Next-generation Gibbs-type Samplers: Combining Strategies to Boost Efficiency

Xiyun Jiao

Statistic Section
Department of Mathematics
Imperial College London

Joint work with David van Dyk, Roberto Trotta & Hikmatali Shariff

ICHASC Talk
Feb 02, 2016

Imperial College London

Outline

Imperial College London

Problem Setting

- **Goal**: sample from posterior distribution $p(\psi|Y)$ using Gibbs-type samplers.

- **Special case**: Data Augmentation (DA) Algorithm[1]
  $\psi = (Y_{\text{mis}}, \theta)$. DA algorithm proceeds as:

$$[Y_{\text{mis}}|\theta'] \longleftrightarrow [\theta|Y_{\text{mis}}].$$

  Stationary distribution: $p(Y_{\text{mis}}, \theta|Y)$.
  (Or $[\psi_1|\psi_2'] \longleftrightarrow [\psi_2|\psi_1]$ to sample $p(\psi_1, \psi_2|Y)$.)

DA algorithm and Gibbs samplers are easy to implement, but...

Converge slowly!

[1] Tanner, M. A. and Wong, W. H. (1987)

Imperial College London

Problem Setting

- **Goal**: sample from posterior distribution $p(\psi|Y)$ using Gibbs-type samplers.

- **Special case**: Data Augmentation (DA) Algorithm[1]

  $\psi = (Y_{\rm mis}, \theta)$. DA algorithm proceeds as:

$$[Y_{\rm mis}|\theta'] \longleftrightarrow [\theta|Y_{\rm mis}].$$

  Stationary distribution: $p(Y_{\rm mis}, \theta|Y)$.

  (Or $[\psi_1|\psi_2'] \longleftrightarrow [\psi_2|\psi_1]$ to sample $p(\psi_1, \psi_2|Y)$.)

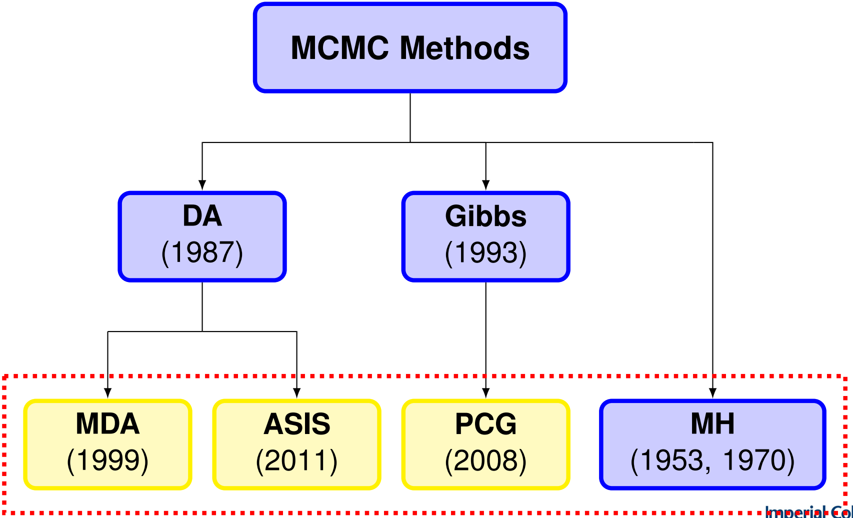
DA algorithm and Gibbs samplers are easy to implement, but...

Converge slowly!

[1] Tanner, M. A. and Wong, W. H. (1987)

Imperial College London

Existing Gibbs-type Algorithms



Marginal Data Augmentation

Marginal Data Augmentation (MDA)[2]

- MDA introduces a working parameter α into $p(Y, Y_{\mathrm{mis}}|\theta)$ via Y_{mis} [e.g., $\tilde{Y}_{\mathrm{mis}} = \mathcal{F}_\alpha(Y_{\mathrm{mis}})$], s.t.,

$$\int p(\tilde{Y}_{\mathrm{mis}}, Y|\theta, \alpha)\mathrm{d}\tilde{Y}_{\mathrm{mis}} = p(Y|\theta).$$

- Standard DA: $[\tilde{Y}_{\mathrm{mis}}|\theta', \alpha = \alpha_0] \longleftrightarrow [\theta|\tilde{Y}_{\mathrm{mis}}, \alpha = \alpha_0]$.
- If the prior distribution of α is proper, MDA proceeds as:

$$[\alpha^\star, \tilde{Y}_{\mathrm{mis}}|\theta'] \longleftrightarrow [\alpha, \theta|\tilde{Y}_{\mathrm{mis}}].$$

- MDA improves convergence by increasing variability in augmented data and reducing augmented information.

