Complexity and optimization of the Gibbs Sampler for multilevel linear models

Giacomo Zanella
joint work with Omiros Papaspiliopoulos and Gareth Roberts

Department of Decision Sciences, BIDSA and IGIER
Bocconi University

AUEB
3rd May 2018

Context: Bayesian multilevel models

- Complex models built via combination of local and simpler distributions
- Extremely powerful and successful paradigm: flexibility, interpretability, borrowing of information,...[1]
- Naturally lend themselves to Gibbs Sampling schemes where you update a subset of variables conditional on the others

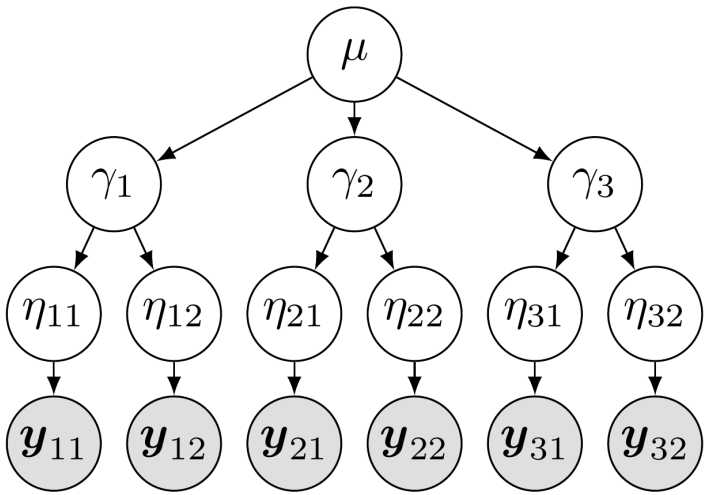


Figure: Hierarchical structure induced by a multilevel model

[1] Gelman&Hill (2006) Data analysis using regression and multilevel/hierarchical models. Cambridge U.Press

Complexity&optimization of MCMC for multilevel models

Aim: improve theoretical understanding and methodological guidance for MCMC on multilevel models.

This talk:

- consider the Gibbs Sampler and multilevel Gaussian models
- explore the interaction between model structure and algorithms' behavior
- Provide *quantitative* theory with *methodological implications*, e.g.
  1. complexity statements
  2. guidance on optimal implementations

NB: large literature on MCMC theory deals with generic target distributions, here we consider *structured* data.

Overview of the talk

1. Introduction
2. Nested linear models
   - Introduce multigrid decomposition
   - Hierarchical ordering
   - Reparametrizations
3. Crossed effect models
   - Multigrid analysis
   - Recovering scalability
   - Effect of sparsity
4. Conclusions and future work

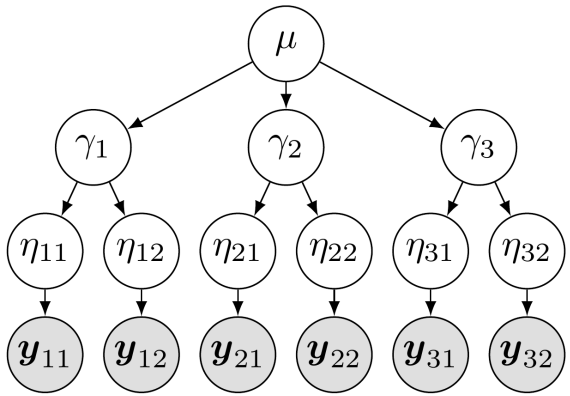


Figure: Nested effects models

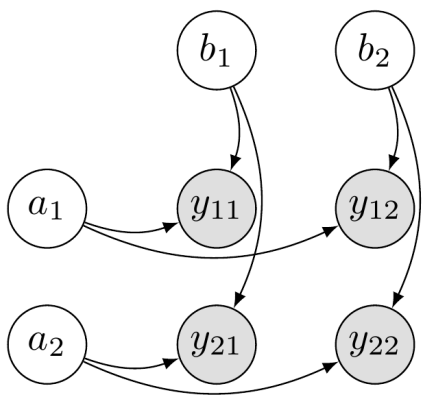


Figure: Crossed effects models

Nested linear models

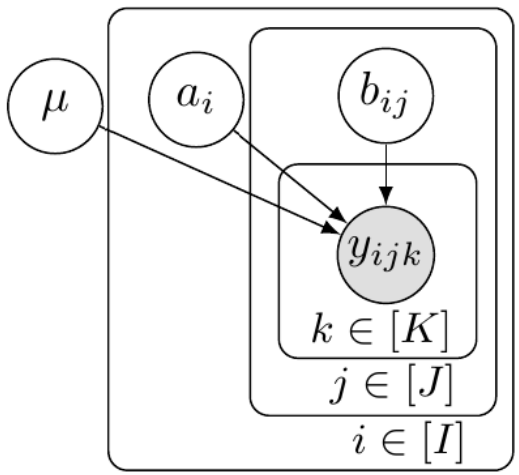
3-level nested model:

Likelihood: $y_{ijk}|\mu, a, b \sim N(\mu + a_i + b_{ij}, \tau_e^{-1}) \quad i\in[I], j\in[J], k\in[K]$

Prior: $b_{ij} \overset{iid}{\sim} N(0, \tau_b^{-1}), a_i \overset{iid}{\sim} N(0, \tau_a^{-1}), p(\mu) \propto 1.$

Standard Gibbs Sampler for $(\mu, a, b)|y$

1. Sample $\mu \sim p(\mu|a, b, y)$
2. Sample $a_i \sim p(a_i|\mu, b, y)$ for all i
3. Sample $b_{ij} \sim p(b_{ij}|\mu, a, y)$ for all i, j



Question: what is the computational complexity of GS?

NB: we are considering the fixed-variance scenario. Typically variance parameters are given a prior distribution and GS is embedded in a scheme updating also those.

Complexity of MCMC

For iterative sampling algorithms like MCMC

$$Cost_{alg} = Cost_{iter} \cdot T_{mix}$$

$Cost_{iter}$ typically easy to compute. For Gibbs often $Cost_{iter} = O(N)$

Technically challenging part: quantify T_{mix}.

We seek algorithms with good scalability, e.g. $Cost_{alg} \leq O(N)$

Approach and main technical tool

There are different notions of T_{mix}. In this talk, we will consider the following.
Definition: The rate of convergence of a Markov chain $X_1, X_2 \dots$ is the smallest number ρ such that

$$\|\mathcal{L}(X_t|X_0 = x) - \pi\| \leq C(x)\rho^t$$

The rate of convergence can be interpreted in terms of convergence time as

$$T_{mix} = \frac{1}{1-\rho}$$

Intuition: $T_{mix} \approx$ number of iterations needed to get each iid sample.
Example: $\rho = 0.999 \Rightarrow T_{mix} \approx 1000$

Gaussian Gibbs Samplers

Many proofs of $\rho < 1$ (i.e. geometric ergodicity) under mild assumptions. However, computing ρ exactly (or even bounding it) is very difficult in practice! An important exception is given by Gaussian autoregressions.

