMC convergence assessment

are not big issues for binary regression models.

More important issues for us are:

- chain autocorrelations
- difficulty to parallelise

Gibbs

Well-known, based on data augmentation:

$$z_i = \boldsymbol{\beta}^T \boldsymbol{x}_i + \epsilon_i$$
$$y_i = \mathrm{sgn}(z_i)$$

then sample iteratively (probit/Gaussian case):

1. $\boldsymbol{\beta}|\boldsymbol{z}$ (regression posterior, tractable)
2. $\boldsymbol{z}|\boldsymbol{\beta}, \boldsymbol{y}$ (product of truncated Gaussians)

Gibbs is particularly **not generic**: any change in the prior of F requires deriving a new algorithm. This can also change the complexity (e.g. from $\mathcal{O}(p^2)$ to $\mathcal{O}(p^3)$ when using a Student prior).

Random walk Metropolis-Hastings

One iteration of RWMH

Input: $\boldsymbol{\beta}$

Output: $\boldsymbol{\beta}'$

1. Sample $\boldsymbol{\beta}^\star \sim \mathrm{N}_p(\boldsymbol{\beta}, \boldsymbol{\Sigma})$
2. With probability $1 \wedge r$,

$$r = \frac{p(\boldsymbol{\beta}^\star)p(\mathcal{D}|\boldsymbol{\beta}^\star)}{p(\boldsymbol{\beta})p(\mathcal{D}|\boldsymbol{\beta})},$$

set $\boldsymbol{\beta}' = \boldsymbol{\beta}^\star$; otherwise set $\boldsymbol{\beta}' = \boldsymbol{\beta}$

In practice, choose $\boldsymbol{\Sigma}$ as some fraction of $\boldsymbol{\Sigma}_q$.

HMC

Consider $(\boldsymbol{\beta}, \boldsymbol{\alpha})$, $\boldsymbol{\beta} \sim p(\boldsymbol{\beta}|\mathcal{D})$, $\boldsymbol{\alpha} \sim N_p(0, M^{-1})$, with joint un-normalised density $\exp\{-H(\boldsymbol{\beta}, \boldsymbol{\alpha})\}$,

$$H(\boldsymbol{\beta}, \boldsymbol{\alpha}) = E(\boldsymbol{\beta}) + \frac{1}{2}\boldsymbol{\alpha}^T \boldsymbol{M}\boldsymbol{\alpha}, \quad E(\boldsymbol{\beta}) = -\log\{p(\boldsymbol{\beta})p(\mathcal{D}|\boldsymbol{\beta})\}.$$

The physical interpretation of HMC is that of a particle at position $\boldsymbol{\beta}$, with velocity $\boldsymbol{\alpha}$, potential energy $E(\boldsymbol{\beta})$, kinetic energy $\frac{1}{2}\boldsymbol{\alpha}^T M\boldsymbol{\alpha}$, and thus total energy given by $H(\boldsymbol{\beta}, \boldsymbol{\alpha})$. The particle is expected to follow a trajectory such that $H(\boldsymbol{\beta}, \boldsymbol{\alpha})$ remains constant over time.

HMC iteration

One iteration of HMC

Input: $\boldsymbol{\beta}$

Output: $\boldsymbol{\beta}'$

1. Sample momentum $\boldsymbol{\alpha} \sim \mathrm{N}_p(0, \boldsymbol{M})$.

2. Perform L leap-frog steps, starting from $(\boldsymbol{\beta}, \boldsymbol{\alpha})$; call $(\boldsymbol{\beta}^\star, \boldsymbol{\alpha}^\star)$ the final position.

3. With probability $1 \wedge r$, $r = \exp\left\{H(\boldsymbol{\beta}, \boldsymbol{\alpha}) - H(\boldsymbol{\beta}^\star, \boldsymbol{\alpha}^\star)\right\}$ set $\boldsymbol{\beta}^{'} = \boldsymbol{\beta}^\star$; otherwise set $\boldsymbol{\beta}^{'} = \boldsymbol{\beta}$.