[2] Meng, X.-L. and van Dyk, D. A. (1999); Liu, J. S. and Wu, Y. N. (1999)

Imperial College London

Ancillarity-Sufficiency Interweaving Strategy

Ancillarity-Sufficiency Interweaving Strategy (ASIS)[3]

- ASIS considers a pair of special DA schemes:
  - Sufficient augmentation $Y_{\mathrm{mis,S}}$: $p(Y|Y_{\mathrm{mis,S}},\theta)$ is free of θ.
  - Ancillary augmentation $Y_{\mathrm{mis,A}}$: $p(Y_{\mathrm{mis,A}}|\theta)$ is free of θ.
- Given θ, $Y_{\mathrm{mis,A}} = \mathcal{F}_\theta(Y_{\mathrm{mis,S}})$. ASIS proceeds as

Interweave $[\theta|Y_{\mathrm{mis,S}}]$ into DA algorithm w.r.t. $Y_{\mathrm{mis,A}}$

$$\Downarrow$$

$$[Y_{\mathrm{mis,S}}|\theta'] \rightarrow \boxed{[\theta^\star|Y_{\mathrm{mis,S}}] \rightarrow [Y_{\mathrm{mis,A}}|Y_{\mathrm{mis,S}},\theta^\star]} \rightarrow [\theta|Y_{\mathrm{mis,A}}]$$

$$\Updownarrow$$

$$[Y_{\mathrm{mis,S}}|\theta'] \rightarrow \boxed{[Y_{\mathrm{mis,A}}|Y_{\mathrm{mis,S}}]} \rightarrow [\theta|Y_{\mathrm{mis,A}}]$$

- ASIS obtains more efficiency by taking advantage of the "beauty-and-beast" feature of two parent DA algorithms.

[3] Yu, Y. and Meng, X.-L. (2011)

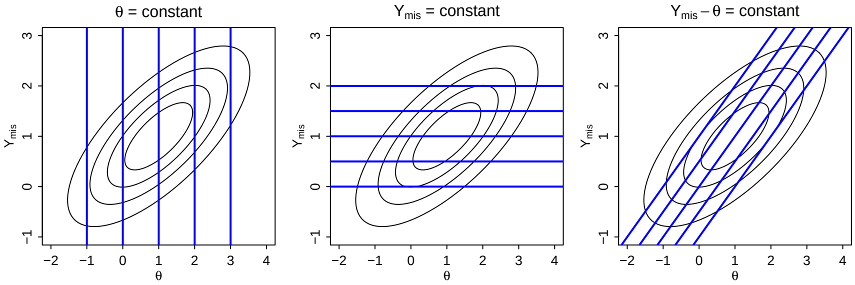
Understanding ASIS

- Model:

$$Y|(Y_{\text{mis}}, \theta) \sim \mathrm{N}(Y_{\text{mis}}, 1),\ Y_{\text{mis}}|\theta \sim \mathrm{N}(\theta, V),\ p(\theta) \propto 1.$$

- ASIS: $Y_{\text{mis,S}} = Y_{\text{mis}},\ Y_{\text{mis,A}} = Y_{\text{mis}} - \theta.$

$$[Y_{\text{mis,S}}|\theta'] \to [\theta^\star|Y_{\text{mis,S}}] \to [Y_{\text{mis,A}}|Y_{\text{mis,S}}, \theta^\star] \to [\theta|Y_{\text{mis,A}}]$$



More directions: efficient and easy to implement.