A Gibbs Sampler targeting $N(0, \Sigma)$ becomes a simple AR(1) process

$$X_t = BX_{t-1} + noise$$

where B is an explicit function of Σ. In this context, the Gibbs Sampler rate of convergence coincide with the largest eigenvalue of B, $\rho(B)$. [2] [3] Issue in practice is the high-dimensionality of B, which equals the number of parameters p.

[2] Amit (1996) Convergence properties of the Gibbs Sampler for perturbations of Gaussians.Ann.Statist.

[3] Roberts&Sahu(1997)Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. JRSS-B

Back to nested models

Model: $y_{ijk}|\mu, a, b \sim N(\mu + a_i + b_{ij}, \tau_e^{-1})$

MCMC: the Markov chain $((\mu, a, b)(t))_{t=0}^{\infty}$ induced by the Gibbs Sampler is a Gaussian auto-regression However, it is high-dimensional (1+I+IJ).

Basic idea: find a decomposition of $(\mu, a, b)(t)$ into easier and lower-dimensional chains that allows direct analysis

Multigrid decomposition

Map $(\mu, a, b) \mapsto (\delta^{(0)}, \delta^{(1)}, \delta^{(2)})$ by

1. decomposing (μ, a, b) into residuals at different levels of granularity:

$$\begin{aligned} b_{ij} &= \bar{b} + (\bar{b}_i - \bar{b}) + (b_{ij} - \bar{b}_i) = \delta^{(0)} b + \delta^{(1)} b_i + \delta^{(2)} b_{ij} \\ a_i &= \bar{a} + (a_i - \bar{a}) \qquad\qquad = \delta^{(0)} a + \delta^{(1)} a_i \\ \mu &= \mu \qquad\qquad\qquad\quad = \delta^{(0)} \mu \end{aligned}$$

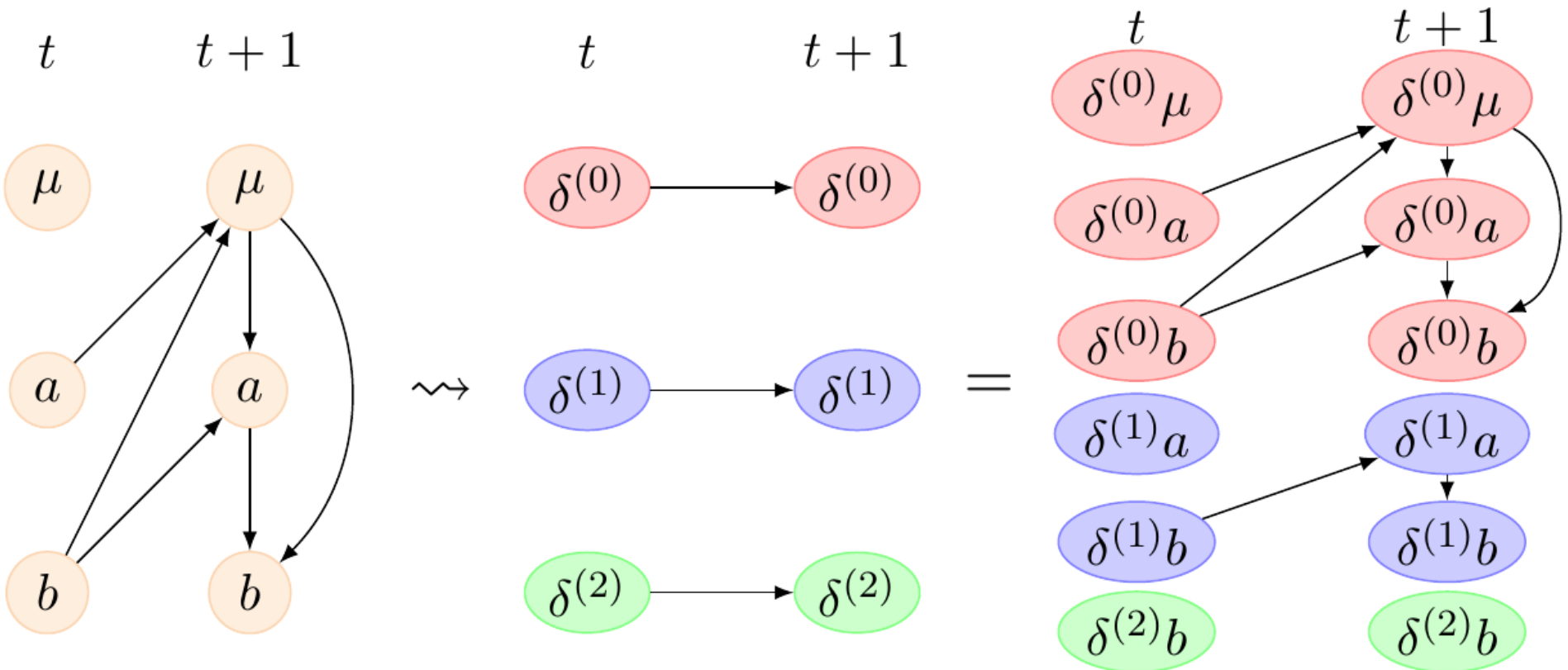
   where $\bar{a} = \frac{1}{I} \sum_i a_i$, $\bar{b} = \frac{1}{IJ} \sum_{ij} b_{ij}$ and $\bar{b}_i = \frac{1}{J} \sum_j b_{ij}$.

2. re-arrange terms and consider

$$\begin{aligned} \delta^{(0)} &= (\delta^{(0)} \mu, \delta^{(0)} a, \delta^{(0)} b) && \in \mathbb{R}^3 \\ \delta^{(1)} &= (\delta^{(1)} a_i, \delta^{(1)} b_i)_i && \in \mathbb{R}^{2I} \\ \delta^{(2)} &= (\delta^{(2)} b_{ij})_{ij} && \in \mathbb{R}^{IJ} \end{aligned}$$

Theorem (Multigrid decomposition of GS)

Let $((\mu, a, b)(t))_{t=0}^{\infty}$ be the Markov chain generated by the Gibbs Sampler. Then $\delta^{(0)}(t)$, $\delta^{(1)}(t)$ and $\delta^{(2)}(t)$ are three independent Markov chains.



Corollary: The mixing time of GS is $T_{gibbs} = \max\{T(\delta^{(0)}), T(\delta^{(1)}), T(\delta^{(2)})\}$

Target decomposition $\neq$ MCMC decomposition

Toy example

(x, y) bivariate gaussian with correlation ρ. Then:

- x and $z = y - \rho x$ are independent r.v.s under the target, *but*
- the stochastic processes $x(t)$ and $z(t)$ induced by the Gibbs Sampler are not independent Markov chains.

