The Hamming Ball Sampler

Michalis K. Titsias
Department of Informatics
Athens University of Economics and Business, Greece

Christopher Yau
Wellcome Trust Centre for Human Genetics
University of Oxford

Motivation

Many statistical models require inference over high dimensional discrete spaces:

Sparse regression models using binary inclusion variables indicating whether the covariates affect the response

Factorial Hidden Markov models involving multiple unobserved discrete-valued latent chains contributing to a single observation process (Titsias and Yau NIPS 2014)

Markov random fields

These large data, high-dimensional set-ups are challenging for Markov Chain Monte Carlo (MCMC) methods (e.g. multimodality in the posterior is hard to deal with)

Setup

Consider Bayesian inference using MCMC for an unobserved latent discrete-valued discrete sequence or matrix $\mathbf{X} \in \mathcal{X}$

Each element $x_{ij} \in \{1, \ldots, S\}$

We have N observations $\mathbf{y} = [y_1, \ldots, y_N]$

The observations are conditionally independent given $\mathbf{X}$ and model parameters θ so that the joint distribution factorizes as

$$p(\mathbf{y}, \mathbf{X}, \theta) = \left[\prod_{i=1}^{N} p(y_i|\mathbf{X}, \theta)\right] p(\mathbf{X}, \theta)$$

We assume that the posterior distribution $p(\mathbf{X}, \theta|\mathbf{y})$ has a complex dependence structure

Standard Schemes

MCMC schemes, such as a (Metropolis-within) Gibbs Sampler, use

$$\theta \leftarrow p(\theta|\mathbf{X}, \mathbf{y}), \tag{1}$$
$$\mathbf{X} \leftarrow p(\mathbf{X}|\theta, \mathbf{y}), \tag{2}$$

or a marginal Metropolis-Hastings sampler over θ based on

$$\theta \leftarrow p(\theta|\mathbf{y}) \propto \sum_{\mathbf{X} \in \mathcal{X}} p(\mathbf{y}, \mathbf{X}, \theta), \tag{3}$$

are both **intractable**

Exhaustive summation over the entire state space of $\mathbf{X}$ has exponential complexity

Block Gibbs Sampler

A popular and tractable alternative is to employ block-conditional (Metropolis-within) Gibbs sampling

Subsets $\mathbf{x}_i$ of $\mathbf{X}$ are updated conditional on other elements being fixed using

$$\theta \leftarrow p(\theta|\mathbf{X}, \mathbf{y}), \tag{4}$$

$$\mathbf{x}_i \leftarrow p(\mathbf{x}_i|\mathbf{X}_{-i}, \theta, \mathbf{y}), \forall i, \tag{5}$$

where $\mathbf{X}_{-i}$ denotes the elements excluding those in $\mathbf{x}_i$

Typical block structures might be rows/columns of $\mathbf{X}$, when it is a matrix, or sub-blocks when $\mathbf{X}$ is a vector

Pros and Cons of the Block Gibbs Sampler

Block-conditional sampling often admits closed form updates for Gibbs sampling without resort to Metropolis-Hastings steps

However, major alterations to the configuration of $\mathbf{X}$ maybe difficult to achieve in high dimensional problems

If the elements of $\mathbf{X}$ are strongly correlated and/or with θ, conditional sampling may lead to an inability to escape from local modes

We would like to use larger blocks but the blocks cannot be too large otherwise we cannot do exhaustive enumeration within blocks

The Hamming Ball Sampler

Consider an augmented joint probability model that can be factorized as

$$p(\mathbf{y}, \mathbf{X}, \theta, \mathbf{U}) = p(\mathbf{y}, \mathbf{X}, \theta)p(\mathbf{U}|\mathbf{X})$$

$p(\mathbf{U}|\mathbf{X})$ is a conditional distribution over an auxiliary variable $\mathbf{U}$ which lives in the same space and has the same dimensions as $\mathbf{X}$

$p(\mathbf{U}|\mathbf{X})$ is chosen to be an uniform distribution over a neighborhood set $\mathcal{H}_m(\mathbf{X})$ centered at $\mathbf{X}$,

$$p(\mathbf{U}|\mathbf{X}) = \frac{1}{Z_m}\mathbb{I}(\mathbf{U} \in \mathcal{H}_m(\mathbf{X})),$$

where $\mathbb{I}(\cdot)$ denotes the indicator function and the normalizing constant Z_m is the cardinality of $\mathcal{H}_m(\mathbf{X})$