Imperial College London

Partially Collapsed Gibbs Sampling

Partially Collapsed Gibbs (PCG)[4]

- Model Reduction: PCG reduces conditioning of Gibbs. It replaces some conditional distributions of a Gibbs sampler with conditionals of marginal distributions of the target.

- PCG improves convergence by increasing variance and jump size of conditional distributions.

- Three stages: *Marginalization*, *permutation*, *trimming*.
  - Tools to transform a Gibbs sampler into a PCG one.
  - Maintain the target stationary distribution.

[4] van Dyk, D. A. and Park, T. (2008)

Imperial College London

Examples of PCG Sampling

Example. $\psi = (\psi_1, \psi_2, \psi_3, \psi_4)$; Sample from $p(\psi|Y)$.

Gibbs	PCG I	PCG II
$p(\psi_1\|\psi_2', \psi_3', \psi_4')$	$p(\psi_1\|\psi_2', \psi_3', \psi_4')$	$p(\psi_1\|\psi_2', \psi_4')$
$p(\psi_2\|\psi_1, \psi_3', \psi_4')$	$p(\psi_2, \psi_3\|\psi_1, \psi_4')$	$p(\psi_2, \psi_3\|\psi_1, \psi_4')$
$p(\psi_3\|\psi_1, \psi_2, \psi_4')$	$p(\psi_4\|\psi_1, \psi_2, \psi_3)$	$p(\psi_4\|\psi_1, \psi_2, \psi_3)$
$p(\psi_4\|\psi_1, \psi_2, \psi_3)$		

- Special cases: blocked and collapsed Gibbs, e.g., PCG I.
- More interestingly, a PCG sampler consists of *incompatible conditional distributions*, e.g., PCG II. Modifying the order of steps of PCG II may alter its stationary distribution.

Three Stages to Derive a PCG Sampler

(a) Gibbs

$p(\psi_1|\psi_2', \psi_3', \psi_4')$
$p(\psi_2|\psi_1, \psi_3', \psi_4')$
$p(\psi_3|\psi_1, \psi_2, \psi_4')$
$p(\psi_4|\psi_1, \psi_2, \psi_3)$

(b) Marginalize

$p(\psi_1, \psi_3^\star|\psi_2', \psi_4')$
$p(\psi_2, \psi_3^\star|\psi_1, \psi_4')$
$p(\psi_3|\psi_1, \psi_2, \psi_4')$
$p(\psi_4|\psi_1, \psi_2, \psi_3)$

(c) Permute

$p(\psi_1, \psi_3^\star|\psi_2', \psi_4')$
$p(\psi_2, \psi_3^\star|\psi_1, \psi_4')$
$p(\psi_3|\psi_1, \psi_2, \psi_4')$
$p(\psi_4|\psi_1, \psi_2, \psi_3)$

(d) Trim [PCG II]

$p(\psi_1|\psi_2', \psi_4')$
$p(\psi_2, \psi_3|\psi_1, \psi_4')$
$p(\psi_4|\psi_1, \psi_2, \psi_3)$

"$\star$"—Intermediate Draws

Imperial College London

Outline

Factor Analysis Model

- **Model**

$$Y_i \sim \mathrm{N}_p\left[\beta Z_i, \Sigma = \mathrm{Diag}(\sigma_1^2, \ldots, \sigma_p^2)\right], \text{ for } i = 1, \ldots, n.$$

  - Y_i —p-vector of the ith observation;
    Z_i —q-vector of factors; $Z_i \sim \mathrm{N}_q(0, I)$; $q \ll p$;
    $\beta_{kh} = 0$ and $\beta_{kk} > 0$, $h = k + 1, \ldots, q$ and $k = 1, \ldots, q$.
  - Priors: $p(\beta) \propto 1$; $\sigma_j^2 \sim$ Inv-Gamma$(0.01, 0.01)$, $j = 1, \ldots, p$.

- **Simulation Study**
  - Set $p = 6$, $q = 2$, and $n = 100$.
  - $\sigma_j^2 \sim$ Inv-Gamma$(1, 0.5)$, $(j = 1, \ldots, 6)$;
    $\beta_{hj} \sim \mathrm{N}(0, 3^2)$, $(h = 1, 2; j = 1, \ldots, 6)$.