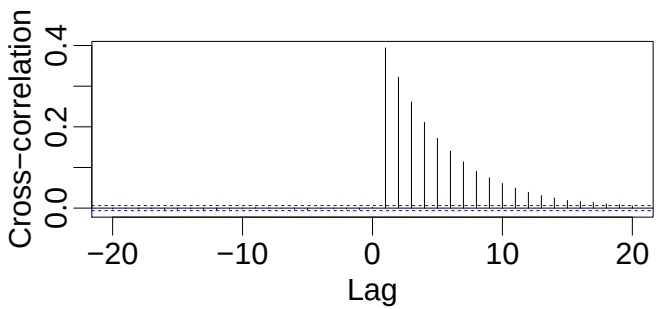


Figure: Cross correlation between $x(t)$ and $z(t)$

For crossed (and nested) random effect models the multigrid decomposition for MCMC has to do with model structure.

Multigrid decomposition - Nested model case

Theorem (Hierarchical ordering of mixing times)

$$T(\delta^{(0)}) \geq T(\delta^{(1)}) \geq T(\delta^{(2)})$$

$\Rightarrow$ convergence behavior of GS is monotonic with granularity (coarsest=slowest)

Corollary

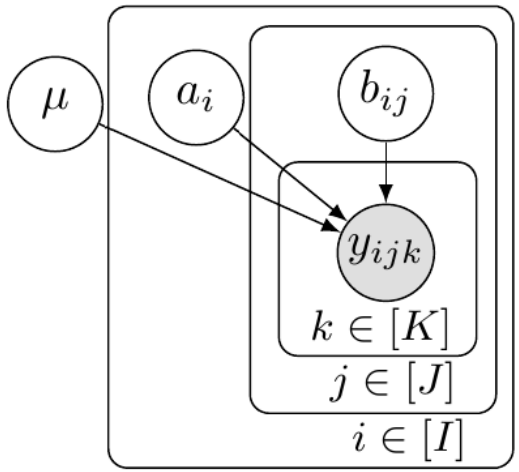
$$T_{gibbs} = T(\delta^{(0)}) = 1 + JK \frac{\tau_e}{\min\{\tau_a, J\tau_b\}}$$

Therefore

$$Cost_{gibbs} = O(JK \cdot N)$$

$\Rightarrow$ mixing deteriorates as model/data size increase and total cost is super-linear!

Reparametrizations



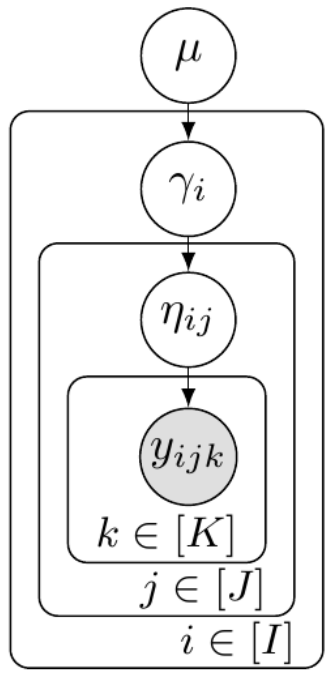
Original model: $y_{ijk} \sim N(\mu + a_i + b_{ij}, \tau_e^{-1})$

Sampler $GS(\mu, a, b)$:

1. Sample $\mu \sim p(\mu|a, b)$
2. Sample $a_i \sim p(a_i|\mu, b)$ for all i
3. Sample $b_{ij} \sim p(b_{ij}|\mu, a)$ for all i, j

Centered parametrization:[4] [5] define $\gamma_i = \mu + a_i$ and $\eta_{ij} = \gamma_i + b_{ij}$. Re-write the model as:

$y_{ijk} \sim N(\eta_{ij}, \tau_e^{-1}), \quad \eta_{ij} \sim N(\gamma_i, \tau_b^{-1}), \quad \gamma_i \sim N(\mu, \tau_a^{-1})$



Sampler $GS(\mu, \gamma, \eta)$:

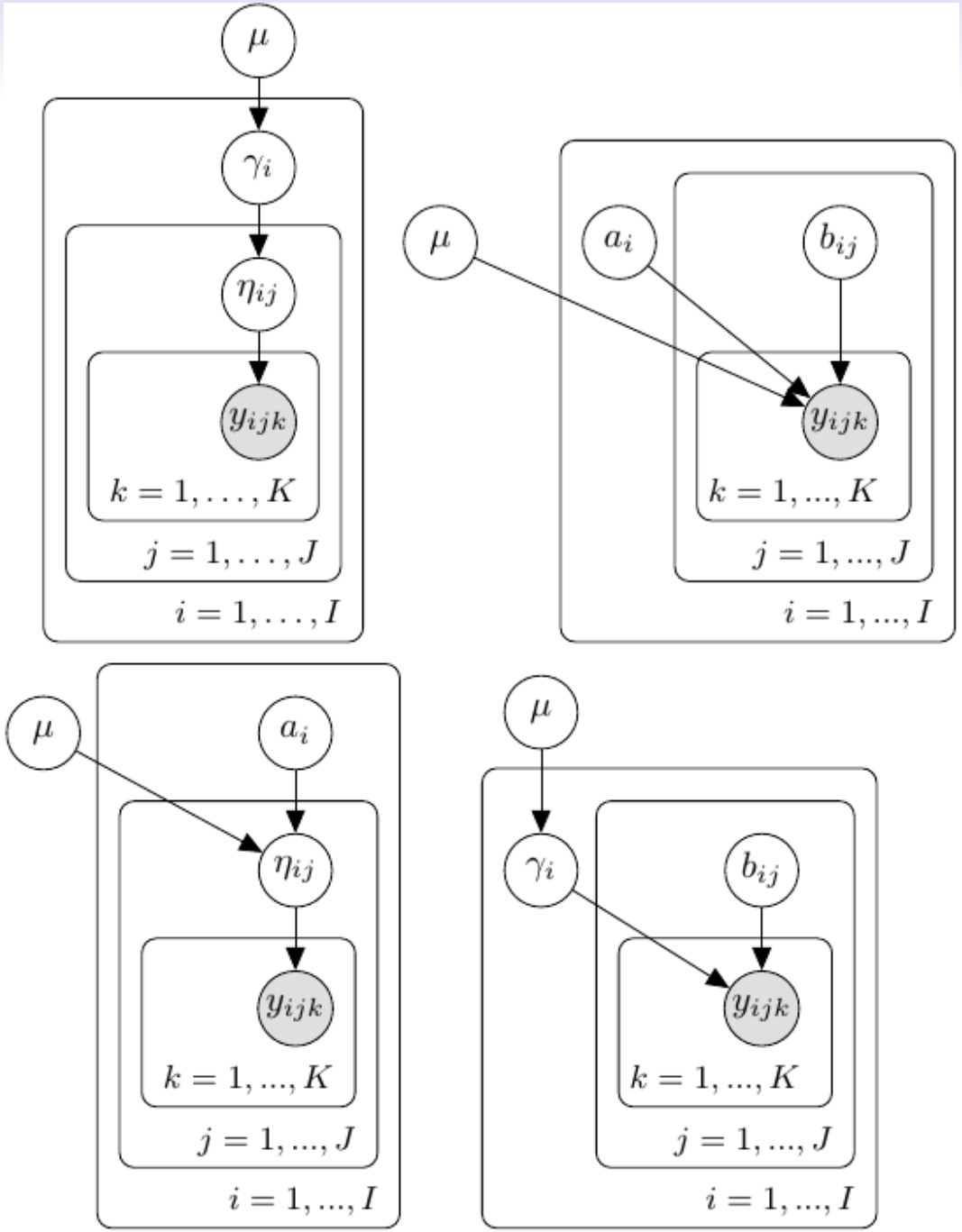
1. Sample $\mu \sim p(\mu|\gamma, \eta)$
2. Sample $\gamma_i \sim p(\gamma_i|\mu, \eta)$ for all i
3. Sample $\eta_{ij} \sim p(\eta_{ij}|\mu, \gamma)$ for all i, j