The Hamming Ball Sampler

The neighborhood set $\mathcal{H}_m(\mathbf{X})$ will be referred to as a **Hamming Ball**

Defined through Hamming distances so that

$$\mathcal{H}_m(\mathbf{X}) = \{\mathbf{U} : d(\mathbf{u}_i, \mathbf{x}_i) \leq m, i = 1, \ldots, P\}$$

The term $d(\mathbf{x}_i, \mathbf{u}_i)$ denotes the Hamming distance $\sum_j \mathbb{I}(u_{ij} \neq x_{ij})$

The pairs $(\mathbf{u}_i, \mathbf{x}_i)$ denote non-overlapping subsets of corresponding entries in $(\mathbf{U}, \mathbf{X})$ such that $\cup_{i=1}^{P} \mathbf{u}_i = \mathbf{U}$ and $\cup_{i=1}^{P} \mathbf{x}_i = \mathbf{X}$

The parameter m denotes the maximal distance or radius of each individual Hamming Ball set (the number of bits we can change at once)

The Hamming Ball Sampler

Example

Pairs can correspond to different matrix columns.

$\mathbf{x}_i$ will be the i-th column of $\mathbf{X}$

$\mathbf{u}_i$ the corresponding column of $\mathbf{U}$

The Hamming Ball $\mathcal{H}_m(\mathbf{X})$ would consist of all matrices whose columns are *at most* m elements different to $\mathbf{X}$

Gibbs Sampling in the Augmented Space

Hamming Ball Sampling is just Gibbs sampling for the augmented joint probability distribution $p(\mathbf{y}, \mathbf{X}, \theta, \mathbf{U})$

The target posterior distribution $p(\mathbf{X}, \theta|\mathbf{y})$ is admitted as a by-product

The Hamming Ball Sampler alternates between the steps:

$$\mathbf{U} \leftarrow p(\mathbf{U}|\mathbf{X}), \tag{6}$$

$$(\theta, \mathbf{X}) \leftarrow p(\theta, \mathbf{X}|\mathbf{y}, \mathbf{U}). \tag{7}$$

The update of $(\theta, \mathbf{X})$ can be implemented as conditional (Gibbs) updates or as a joint M-H step (model-dependent)

Restricted State Space

Conditioning on $\mathbf{U}$ has some computational advantages

For example, $p(\mathbf{X}|\theta, \mathbf{U}, \mathbf{y})$ is given by:

$$p(\mathbf{X}|\theta, \mathbf{U}, \mathbf{y}) = \frac{p(\mathbf{y}, \mathbf{X}, \theta)p(\mathbf{U}|\mathbf{X})}{p(\theta, \mathbf{U}, \mathbf{y})} \propto p(\mathbf{y}, \mathbf{X}, \theta)\mathbb{I}(\mathbf{X} \in \mathcal{H}_m(\mathbf{U}))$$

The normalizing constant is found by exhaustive summation over all **admissible** matrices inside the Hamming Ball $\mathcal{H}_m(\mathbf{U})$

We will normally choose m so that the cardinality of $\mathcal{H}_m(\mathbf{U})$ will be considerably **less than** the cardinality of $\mathcal{X}$

Exhaustive enumeration of all elements inside the Hamming Ball would be **computationally feasible**

Algorithm Overview

To summarise:

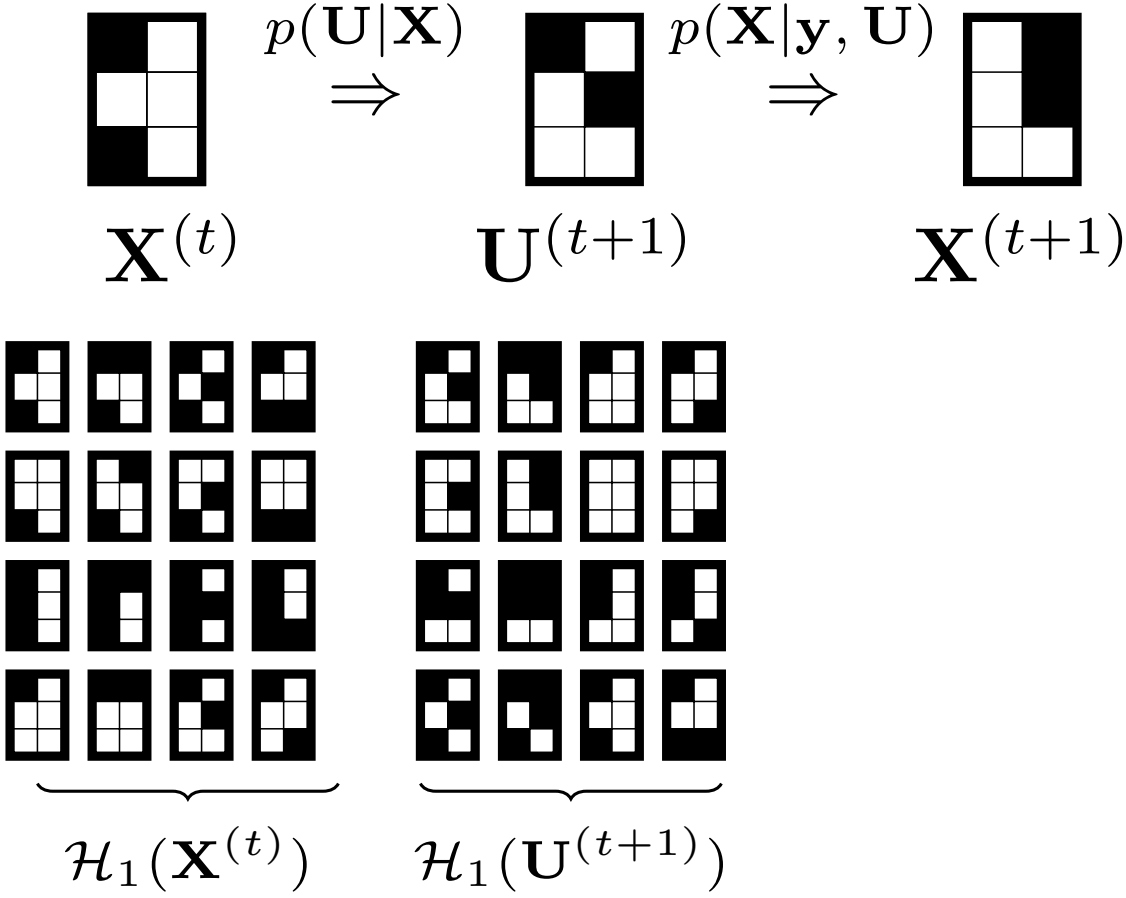
1. Use an auxiliary variable $\mathbf{U}$ to define a slice of the model given by $\mathcal{H}_m(\mathbf{U})$

2. Sampling of $(\theta, \mathbf{X})$ is performed efficiently within this sliced part of the model through $p(\theta, \mathbf{X}|\mathbf{U}, \mathbf{y})$

3. At each iteration, this model slice randomly moves via the re-sampling of $\mathbf{U}$, which simply sets $\mathbf{U}$ to a random element from $\mathcal{H}_m(\mathbf{X})$

The final re-sampling step allows for **random exploration** that is necessary to ensure that the overall sampling scheme is **ergodic**

Algorithm Overview



General Blocking Scheme

The selection of the subsets or blocks $\{\mathbf{x}_1, \ldots, \mathbf{x}_P\}$ will depend on the conditional dependencies specified by the statistical model

In unstructured models ($\mathbf{X}$ is just a large pool of fully dependent discrete variables), we can divide the variables into randomly chosen blocks

