Embedding Supernova Cosmology into a Bayesian Hierarchical Model

Xiyun Jiao

Statistic Section
Department of Mathematics
Imperial College London

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

ICHASC Talk
Jan 27, 2015

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]

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

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

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

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

Converge slowly!

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

Problem Setting

- **Goal**: Sample from posterior distribution $p(\psi|Y)$ using Gibbs-type samplers.
- **Special case**: Data Augmentation (DA) Algorithm[1]

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

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

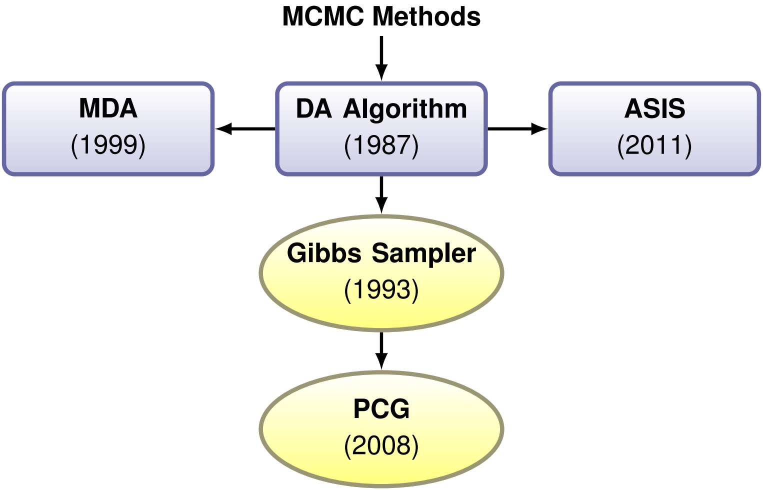
  Stationary distribution: $p(Y_{\text{mis}}, \theta|Y)$.

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

Converge slowly!

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

Algorithm Review



Blue Rectangle—Expand Paramter Space;
Yellow Ellipse—Change Conditioning Strategy

Imperial College London

Marginal Data Augmentation

Marginal Data Augmentation (MDA)[2]

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

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

- If the prior distribution of α is proper, MDA proceeds as:

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

- 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)

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)

Imperial College London

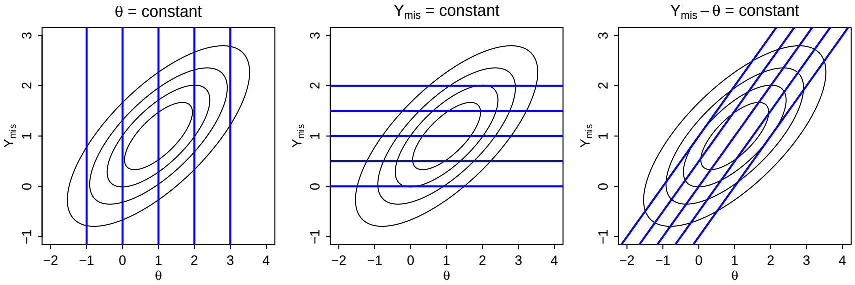
Understanding ASIS

- 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)

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.

Imperial College London

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

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

(c) Permute

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

(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

Imperial College London

Combining Different 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 gained efficiency exceeds extra computational expense.

[5] Gilks et al. (1995)

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

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.

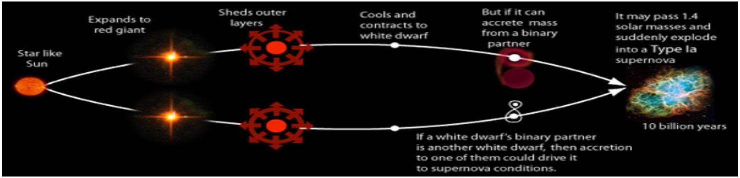


Image credit: http://hyperphysics.phy-astr.gsu.edu/hbase/astro/snovcn.html

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,

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

Distance Modulus

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

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

- 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[7]

- **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.