[4] Gelfand,Sahu&Carlin (1995) Efficient parametrisations for normal linear mixed models. Biometrika

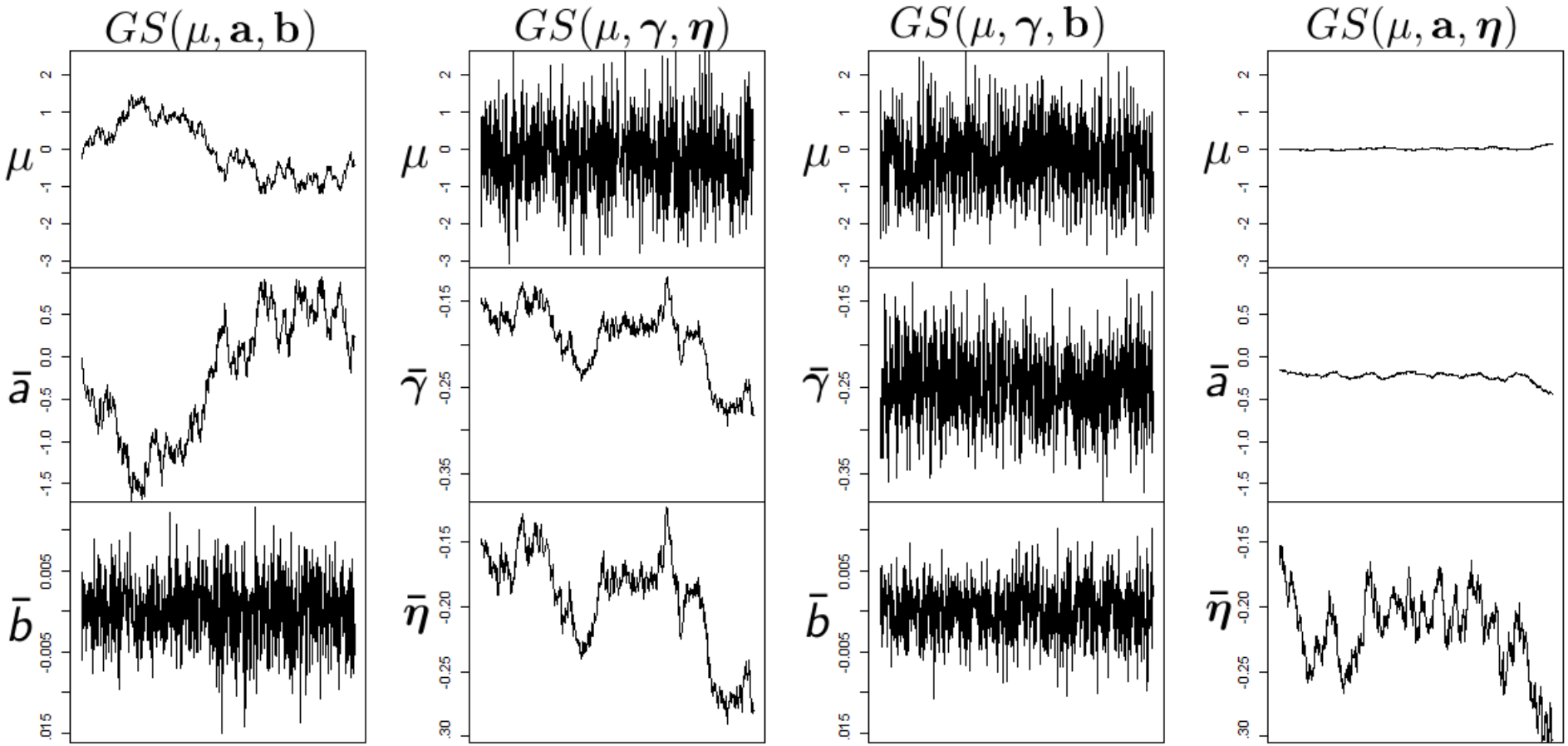
[5] Papaspiliopoulos et al.(2007)A general framework for the parametrization of hier.models. Stat.Science

For 3-level nested models we have four natural parametrizations leading to four Gibbs Samplers:

- $GS(\mu, a, b)$
- $GS(\mu, \gamma, \eta)$
- $GS(\mu, a, \eta)$
- $GS(\mu, \gamma, b)$



Change of parametrizations often have major effects on MCMC convergence!



Multigrid decomposition allows to derive mixing times for all parametrizations

Theorem (Explicit rates for different parametrizations)