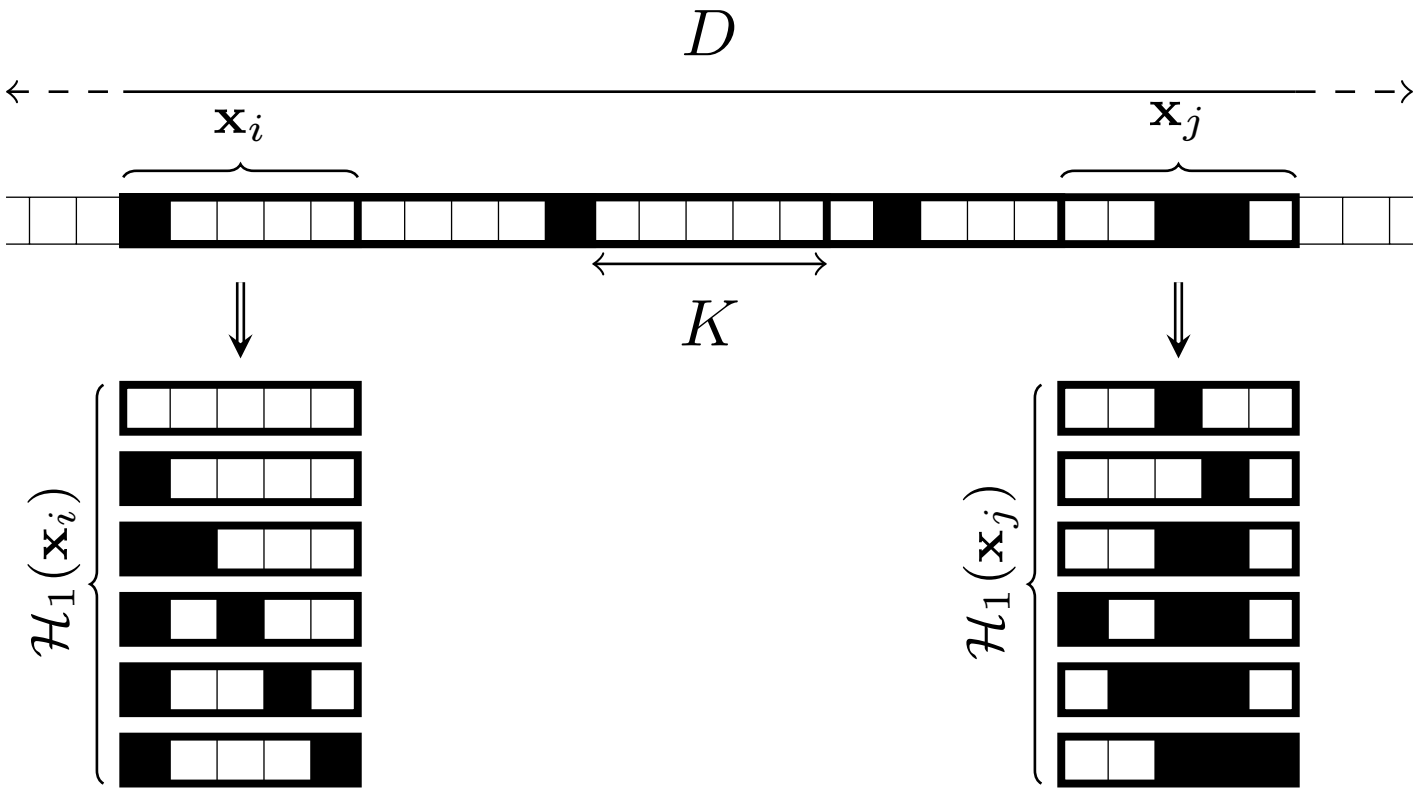
Exact simulation from $p(\mathbf{X}|\theta, \mathbf{U}, \mathbf{y})$ may not be feasible and instead we can use the Hamming Balls to sequentially sample each block

This variant of the algorithm can be based on the iteration on P sequential conditional steps

$$\mathbf{u}_i \leftarrow p(\mathbf{u}_i|\mathbf{x}_i), \ \mathbf{x}_i \leftarrow p(\mathbf{x}_i|\mathbf{X}_{-i}, \theta, \mathbf{u}_i, \mathbf{y}), \forall i. \tag{8}$$

General Blocking Scheme

Use random subsets of $\mathbf{X}$ and apply Hamming Ball Sampling



Unlike a Block Gibbs Sampler, exhaustive summation is not necessary within each block $\Rightarrow$ can choose larger block sizes