[7] March et al. (2011)

Imperial College London

Notation and Data

Notation

- $X_{(3n\times 1)}$—$(c_1, x_1, M_1^\epsilon, \ldots, c_n, x_n, M_n^\epsilon)$;
- $b_{(3\times 1)}$—(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).

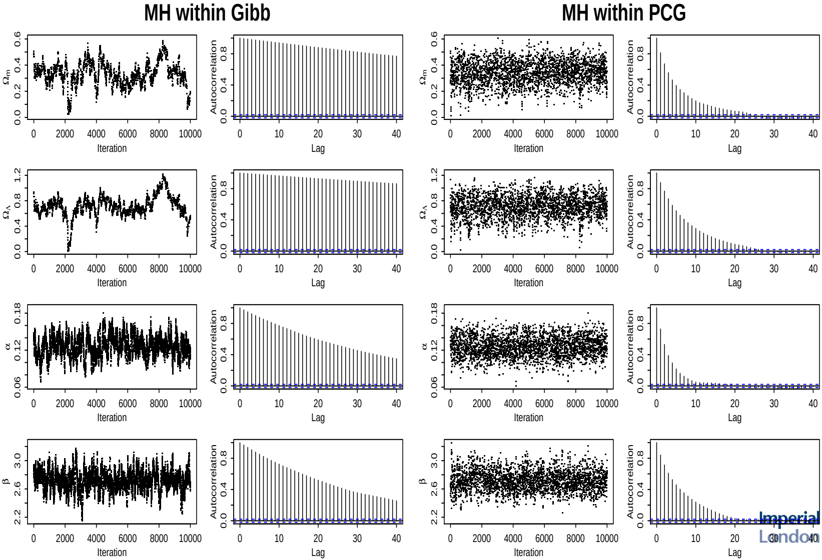
Algorithms for Cosmological Herarchical Model

- **MH within Gibbs sampler**: Update of $(\Omega_m, \Omega_\Lambda)$ needs MH.
- **MH within PCG sampler**:
  - Sample $(\Omega_m, \Omega_\Lambda)$ and (α, β) without conditioning on (X, b).
  - Updates of both $(\Omega_m, \Omega_\Lambda)$ and (α, β) need MH.
- **ASIS sampler**: $Y_{\mathrm{mis},S}$ for $(\Omega_m, \Omega_\Lambda)$ and (α, β): $AX + L$;
  $Y_{\mathrm{mis},A}$ for $(\Omega_m, \Omega_\Lambda)$ and (α, β): X.
- **MH within PCG+ASIS sampler**:
  - Given (α, β), sample $(\Omega_m, \Omega_\Lambda)$ with MH within PCG;
  - Given $(\Omega_m, \Omega_\Lambda)$, sample (α, β) with ASIS.

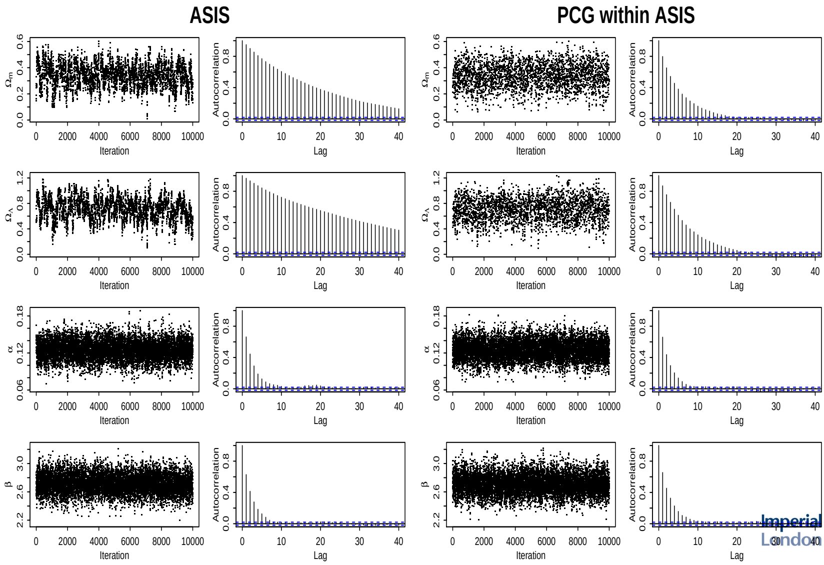
For each sampler, run 11,000 iterations with a burn-in of 1,000.