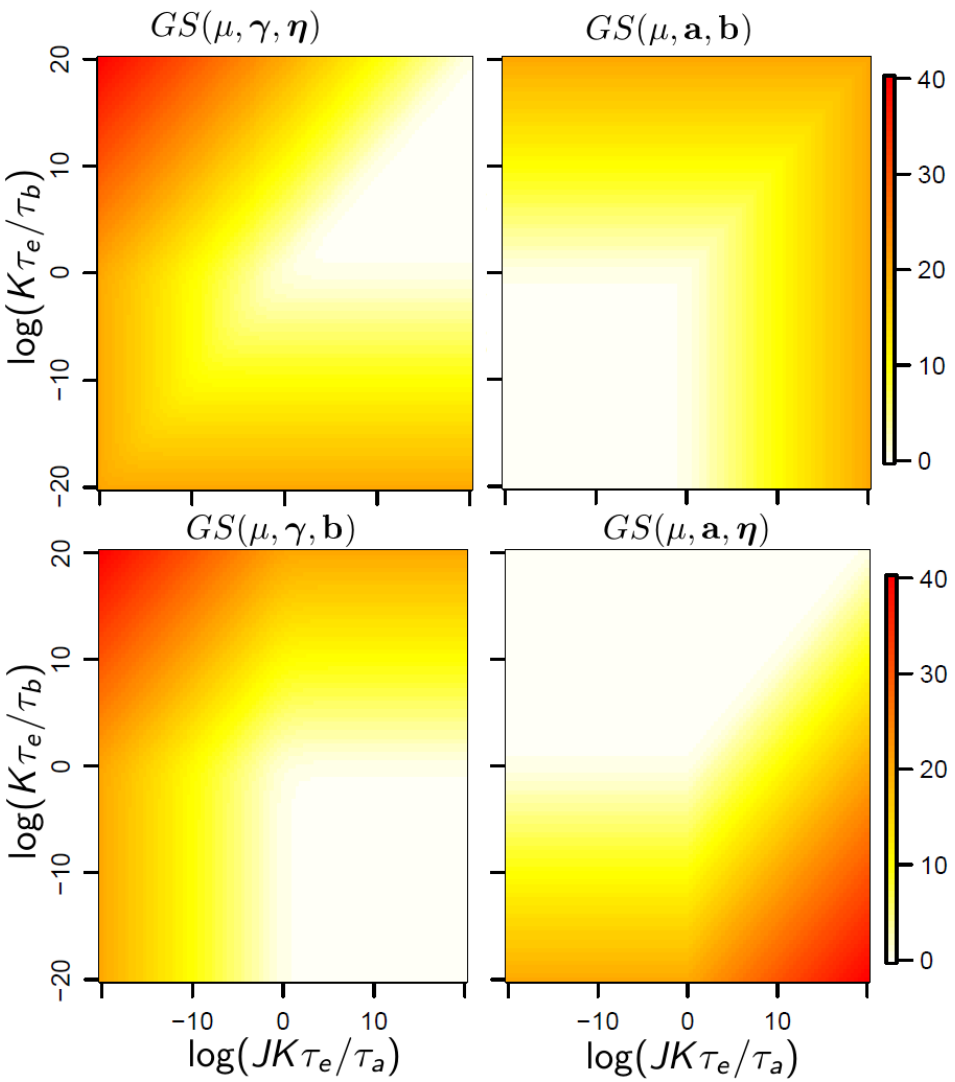
$$T_{(\mu,\gamma,\eta)} = \left(1+\frac{\tau_a}{J\tau_b}\right)\left(1+\frac{\tau_b}{K\tau_e}\right)$$

$$T_{(\mu,a,b)} = 1 + JK\frac{\tau_e}{\min\{\tau_a, J\tau_b\}}$$

$$T_{(\mu,\gamma,b)} = \left(1+\frac{\tau_a}{JK\tau_e}\right)\left(1+\frac{K\tau_e}{\tau_b}\right)$$

$$T_{(\mu,a,\eta)} = 1 + J\frac{\tau_b}{\min\{\tau_a, JK\tau_e\}}$$

Mixing time for different parametrizations (white=fast)



Corollary

To obtain the parametrization with the smallest mixing time

replace a with γ *iff* $Var(\bar{a}) \geq Var(\bar{b}) + Var(\bar{y})$ $\left(i.e.\ \frac{1}{\tau_a} \geq \frac{1}{J\tau_b} + \frac{1}{JK\tau_e} \right)$

replace b with η *iff* $Var(\bar{b}) \geq Var(\bar{y})$ $\left(i.e.\ \frac{1}{\tau_b} \geq \frac{1}{K\tau_e} \right)$

Under the optimal parametrization $T_{gibbs} \leq 3 \quad \Rightarrow \quad Cost_{gibbs} = O(N)$

In the unknown variances case, the parametrization can be optimized "on the fly"

More details in preprint [6] :

- Generalization to arbitrary tree structure
- Hierarchical ordering of rates for k levels (Cauchy interlacing theorem)
- Bounds for general non-symmetric scenarios
- Analysis of partially non-centered and bespoke parametrizations
- ...

[6] G.Zanella&G.Roberts (2017) Analysis of the Gibbs Sampler for Gaussian hierarchical models via multigrid decomposition. Preprint.

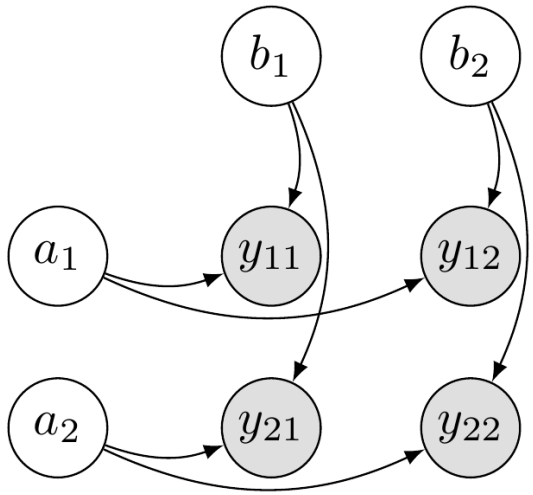
Beyond nested structures: crossed effects

2-factors crossed effects model

Likelihood: $y_{ij} \sim N(\mu + a_i + b_j, \tau_e^{-1}) \qquad i \in [I], j \in [J]$

Prior: $b_j \overset{iid}{\sim} N(0, \tau_b^{-1})$, $a_i \overset{iid}{\sim} N(0, \tau_a^{-1})$, $p(\mu) \propto 1$.

e.g. in recommender systems i denotes items and j users



- Crossed structure has major computational implications: no tree-based algorithms, cost of marginal likelihood and exact sampling is $O(N^{3/2})$, ...
- Cost driven by inversions of large Gaussian precision matrix. Sparse linear algebra techniques could be used, but the precision matrix has no specific structure (e.g. banded) and could even be dense.
- Motivated by recent work on method of moments to get $O(N)$ algorithms [7]
- How does MCMC perform here?

[7] K.Gao&A.Owen (2017) Estimation and Inference for Very Large Linear Mixed Effects Models. EJS.

Crossed Effect - Multigrid analysis

Model: $y_{ij} \sim N(\mu + a_i + b_j, \tau_e^{-1})$

Standard Gibbs Sampler:

1. Sample $\mu \sim p(\mu|a, b)$
2. Sample $a_i \sim p(a_i|\mu, b)$ for all i
3. Sample $b_j \sim p(b_j|\mu, a)$ for all j

Notation: $\bar{a} = \frac{1}{I}\sum_i a_i, \quad \bar{b} = \frac{1}{J}\sum_j b_j, \quad \delta a_i = a_i - \bar{a}, \quad \delta b_j = b_j - \bar{b}$

Theorem (Multigrid decomposition)

Let $(\mu, a, b)(t)$ be the Markov chain generated by the Gibbs Sampler. Then $(\mu, \bar{a}, \bar{b})(t)$, $\delta a(t)$ and $\delta b(t)$ are three independent Markov chains. [8]

Corollary: $T_{gibbs} = T(\mu, \bar{a}, \bar{b})$

[8] Papaspiliopoulos,Roberts&Z. Scalable Bayesian computation for crossed effect models. In preparation

Complexity of standard Gibbs for crossed effects

Corollary

$$T_{gibbs} = T(\mu, \bar{a}, \bar{b}) = 1 + \max\left\{\frac{J\tau_e}{\tau_a}, \frac{I\tau_e}{\tau_b}\right\} = \mathcal{O}(\max\{\#rows, \#columns\})$$