Computational Complexity

For P blocks of size K, the computational complexity of the Hamming Ball Sampler scales with the Hamming radius m, block size K and P according to $O(MP)$ where

$$M = \sum_{j=0}^{m} (S-1)^j \binom{K}{j}$$

The block Gibbs Sampler has computational complexity of $O(S^K P)$

$$S^K = \sum_{j=0}^{K} (S-1)^j \binom{K}{j}$$

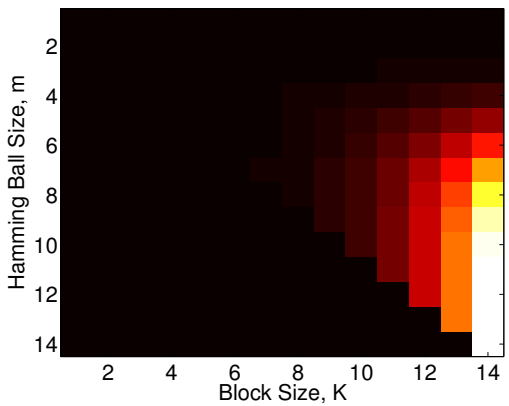
and it is only applicable for small values of block size K

Block Gibbs Sampler is a **special case** of a Hamming Ball Sampler

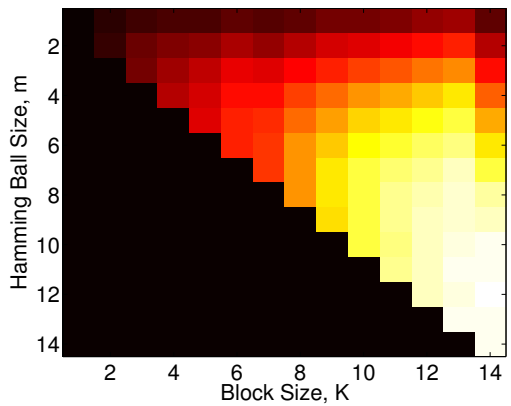
Hamming Ball sampling is **more flexible** by allowing control over the computational cost through both K and m

Computational Complexity and Sampling Efficiency

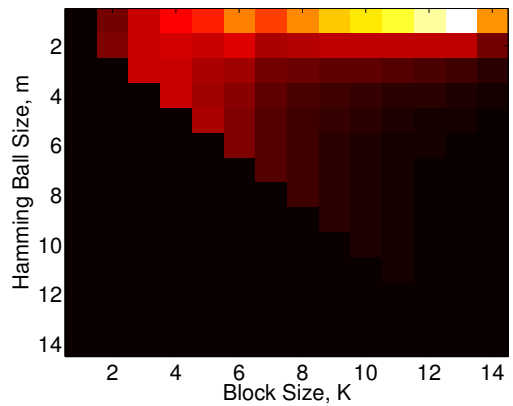
For fixed blocksize K all Hamming Ball schemes can have radius $m \leq K$ (we can visualize this as upper triangular matrix)



Complexity



Sampling Efficiency



Overall Efficiency

The diagonal elements corresponds to standard Block Gibbs Samplers. The other ones ($m < K$) are pure Hamming Ball schemes

Sparse Linear Regression

We have the $N \times 1$ vector $\mathbf{y}$ of responses (normalized to have zero mean) and the $N \times D$ design matrix $\mathbf{Z}$ with covariates

The $D \times 1$ latent vector $\mathbf{X}$ encodes variable inclusion

$$x_d \sim \text{Bernoulli}(x_d, \pi_0), \quad d = 1, \ldots, D, \quad \pi_0 \sim \text{Beta}(\pi_0 | \alpha_{\pi_0}, b_{\pi_0})$$

Responses are generated from a Gaussian linear regression model

$$\mathbf{y} = \mathbf{Z}_\mathbf{X} \beta_\mathbf{X} + \eta, \quad \eta \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_N),$$

where $\mathbf{Z}_\mathbf{X}$ is the $N \times D_\mathbf{X}$ design sub-matrix, with columns corresponding to $x_d = 1$ and $\beta_\mathbf{X}$ is the $D_\mathbf{X} \times 1$ vector of regression coefficients

Conjugate g-prior on $\beta_\mathbf{X}$

$$p(\beta_\mathbf{X}, \sigma^2 | \mathbf{X}) = \mathcal{N}(\beta_\mathbf{X} | \mathbf{0}, g(\mathbf{Z}_\mathbf{X}^T \mathbf{Z}_\mathbf{X})^{-1}) \text{InvGa}(\sigma^2 | \alpha_\sigma, b_\sigma)$$