Leapfrog step

Leapfrog step

Input: $(\boldsymbol{\beta}, \boldsymbol{\alpha})$
Output: $(\boldsymbol{\beta}_1, \boldsymbol{\alpha}_1)$
1. $\boldsymbol{\alpha}_{1/2} \leftarrow \boldsymbol{\alpha} - \frac{\epsilon}{2} \nabla_{\boldsymbol{\beta}} E(\boldsymbol{\beta})$
2. $\boldsymbol{\beta}_1 \leftarrow \boldsymbol{\beta} + \epsilon \boldsymbol{\alpha}_{1/2}$
3. $\boldsymbol{\alpha}_1 \leftarrow \boldsymbol{\alpha}_{1/2} - \frac{\epsilon}{2} \nabla_{\boldsymbol{\beta}} E(\boldsymbol{\beta}_1)$

HMC variants

- Riemanian HMC (Girolami and Calderhead, 2011): simply too expensive
- NUTS (No U-Turn Sampler, Hoffman & Gelman, 2013): HMC with on-the-fly calibration of L and ϵ. Included in our comparisons.

SMC

We consider tempering SMC, i.e. SMC for sequence

$$\pi_t(\boldsymbol{\beta}) \propto q(\boldsymbol{\beta})^{1-\delta_t} \left\{p(\boldsymbol{\beta})p(\mathcal{D}|\boldsymbol{\beta})\right\}^{\delta_t}$$

with $0 = \delta_0 < \ldots < \delta_T = 1$.

Principle: sequence of importance sampling steps, from π_{t-1} to π_t. When weight degeneracy becomes too high, resample, and move particles through MCMC (e.g. random walk Metropolis).

The algorithm can choose the δ_j on the fly (Jasra et al, 2011).

SMC algorithm

0. Sample $\boldsymbol{\beta}_n \sim q(\boldsymbol{\beta})$ and set $\underline{\delta} \leftarrow 0$.
1. Let, for $\delta \in [\underline{\delta}, 1]$,

$$\mathrm{EF}(\delta) = \frac{1}{N} \frac{\left\{\sum_{n=1}^{N} w_\gamma(\boldsymbol{\beta}_n)\right\}^2}{\left\{\sum_{n=1}^{N} w_\gamma(\boldsymbol{\beta}_n)^2\right\}}, \quad u_\delta(\boldsymbol{\beta}) = \left\{\frac{p(\boldsymbol{\beta})p(\mathcal{D}|\boldsymbol{\beta})}{q(\boldsymbol{\beta})}\right\}^\delta.$$

If $\mathrm{EF}(1) \geq \tau$, stop and return $(\boldsymbol{\beta}_n, w_n)_{n=1:N}$, $w_n = u_1(\boldsymbol{\beta}_n)$. Otherwise, use bisection method to solve in δ equation $\mathrm{EF}(\gamma) = \tau$.

2. Resample according to normalised weights $W_n = w_n / \sum_{m=1}^{N} w_m$; with $w_n = u_\delta(\boldsymbol{\beta}_n)$.
3. Update the $\boldsymbol{\beta}_n$'s through m MCMC steps (w.r.t. $\pi_t(\boldsymbol{\beta})$).
4. Set $\underline{\delta} \leftarrow \delta$. Go to Step 1.

Remarks on SMC

- Completely automatic: we can use the current set of particles to adjust the random walk proposal, the number of MCMC steps, and so on.
- Will often collapse to a **single** IS step (when ESS from q to posterior is not too low)

First set of datasets

Dataset	$n_{\mathcal{D}}$	p
Pima (Indian diabetes)	532	8
German (credit)	999	25
Heart (Statlog)	270	14
Breast (cancer)	683	10
Liver (Indian Liver patient)	579	11
Plasma (blood screening data)	32	3
Australian (credit)	690	15
Elections	2015	52

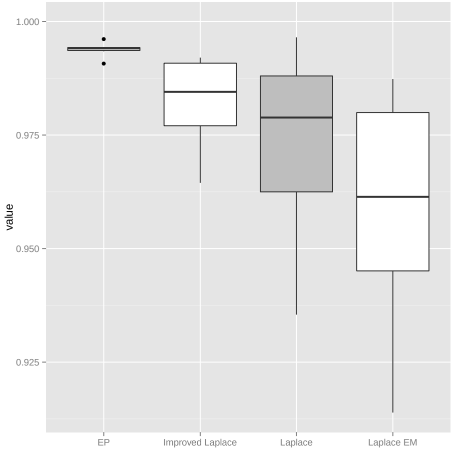
This is a superset of datasets considered in most papers.