Thus $T_{gibbs} \geq \mathcal{O}(N^{1/2})$ and $Cost_{gibbs} = \mathcal{O}(N) \quad \Rightarrow \quad Cost_{gibbs} \geq \mathcal{O}(N^{3/2})$

K factors case

$$y_{i_1 \dots i_K} \sim N(\mu + a^{(1)}_{i_1} + \cdots + a^{(K)}_{i_K}, \tau_e^{-1}) \qquad i_k = 1, \dots, I_k;\ k = 1, \dots, K\,.$$

Theorem

Let $(\mu, a^{(1)}, \dots, a^{(K)})(t)$ *be the Markov chain generated by the Gibbs Sampler. Then*

1. $(\mu, \bar{a}^{(1)}, \dots, \bar{a}^{(K)})(t)$ *and* $(\delta a^{(1)}, \dots, \delta a^{(K)})(t)$ *are independent Markov chains.*
2. $T(\mu, \bar{a}^{(1)}, \dots, \bar{a}^{(K)})(t) \geq T(\delta a^{(1)}, \dots, \delta a^{(K)})(t)$

Corollary

$$T_{Gibbs} = 1 + \max_{k=1,\dots,K} \frac{N\tau_e}{I_k \tau_k} = \mathcal{O}\left(\frac{N}{\min_k I_k}\right) \geq \mathcal{O}\left(N^{1-1/K}\right)$$

Reparametrizations

If replace a_i with $\gamma_i = \mu + a_i$ or $\eta_j = \mu + b_j$ then

$$T_{a-centred} = \left(1 + \frac{\tau_a}{J\tau_e}\right)\left(1 + \frac{I\tau_e}{\tau_b}\right) = O(I)$$

$$T_{b-centred} = \left(1 + \frac{\tau_b}{I\tau_e}\right)\left(1 + \frac{J\tau_e}{\tau_a}\right) = O(J)$$

$\Rightarrow$ Reparametrizations do not solve the problem here
Things even worse for $K > 2$

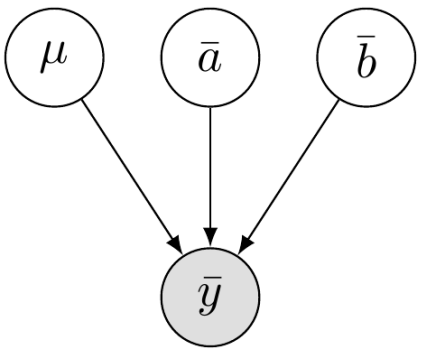
Alternative methodological trick to get $T_{mix} = O(1)$?

Slow mixing - Geometric intuition

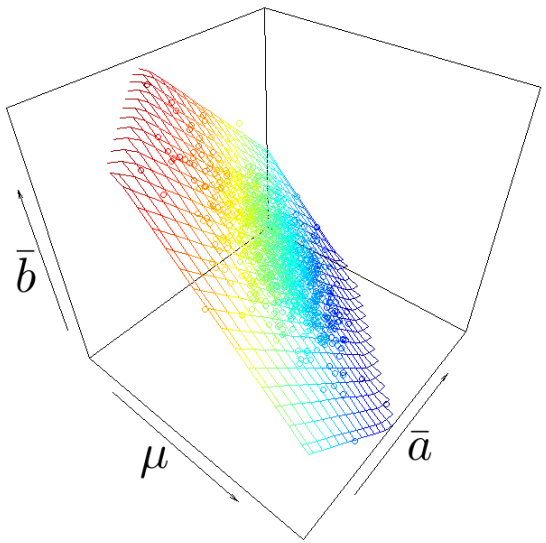
Crossed effects induce a strong negative correlation due to $\mu + \bar{a} + \bar{b} \approx \bar{y}$

Nested models induce positive correlation due to $\mu \approx \bar{\gamma} \approx \bar{\eta} \approx \bar{y}$

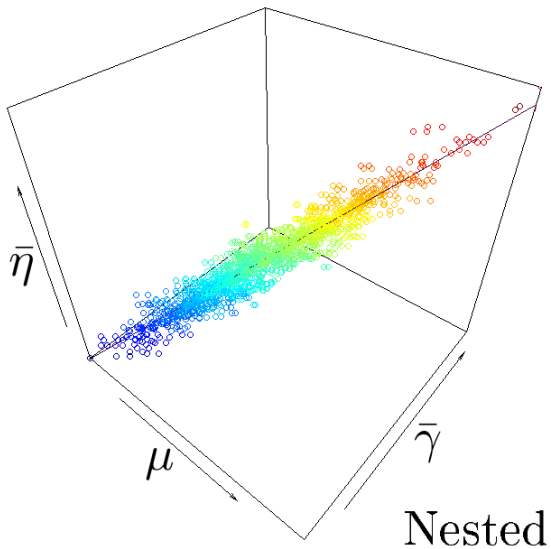
Negative correlation structure

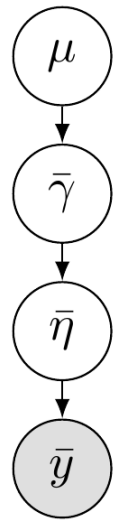


Crossed effects



Positive correlation structure





Nested effects

Crossed: as data increase posterior concentrates on an (hyper)plane of co-dimension 1 $\Rightarrow$ it is sufficient to collapse one variable to break correlation!

Crucially, while collapsing *a* or *b* is computationally expensive (large matrix inversions), collapsing μ is straightforward (one dimensional parameter)

Collapsed Gibbs Sampler

Collapsed Gibbs Sampler:

1. Sample $(\mu, a) \sim p(\mu, a|b)$
2. Sample $(\mu, b) \sim p(\mu, b|a)$

$\overset{equivalent}{\longleftrightarrow}$

1. Sample $\mu \sim p(\mu|b)$
2. Sample $a_i \sim p(a_i|\mu, b)$ for all i
3. Sample $\mu \sim p(\mu|a)$
4. Sample $b_j \sim p(b_j|\mu, a)$ for all j

The collapsed version has basically the same cost per iteration as the original Gibbs Sampler, but the mixing time is drastically different.

Theorem: for the crossed effect model under consideration, the collapsed Gibbs sampler produces iid samples from $\mu, a, b|y$.