Sparse Linear Regression

Analytically marginalize out the parameters $\theta = (\pi_0, \boldsymbol{\beta}_{\mathbf{x}}, \sigma^2)$ (as in Bottolo and Richardson, 2010):

$$p(\mathbf{y}, \mathbf{X}|\cdot) \propto C \left(2b_\sigma + S(\mathbf{X})\right)^{-(2\alpha_\sigma + N - 1)/2},$$

where

$$C = (1+g)^{-D_{\mathbf{X}}/2}\Gamma(D_{\mathbf{X}} + \alpha_{\pi_0})\Gamma(D - D_{\mathbf{X}} + b_{\pi_0}),$$
$$S(\mathbf{X}) = \mathbf{y}^T\mathbf{y} - \frac{g}{1+g}\mathbf{y}^T\mathbf{Z}_{\mathbf{X}}(\mathbf{Z}_{\mathbf{X}}^T\mathbf{Z}_{\mathbf{X}})^{-1}\mathbf{Z}_{\mathbf{X}}^T\mathbf{y}$$

and $\Gamma(\cdot)$ denotes the Gamma function

Sparse Linear Regression

1. Randomly initialize $\mathbf{X}^{(0)}$ and set $t = 0$.
2. At iteration $t + 1 = 1, \ldots, T$ randomly divide $\mathbf{X}$ into $P = D/K$ blocks $\mathbf{x}_i$. Then for $i = 1, \ldots, P$

   2.1 Sample auxiliary variables $\mathbf{u}_i^{(t+1)}$ from

$$p(\mathbf{u}_i^{(t+1)} | \mathbf{x}_i^{(t)}) = \frac{1}{\sum_{\mathbf{u}_i^{(t+1)} \in \mathcal{H}_m(\mathbf{x}_i^{(t)})} 1}, \forall \mathbf{u}_i^{(t+1)} \in \mathcal{H}_m(\mathbf{x}_i^{(t)})$$

   2.2 Sample $\mathbf{x}_i^{(t+1)}$ from

$$\frac{p(\mathbf{y}, \mathbf{x}_1^{(t+1)}, \ldots, \mathbf{x}_i^{(t+1)}, \mathbf{x}_{i+1}^{(t)}, \ldots, \mathbf{x}_P^{(t)} | \cdot)}{\sum_{\mathbf{x}_i^{(t+1)} \in \mathcal{H}_m(\mathbf{u}_i^{(t+1)})} p(\mathbf{y}, \mathbf{x}_1^{(t+1)}, \ldots, \mathbf{x}_i^{(t+1)}, \mathbf{x}_{i+1}^{(t)}, \ldots, \mathbf{x}_P^{(t)} | \cdot).}$$

We consider schemes with fixed block size $K = 10$ and Hamming radius $m = 1, 2, 3$

Also consider block Gibbs Samplers: jointly sample between $1, 2, 3$ elements of $\mathbf{X} \Rightarrow$ special cases of the Hamming Ball algorithm above