- **Goal**

  Sample from the posterior distribution of Z, β and Σ.

Imperial College London

Samplers for Factor Analysis

- **Standard Gibbs sampler**:

$$\left[Z|\beta', \Sigma'\right] \longrightarrow \left[\sigma_j^2|Z, \beta'\right]_{j=1}^{6} \longrightarrow [\beta|Z, \Sigma].$$

Pros: Easy to implement.
Cons: The convergence of both β and Σ is poor.

- **MH within PCG sampler**: Sampling σ_1^2, σ_2^2, σ_3^2 and σ_4^2 without conditioning on Z. This is facilitated by MH.

Pros: Effective in improving the convergence of Σ.
Cons: Little effect on β.

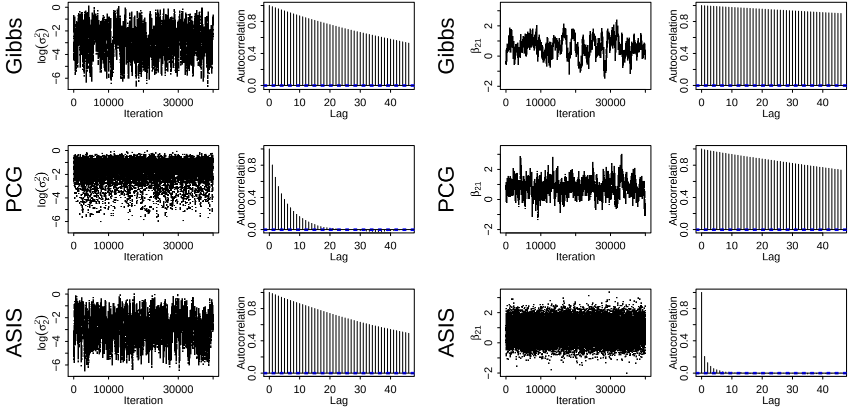
- **ASIS sampler**: Given Σ, for β, $Y_{\mathrm{mis,A}}$: Z_i; $Y_{\mathrm{mis,S}}$: $W_i = \beta Z_i$.

Pros: Effective in improving the convergence of β.
Cons: Hard to derive $Y_{\mathrm{mis,A}}$ and $Y_{\mathrm{mis,S}}$ for both β and Σ.

Imperial College London

Convergence of Gibbs, MH within PCG & ASIS

For each sampler, run 50,000 iterations with burn-in of 10,000.



Imperial College London

Outline

Imperial College London

Solution: Combining Strategies into One Sampler!

Cannot Sample Conditionals?

- Embed Metropolis-Hastings (MH) into Gibbs[5]—standard.
- Embed MH into PCG[6]—subtle implementation!

Further Improvement in Convergence

- Several parameters converge slowly—a strategy is efficient for one parameter, but has little effect on others; Another strategy has opposite effect. By combining, we improve all.
- One strategy alone is useful for all parameters—prefer to use a combination, as long as efficiency gain exceeds added computational expense.

$$1 + 1 > 2$$

[5] Gilks et al. (1995)

[6] van Dyk, D. A. and Jiao, X. (2015)