$\Rightarrow T_{collapsed} = 1 \quad \Rightarrow$ Collapsed GS has complexity $\mathcal{O}(N)$

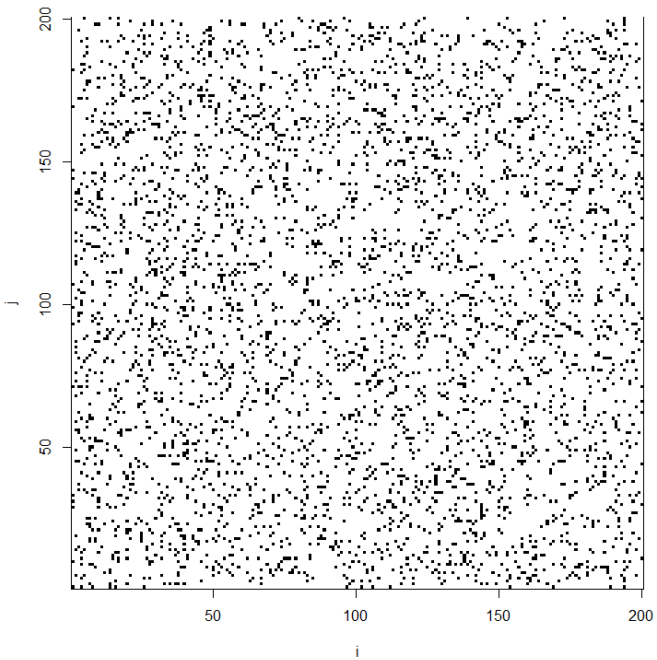
Introducing sparsity in the analysis

So far we assumed a full-matrix of observations
$\rightarrow$ potentially unrealistic simplification!
Can we provide theory that handles sparsity?

Model:

$$y_{ij} \sim N(\mu + a_i + b_j, \tau_e^{-1}) \qquad (i,j) \in S$$

with $S \subsetneq \{1,\dots,I\} \times \{1,\dots,J\}$ and sparsity level $\alpha = \frac{|S|}{IJ}$.



Balancedness assumption: S has constant row sums and column sums (each user sees the same number of movies, each movie seen by the same number of users)

Multigrid analysis for sparse crossed models

Theorem

Let $((\mu, a, b)(t))_{t=0}^{\infty}$ *be the Markov chain generated by the Gibbs Sampler. Then* $(\mu, \bar{a}, \bar{b})(t)$ *and* $(\delta a, \delta b)(t)$ *are two independent Markov chains. Moreover* $T(\mu, \bar{a}, \bar{b}) \geq T(\delta a, \delta b)$.

Corollary : $$T_{gibbs} = T(\mu, \bar{a}, \bar{b}) = 1 + \max\left\{\alpha J \frac{\tau_e}{\tau_a}, \alpha I \frac{\tau_e}{\tau_b}\right\}$$
$$\approx \max\{\#\text{obs.per row}, \#\text{obs.per col.}\}$$

NB: sparsity helps the Gibbs Sampler!
However T_{gibbs} can still grow with N. What can we say about collapsing μ?

Corollary : $$T_{\mu\text{-collapsed}} = T(\delta a, \delta b)$$

μ-collapsing and the residual process

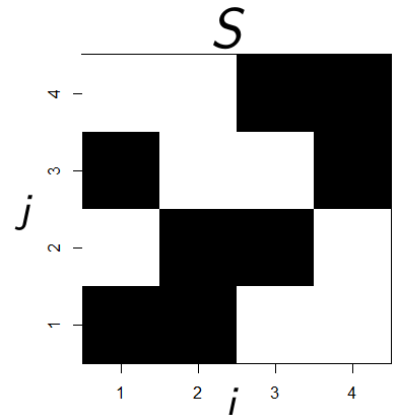
Theorem (Rate of residual process)

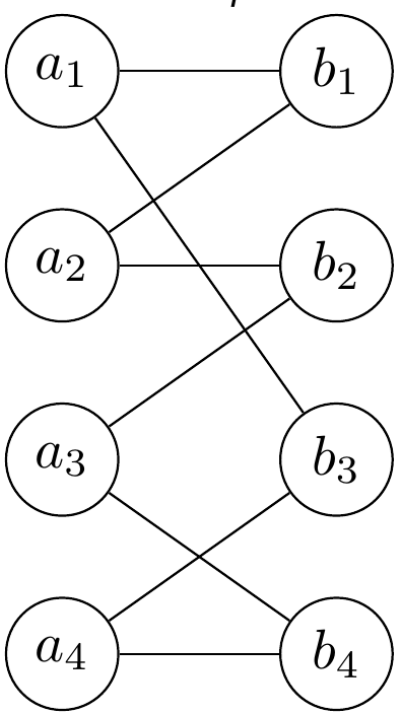
$$\rho_{(\delta a, \delta b)} = \frac{\alpha J \tau_e}{\alpha J \tau_e + \tau_a} \frac{\alpha I \tau_e}{\alpha I \tau_e + \tau_b} \rho_X$$

where X is the simple random walk on the bipartite graph with adjacency matrix S.

Corollary

$$T_{\mu\text{-collapsed}} = \frac{1}{1 - \rho_{(\delta a, \delta b)}}$$
$$\leq 1 + \min\left\{\alpha J \frac{\tau_e}{\tau_a}, \alpha I \frac{\tau_e}{\tau_b}, T_X\right\}$$



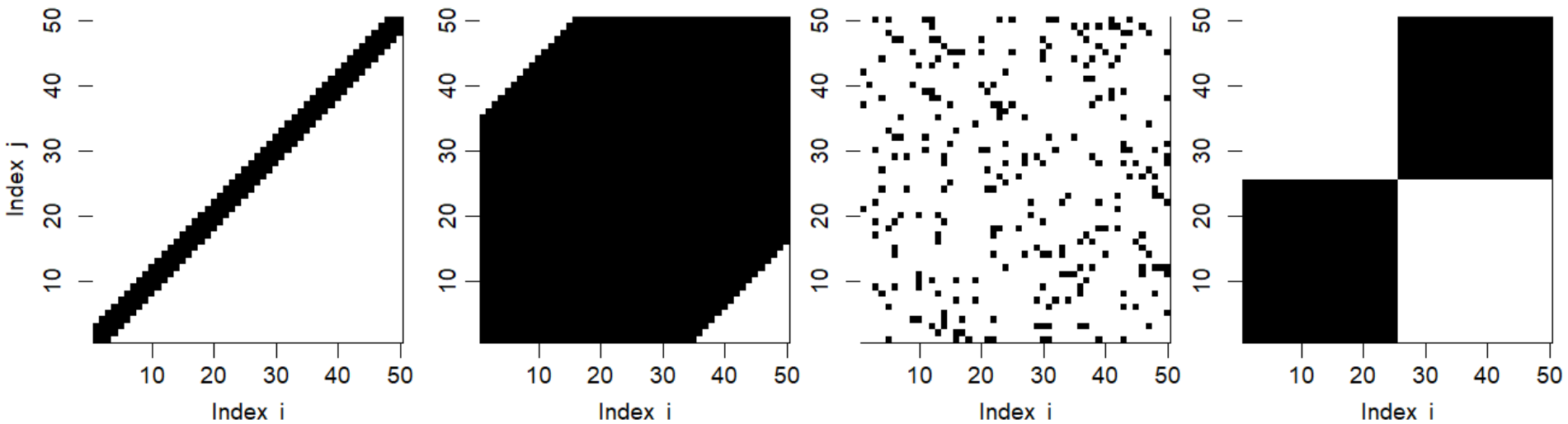


T_X mixing time of the auxiliary random walk on the bipartite graph

$$T_{gibbs} \approx \max\{\#\text{obs.per row}, \#\text{obs.per col.}\}$$
$$T_{\mu\text{-collapsed}} \approx \min\{\#\text{obs.per row}, \#\text{obs.per col.}, T_X\}$$