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

Factor Analysis Model

- **Model**

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

  - Y_i —$(1 \times p)$ vector of the ith observation;
    Z_i —$(1 \times q)$ vector of factors; $Z_i|\beta \sim \mathrm{N}(0, I)$; $q < p$.
  - β and Σ —unknown parameters.
    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$(1, 0.5)$, $(j = 1, \ldots, 6)$;
    $\beta_{hj} \sim \mathrm{N}(0, 3^2)$, $(h = 1, 2; j = 1, \ldots, 6)$.

Algorithms for Factor Analysis

- **Standard Gibbs sampler**:

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

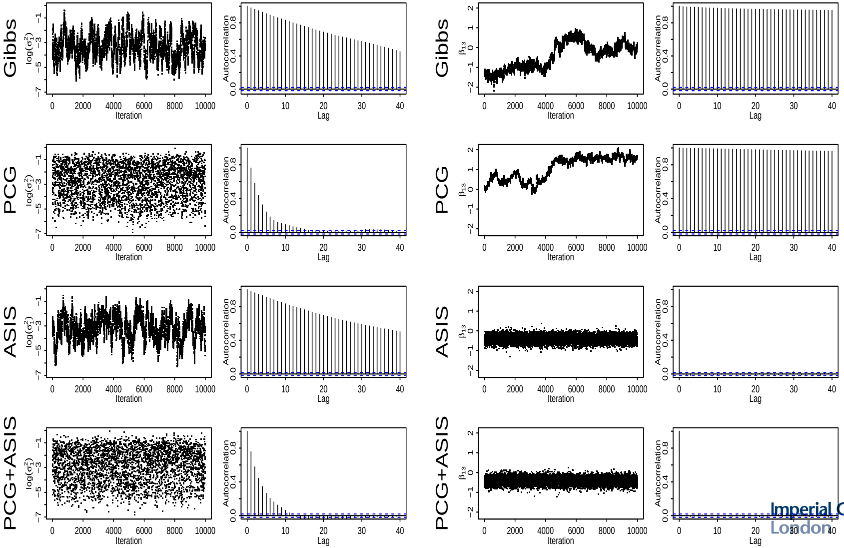
- **MH within PCG sampler**: sampling σ_1^2, σ_3^2 and σ_4^2 without conditioning on Z. This should be facilitated by MH.

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

- **MH within PCG+ASIS sampler**:
  - Given β, update Σ with MH within PCG;
  - Given Σ, update β with ASIS.

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

Convergence Results of Factor Analysis Model



Effective Sample Size (ESS) per Second

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

	Gibbs	PCG	ASIS	PCG + ASIS
$\log(\sigma_1^2)$	0.18	2.17	0.15	1.91
β_{13}	0.0087	0.0090	17.54	15.37

Imperial College London

Outline

Imperial College London

Bivariate Surrogate Distribution

Target distribution: $p(\psi_1, \psi_2)$.

Surrogate distribution: $\pi(\psi_1, \psi_2)$.

$\pi(\psi_1) = p(\psi_1)$, $\pi(\psi_2) = p(\psi_2)$; The correlation between ψ_1 and ψ_2 is lower for π than for p.