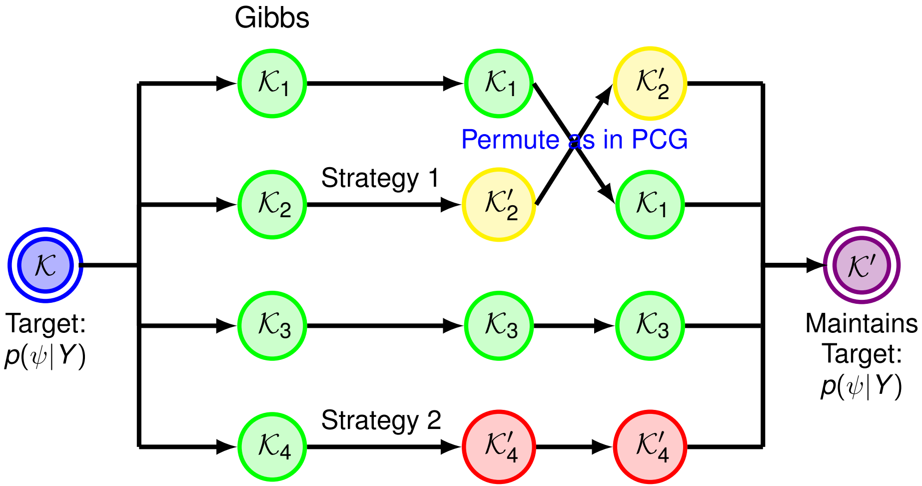
Algorithm Description

To sample from the posterior $p(\psi|Y)$, partition the unknown variable ψ into N components, i.e., $\psi = (\psi_1, \ldots, \psi_N)$.

- Start with Gibbs sampler: transition kernel $\mathcal{K} = \prod_{j=1}^{N} \mathcal{K}_j$ (stationary distribution of each $\mathcal{K}_j$ is a conditional of the target, $p(\psi|Y)$);
- Use one acceleration strategy or a combination of multiple strategies: replace $\mathcal{K}_j$ with $\mathcal{K}'_j$, for some j ($\mathcal{K}'_j$ may not have the target stationary distribution);
- Assuming irreducibility, permute the order of component kernels to guarantee that the new joint kernel $\mathcal{K}'$ maintains $p(\psi|Y)$ as its stationary distribution. (Principle: if input to $\mathcal{K}'_1$ follows the target distribution, the last component kernel should produce a draw from $p(\psi|Y)$.)

Imperial College London

Illustration Example ($N = 4$)



Transition Kernels for PCG, MDA & ASIS

For simplicity, set $N=2$, and suppress Y.

- **PCG**: sample ψ_1 without conditioning on ψ_2;
  Replace $\mathcal{K}_1$ with $\mathcal{K}_1' = p(\psi_1)$.
- **MDA**: with $\psi_1 = Y_{\text{mis}}$, set $\tilde{\psi}_1 = \mathcal{F}_\alpha(\psi_1)$ (improper prior for α);
  Replace $\mathcal{K}_1$ and $\mathcal{K}_2$ with
  $(\mathcal{K}_1, \mathcal{K}_2)' = \int p(\tilde{\psi}_1|\alpha', \psi_2')p(\alpha, \psi_1, \psi_2|\tilde{\psi}_1)\mathrm{d}\alpha\mathrm{d}\tilde{\psi}_1$.
- **ASIS**: with $\psi_1 = Y_{\text{mis,S}}$, set $\tilde{\psi}_1 = \mathcal{F}_{\psi_2}(\psi_1) = Y_{\text{mis,A}}$;
  Replace $\mathcal{K}_1$ and $\mathcal{K}_2$ with
  $(\mathcal{K}_1, \mathcal{K}_2)' = \int\int p(\psi_1^\star|\psi_2')p(\tilde{\psi}_1|\psi_1^\star)p(\psi_1, \psi_2|\tilde{\psi}_1)\mathrm{d}\psi_1^\star\mathrm{d}\tilde{\psi}_1$.

$\mathcal{K}'$ introduced by each of MDA, ASIS and PCG has smaller *cyclic-permutation bound* than $\mathcal{K}$[7].

Imperial College London

[7] van Dyk, D. A. and Park, T. (2008)

Outline

1. Algorithm Review
2. Motivation
3. Combining Strategies
4. Examples
5. Conclusion

Imperial College London

Factor Analysis (Cont.)