Fast approximations

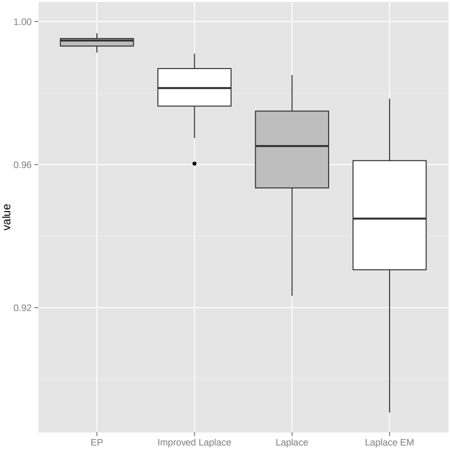
Logit/Cauchy scenario. We compare: Laplace, Improved Laplace, EM-Laplace, and EP, in term of

- marginal accuracies (one minus half the L_1 distance between approximate and true marginals)
- approximation error for marginal likelihood

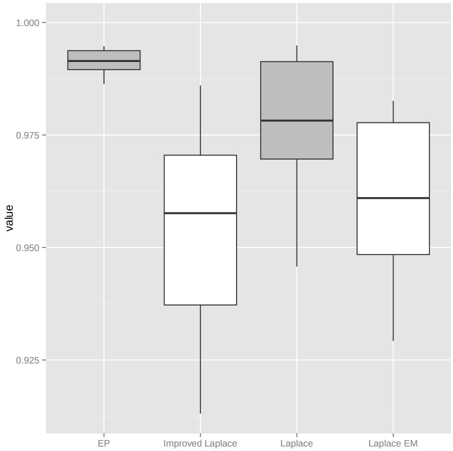
Pima


1.000
0.975
0.950
0.925
value
EP
Improved Laplace
Laplace
Laplace EM

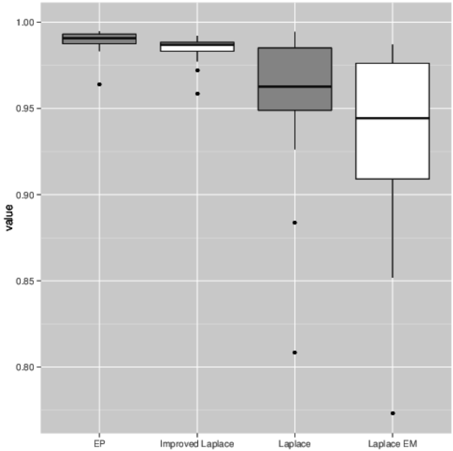
Heart



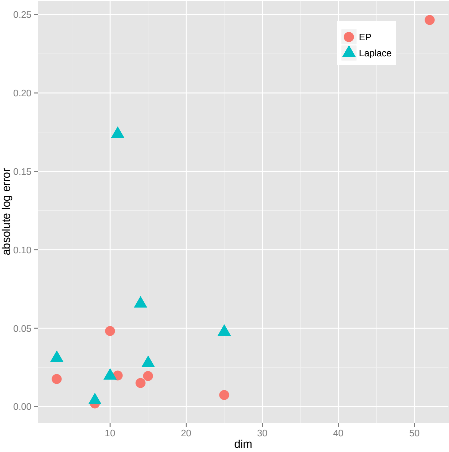
Breast


1.000
0.975
0.950
0.925
value
EP
Improved Laplace
Laplace
Laplace EM

German credit



Marginal likelihoods

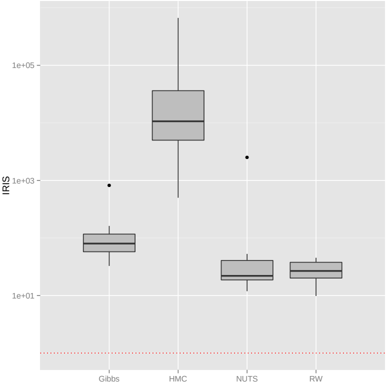


Sampling-based methods: importance sampling

	IS			IS-QMC	
Dataset	EF $= \mathrm{ESS}/N$	CPU time	MT speed-up	MSE x (expect)	MSE x (evid)
Pima	99.5%	37.54 s	4.39	28.9	42.7
German	97.9%	79.65 s	4.51	13.2	8.2
Breast	82.9%	50.91 s	4.45	2.6	6.2
Heart	95.2%	22.34 s	4.53	8.8	9.3
Liver	74.2 %	35.93 s	4.76	7.6	11.3
Plasma	90.0%	2.32 s	4.28	2.2	4.4
Australian	95.6%	53.32 s	4.57	12	20.3
Elections	21.39%	139.48 s	3.87	617.9	3.53