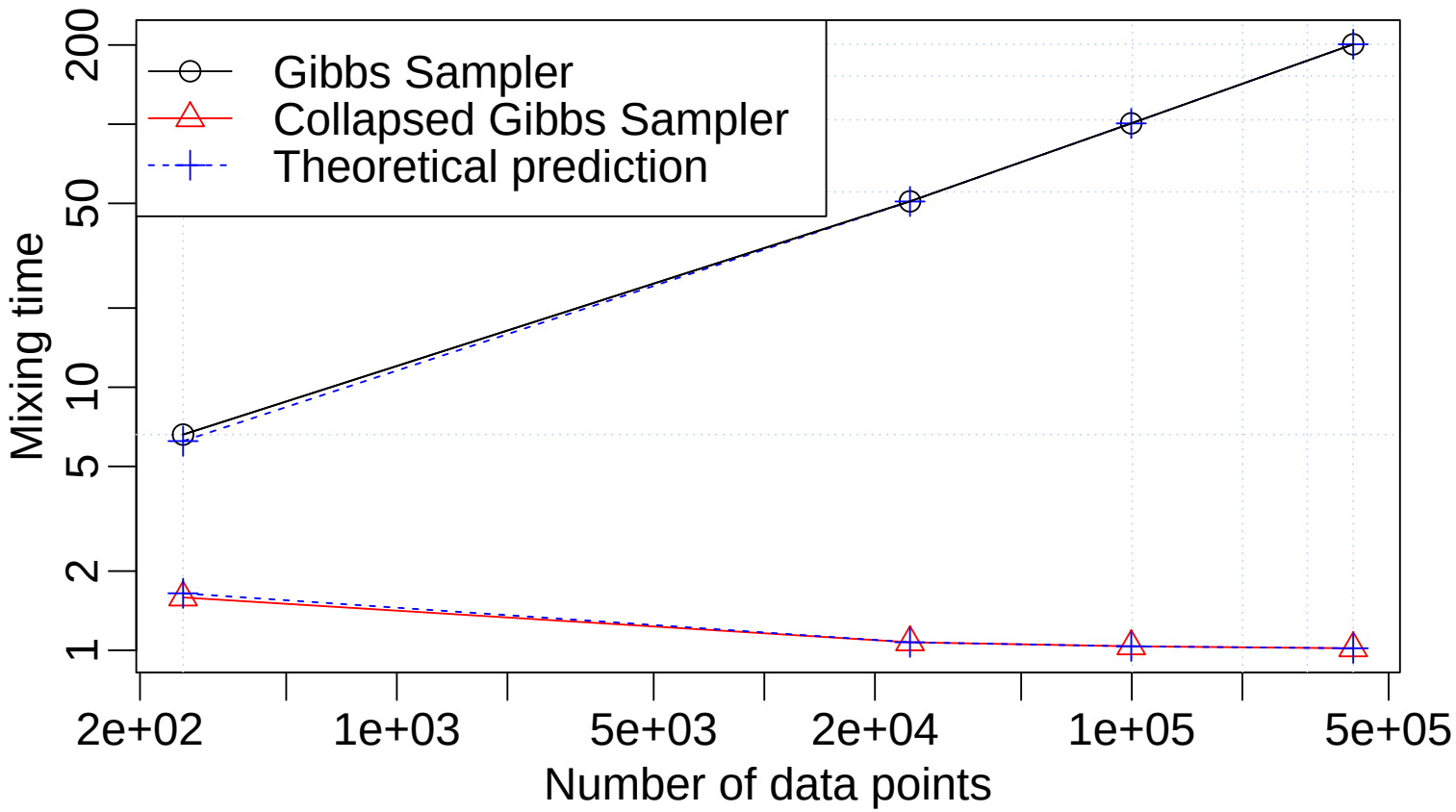
Crucially, as #obs.per row/col. grow T_X decreases, so *min* stays small.

Examples of graphs



Simulated data

$K = 2$, $I = J \to \infty$. Observe each y_{ij} with prob. 0.1 independently of the rest.



ETH Instructors evaluation dataset

Standard data-set available as *InstEval* from *lme4* R package
Collects university lecture evaluations by students at ETH Zurich.
$N = 73.421$ data-points, $K = 5$ factors, $(I_1, \ldots, I_5) = (2972, 1128, 4, 6, 14)$

Fixed variances

Factors included	T_{mix} (numerical) GS	T_{mix} (numerical) collapsed GS	T_{mix} (theory prediction) GS	T_{mix} (theory prediction) collapsed GS
1 and 2	68.9	7.8	66.1	8.3
1 and 5	5245.6	4.8	5245.4	5.0
all	36687	137.2	36711.5	

Table: Mixing times (computed numerically or "predicted" with theory)

ETH Instructors evaluation dataset

Unknown variances

Scheme	time per 1000 iter.	min(ESS)/ time
vanilla GS	13.2s	0.07
collapsed GS	14.2s	2.51
GS+PXDA	13.5s	0.06
cGS+PXDA	14.4s	2.96
HMC	1112.6s	0.08

Table: Numbers are averaged over 10 runs of 10000 iterations for each scheme, discarding the first 1000 samples as burn-in.

Collapsed version improves by 1-2 orders of magnitude over standard Gibbs or HMC.
For comparison *lme4* package took 40.9 seconds to fit the same model.
NB: times for various Gibbs Samplers correspond to basic R implementation.

Conclusions

Contributions

- Multigrid approach to analyze Gibbs Samplers in multilevel linear models
- Complexity statements and quantitative guidance on centering and collapsing
- $O(N)$ sampler for crossed effect models
- Neat connection between MCMC behavior and model's graphical structure

Missing to get a clearer picture

- Analyze case of unknown variances
- Quantify the impact of unbalancedness

Take-home message

- For large "random-effect" models Bayes with linear complexity is achievable
- Need to exploit models structure

Conclusions

More broadly

- Assess relevance of proposed methodology in non-gaussian cases?
- General theory and connections with design of experiments literature?
- Exploit collapsing trick in other contexts (e.g. probabilistic matrix factoriz.)

Arxiv preprints:

- G.Zanella & G.Roberts (2017) Analysis of the Gibbs Sampler for Gaussian hierarchical models via multigrid decomposition. Arxiv preprint.
- O.Papaspiliopoulos & G.Zanella (2017) A note on MCMC for nested multilevel regression models via belief propagation. Arxiv preprint.
- O.Papaspiliopoulos, G.Roberts & G.Zanella (2018) Scalable inference for crossed random effects models. Arxiv preprint.