Gibbs

$p(Z_i | Y, \beta', \Sigma')_{i=1}^{100}$

$p(\sigma_j^2 | Y, Z, \beta')_{j=1}^{6}$

$p(\beta_j | Y, Z, \Sigma)_{j=1}^{6}$

MH within PCG

$\mathcal{M}(\sigma_j^2 | Y, \sigma_{<j}^2, \sigma_{\geq j}^{2\prime}, \beta')_{j=1}^{4}$

$p(Z_i | Y, \sigma_{\leq 4}^2, \sigma_{\geq 5}^{2\prime}, \beta')_{i=1}^{100}$

$p(\sigma_j^2 | Y, Z, \beta')_{j=5}^{6}$

$p(\beta_j | Y, Z, \Sigma)_{j=1}^{6}$

ASIS

$p(Z_i^\star | Y, \beta', \Sigma')_{i=1}^{100}$

$p(\sigma_j^2 | Y, Z^\star, \beta')_{j=1}^{6}$

$p(\beta_j^\star | Y, Z^\star, \Sigma)_{j=1}^{6}$

$\{W_i = \beta^\star Z_i^\star\}_{i=1}^{100}$

$p(\beta, Z | Y, W, \Sigma)$

MH within PCG+ASIS

$\mathcal{M}(\sigma_j^2 | Y, \sigma_{<j}^2, \sigma_{\geq j}^{2\prime}, \beta')_{j=1}^{4}$

$p(Z_i^\star | Y, \sigma_{\leq 4}^2, \sigma_{\geq 5}^{2\prime}, \beta')_{i=1}^{100}$

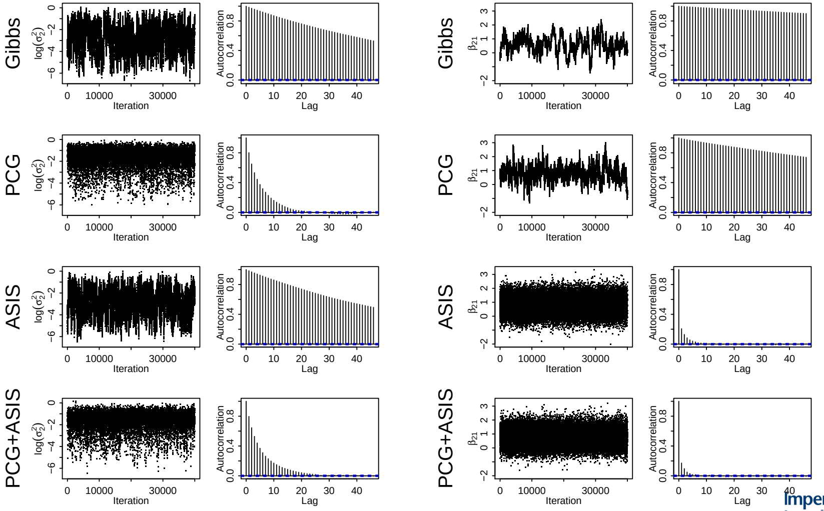
$p(\sigma_j^2 | Y, Z^\star, \beta')_{j=5}^{6}$

$p(\beta_j^\star | Y, Z^\star, \Sigma)_{j=1}^{6}$

$\{W_i = \beta^\star Z_i^\star\}_{i=1}^{100}$

$p(\beta, Z | Y, W, \Sigma)$

Convergence Results of Factor Analysis Model



Imperial College London

Effective Sample Size (ESS) per Second

The larger the ESS/sec, the more efficient the algorithm.

	Gibbs	PCG	ASIS	PCG + ASIS
$\log(\sigma_2^2)$	0.141	1.776	0.137	1.500
β_{21}	0.022	0.062	8.162	9.376

Imperial College London

Astrophysics Background

- Physics Nobel Prize (2011): discovery of acceleration of expansion of the universe.
- The acceleration is attributed to existence of dark energy.
- Type Ia supernova (SNIa) observations: critical to quantify characteristics of dark energy.

  Mass $>$ "Chandrasekhar threshold" ($1.44\ M_{\odot}$) $\Longrightarrow$ SN explosion.