(Probit/Gaussian scenario, to make like easier for Gibbs)

comparison with MCMC



IRIS = Inefficiency relative to IS

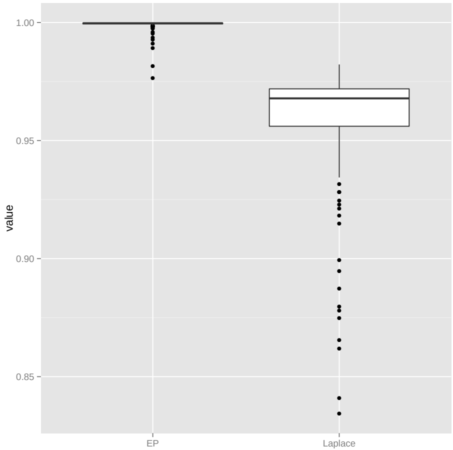
Bigger datasets

Dataset	$n_{\mathcal{D}}$	p
Musk	476	95
Sonar	208	61
DNA	400	180

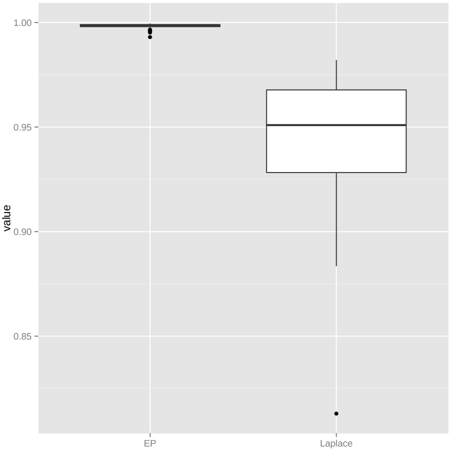
Bigger datasets, but also with higher correlations between predictors. We will look at the probit/Gaussian case.

IS no longer an option.