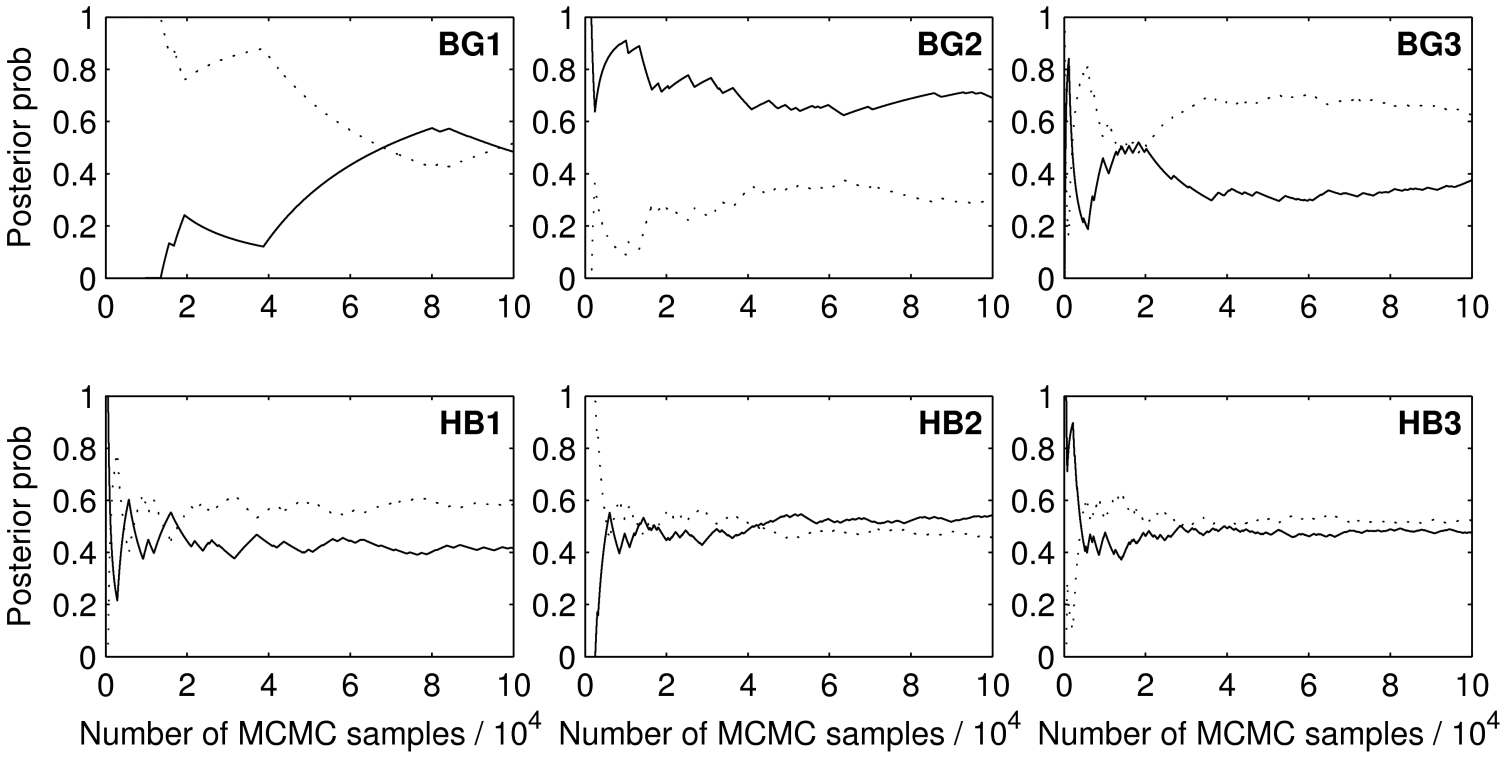
Sparse Linear Regression

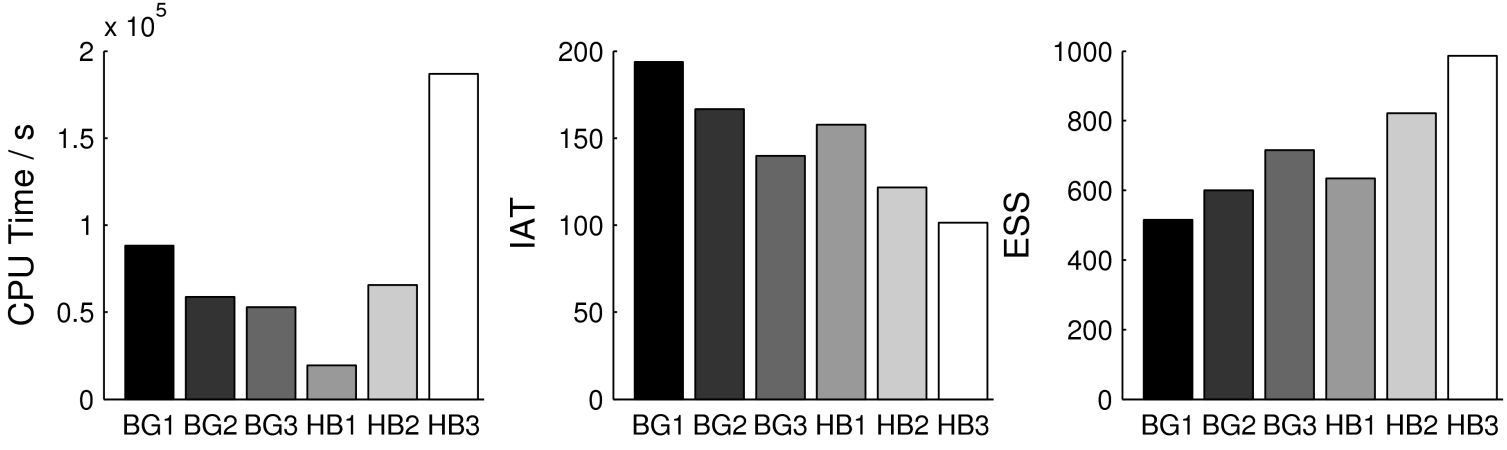
We simulated a regression dataset with $N = 100$ responses and $D = 1,200$ covariates