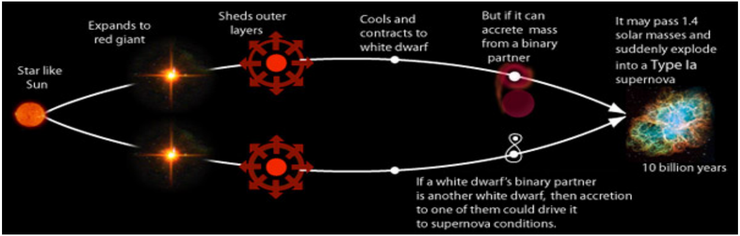


Image credit: http://hyperphysics.phy-astr.gsu.edu/hbase/astro/snovcn.html

Imperial College London

“Standardizable Candles”

Common history $\Longrightarrow$ similar absolute magnitudes for SNIa, i.e.,

$$M_i \sim \mathrm{N}(M_0, \sigma^2_{\mathrm{int}})$$

$\Longrightarrow$ SNIa are “standardizable candles”.

Phillips corrections:

$$M_i = M_i^{\epsilon} - \alpha x_i + \beta c_i,\ M_i^{\epsilon} \sim \mathrm{N}(M_0, \sigma^2_{\epsilon});$$

x_i—stretch correction, c_i—color correction,

$$\sigma^2_{\epsilon} \leq \sigma^2_{\mathrm{int}}$$

Imperial College London

Distance Modulus

Apparent Magnitude $-$ Absolute Magnitude $=$ Distance Modulus:

$$m_B - M = \mu = 5\log_{10}[\mathrm{distance}(\mathrm{Mpc})] + 25.$$

- Nearby SN: $\mathrm{distance} = zc/H_0$;
- Distant SN: $\mu = \mu(z, \Omega_m, \Omega_\Lambda, H_0)$;
  - c—speed of light
  - H_0—Hubble constant
  - z—redshift
  - Ω_m—total matter density
  - Ω_Λ—dark energy density

Imperial College London

Bayesian Hierarchical Model[8]

- **Level 1**: Errors-in-variables regression:

$$m_{Bi} = \mu_i + M_i^\epsilon - \alpha x_i + \beta c_i;$$

$$\begin{pmatrix} \hat{c}_i \\ \hat{x}_i \\ \hat{m}_{Bi} \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} c_i \\ x_i \\ m_{Bi} \end{pmatrix}, \hat{C}_i \right], \; i = 1, \ldots, n.$$

- **Level 2**:
  $M_i^\epsilon \sim \mathrm{N}(M_0, \sigma_\epsilon^2);\; x_i \sim \mathrm{N}(x_0, R_x^2);\; c_i \sim \mathrm{N}(c_0, R_c^2).$
- **Priors**:
  Gaussian for M_0, x_0, c_0;
  Uniform for Ω_m, Ω_Λ, α, β, $\log(R_x)$, $\log(R_c)$, $\log(\sigma_\epsilon)$.
  z and H_0 fixed.

[8] March et al. (2011)

Notation and Data

Notation:

- $\Omega = (\Omega_m, \Omega_\Lambda)$, $S = (\sigma_\epsilon^2, R_x^2, R_c^2)$;
- $X_{(3n\times 1)}$—$(c_1, x_1, M_1^\epsilon, \ldots, c_n, x_n, M_n^\epsilon)$;
- $\xi_{(3\times 1)}$—(c_0, x_0, M_0);
- $L_{(3n\times 1)}$—$(0, 0, \mu_1, \ldots, 0, 0, \mu_n)$;
- $T_{(3\times 3)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \beta & -\alpha & 1 \end{bmatrix}$, and $A_{(3n\times 3n)} = \mathrm{Diag}(T, \ldots, T)$.

Data: A sample of 288 SNIa compiled by Kessler et al. (2009).