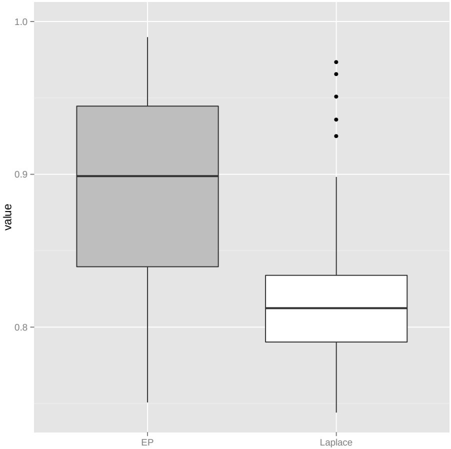
Approximations: Musk



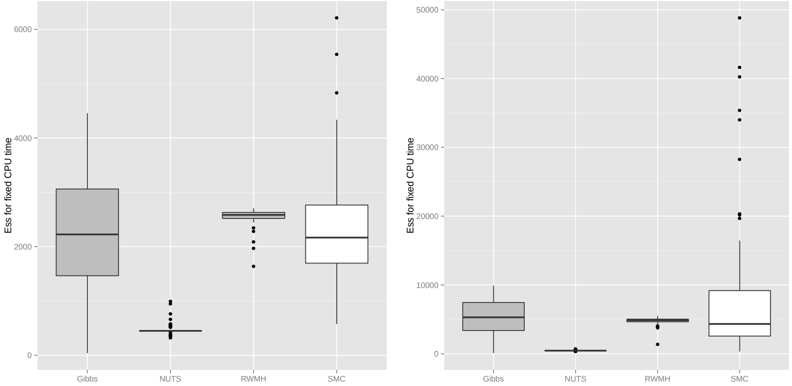
Approximations: Sonar



Approximations: DNA

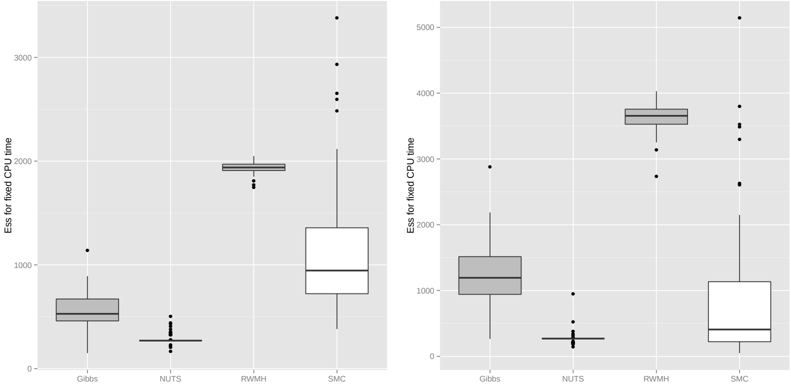


Sampling-based methods: Musk



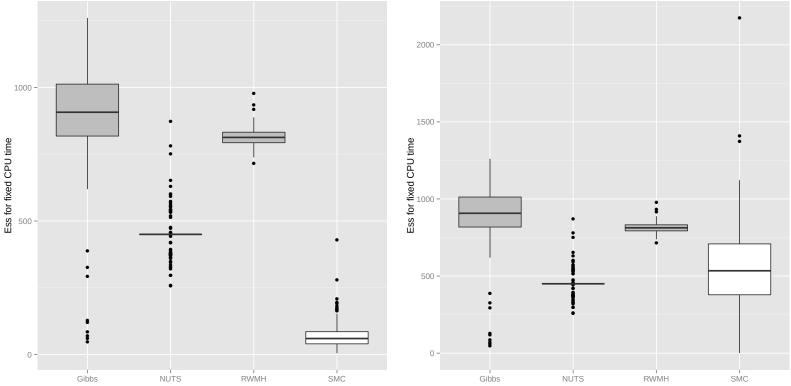
Left: posterior expectations, Right: posterior variances

Sampling-based methods: Sonar



Left: posterior expectations, Right: posterior variances

Sampling-based methods: DNA



Left: posterior expectations, Right: posterior variances

Variable selection

Add for each predictor β_j an indicator $\gamma_j \in \{0, 1\}$; prior for $\boldsymbol{\gamma}$ is Uniform over $\{0, 1\}^p$.

The posterior mixes discrete and continuous components; $p(\boldsymbol{\gamma}|\mathcal{D})$ is severely multimodal.

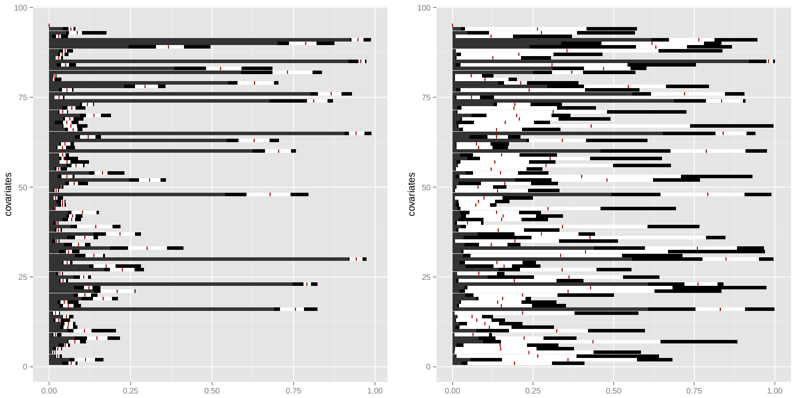
VS: proposed approach

To compute $p(\mathcal{D}|\boldsymbol{\gamma}) = \int p(\mathcal{D}|\boldsymbol{\gamma}, \boldsymbol{\beta})p(\boldsymbol{\beta}|\gamma)\, d\boldsymbol{\beta}$, use:

1. either Laplace
2. or IS based on Laplace

To simulate from $p(\boldsymbol{\gamma}|\mathcal{D})$, adapt the tempering SMC sampler of Schafer and Chopin (2013), for sampling binary vectors.

Results


100
75
50
25
0
covariates
0.00
0.25
0.50
0.75
1.00
100
75
50
25
0
covariates
0.00
0.25
0.50
0.75
1.00

Recommendations to end users (who wish to fit a binary regression model)

- EP is fast and accurate even in difficult cases.
- to improve on EP, one might run SMC; often this will collapse to IS and outperforms everything else significantly.
- That said, for large p, RWHM performs surprising well.
- HMC algorithms seem very difficult to calibrate.

Benchmarks for specialised algorithms

For specialised algorithms (Gibbs), benchmark=dataset.

It is not very clear that the Gibbs samplers developped for binary regression are very useful: corresponding papers tend to showcase these algorithms on datasets with $p < 50$, for which more generic methods fare much better.

Benchmarks for generic algorithms

For generic algorithms (e.g. RWHM), benchmark=posterior.

A binary regression posterior of dimension < 50 is very close to a Gaussian; i.e. it does not represent a very challenging benchmark. However, it is an useful **sanity check**.

More challenging benchmarks: $p \geq 100$, hierarchical regression, spike and slab prior, ...

More general remarks

Beware ML fast approximation schemes; they are fast and getting better and better...

Always compare new methods to well calibrated simple algorithms, like IS and RWHM.

Final word

Comments most welcome!