There were two relevant covariates that fully explain the data while the reminder were noisy redundant inputs

A challenging model exploration problem as only two out of 2^{1200} possible models represent the possible truth

Sparse Linear Regression

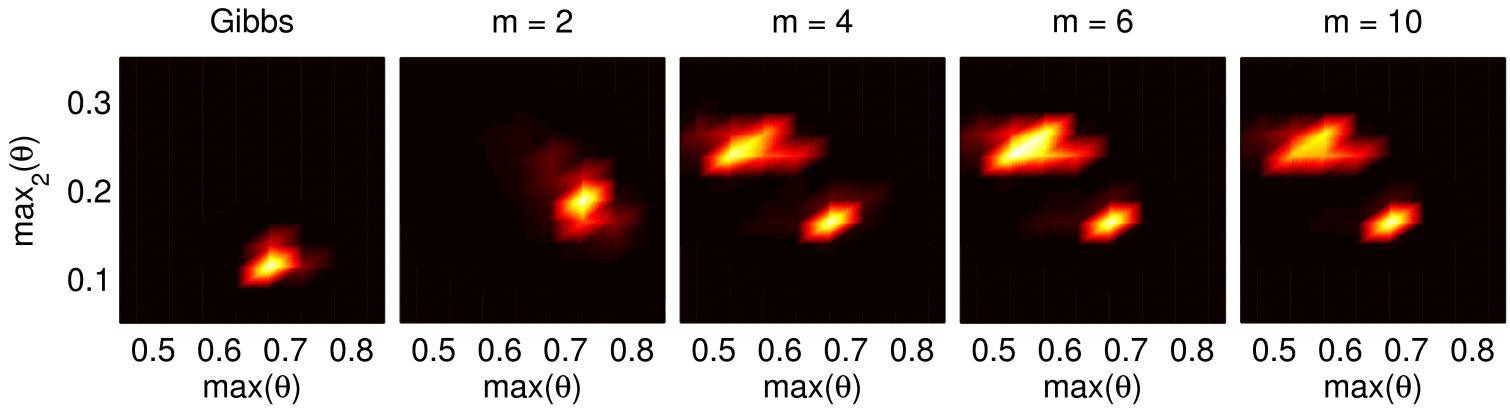




Breast Cancer

Application to 12 tumors from a single breast cancer patient (Zare et. al 2014)

Marginal parameter posteriors for one of the tumor samples:



In the original study, only a single genetic architecture was reported

Here also **standard Gibbs get stuck to a single posterior mode** and we cannot conclude the alternative explanation of the data based on its output

Our analysis suggests that there is ambiguity (multimodality in the posterior) in the genetic architecture of this particular tumor sample

Conclusions

Hamming Ball Sampling generalizes Block Gibbs Sampling

Adds flexibility in terms of balancing computational complexity and statistical efficiency

Requires simple changes to pre-existing implementations

HB Sampling makes efficient MCMC inference for these types of models feasible for large problems (otherwise MCMC may not be viable at all)

Properties not discussed: choice of $p(\mathbf{U}|\mathbf{X})$ (is does not have to be uniform!), M-H extensions, choosing m

More details: Titsias and Yau (2015). The Hamming Ball Sampler. arXiv