Imperial College London

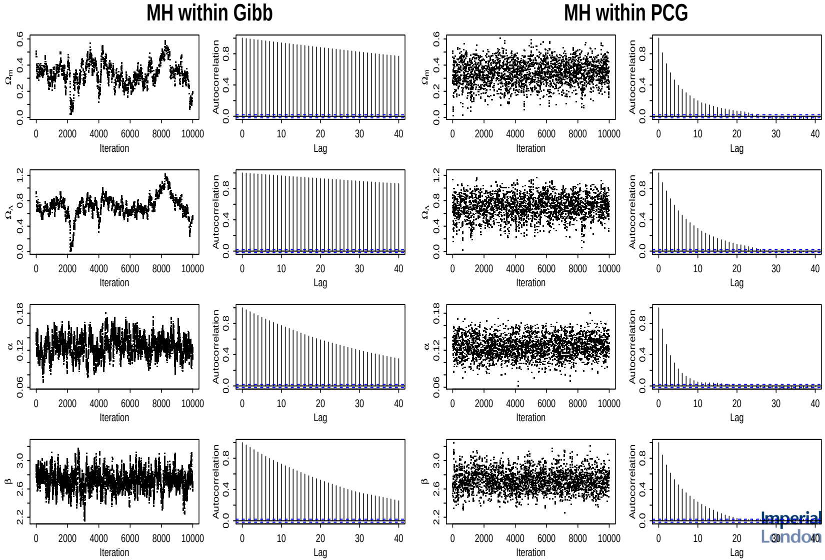
Algorithms for Cosmological Hierarchical Model

- **MH within Gibbs sampler**:
  - Sample each of (X, ξ), Ω, (α, β) and S from their complete conditionals.
  - Update of Ω needs MH.
- **MH within PCG sampler**:
  - Sample Ω and (α, β) without conditioning on (X, ξ).
  - Updates of both Ω and (α, β) require MH.
- **ASIS sampler**: Given ξ and S, for both Ω and (α, β),
  $Y_{\mathrm{mis,S}} = AX + L$; $Y_{\mathrm{mis,A}} = X$.
- **MH within PCG+ASIS sampler**:
  - Given (α, β), sample Ω by MH within PCG;
  - Given Ω, sample (α, β) by ASIS.

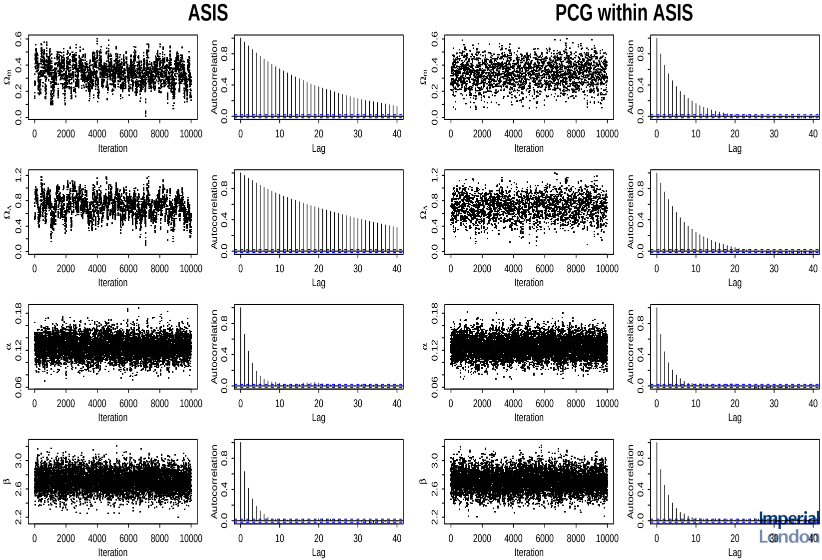
For each sampler, run 11,000 iterations with a burn-in of 1,000.

Imperial College London

Convergence Results of Gibbs and PCG



Convergence Results of ASIS and Combining



Effective Sample Size (ESS) per Second

The larger the ESS/sec, the more efficient the algorithm.

	Gibbs	PCG	ASIS	PCG+ASIS
Ω_m	0.00166	0.0302	0.0103	0.0392
Ω_Λ	0.000997	0.0232	0.00571	0.0282
α	0.00712	0.0556	0.0787	0.0826
β	0.00874	0.0264	0.0830	0.0733

Imperial College London

Outline

Imperial College London

Conclusion

- **Summary**
  - Combining different accelerating strategies into one sampler is useful to produce more efficiency in terms of convergence properties.
  - The hierarchical Gaussian model reflects the underlying physical understanding of supernova cosmology.
- **Future Work**
  - More numerical examples to illustrate the new algorithm.
  - Extend the combining strategy to a more general framework (surrogate distribution) to further improve convergence.