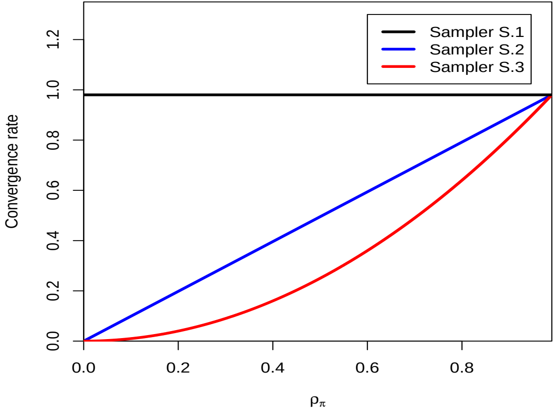
Sampler S.1	**Sampler S.2**	**Sampler S.3**
$p(\psi_1 \vert \psi_2')$	$\pi(\psi_1 \vert \psi_2')$	$\pi(\psi_1 \vert \psi_2')$
$p(\psi_2 \vert \psi_1)$	$p(\psi_2 \vert \psi_1)$	$\pi(\psi_2 \vert \psi_1)$

- Stationary distribution of Samplers S.1 and S.2: $p(\psi_1, \psi_2)$. Stationary distribution of Sampler S.3: $\pi(\psi_1, \psi_2)$.
- Condition for Sampler S.2 maintaining the target: $\pi(\psi_1) = p(\psi_1)$, $\pi(\psi_2) = p(\psi_2)$; Step order is fixed.

Comparison of Samplers S.1–S.3

Example.

$$p(\psi_1, \psi_2) : \mathrm{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.99 \\ 0.99 & 1 \end{pmatrix}\right]; \pi(\psi_1, \psi_2) : \mathrm{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_\pi \\ \rho_\pi & 1 \end{pmatrix}\right].$$



Ways to Derive Surrogate Distributions

- **ASIS**: $[Y_{\mathrm{mis},S}|\theta'] \to [Y_{\mathrm{mis},A}|Y_{\mathrm{mis},S}] \to [\theta|Y_{\mathrm{mis},A}]$.
  $\pi(\theta|Y_{\mathrm{mis},S}) = \int p(Y_{\mathrm{mis},A}|Y_{\mathrm{mis},S})p(\theta|Y_{\mathrm{mis},A})\mathrm{d}Y_{\mathrm{mis},A}$;
  $\pi(\theta, Y_{\mathrm{mis},S}) = \pi(\theta|Y_{\mathrm{mis},S})p(Y_{\mathrm{mis},S})$.
- **PCG**: *intermediate stationary distributions*.
  PCG II: $[\psi_1|\psi_2', \psi_4'] \to [\psi_2, \psi_3|\psi_1, \psi_4'] \to [\psi_4|\psi_1, \psi_2, \psi_3]$.
  Intermediate stationary ending with Step 1:
  $\pi(\psi_1, \psi_2, \psi_3, \psi_4) = p(\psi_2, \psi_3, \psi_4)p(\psi_1|\psi_2, \psi_4)$.
- **MDA**: $[\alpha^\star, \tilde{Y}_{\mathrm{mis}}|\theta'] \longrightarrow [\alpha, \theta|\tilde{Y}_{\mathrm{mis}}]$.
  $p(\theta|\tilde{Y}_{\mathrm{mis}}) = \int p(\alpha, \theta|\tilde{Y}_{\mathrm{mis}})\mathrm{d}\alpha \xlongequal{\text{Set } \tilde{Y}_{\mathrm{mis}} \text{ as } Y_{\mathrm{mis}}} \pi(\theta|Y_{\mathrm{mis}})$;
  $\pi(\theta, Y_{\mathrm{mis}}) = \pi(\theta|Y_{\mathrm{mis}})p(Y_{\mathrm{mis}})$.

Advantages of Surrogate Distribution

- Surrogate distribution unifies different strategies under a common framework.
- For ASIS, a sampler involving surrogate distribution, but equivalent to the original ASIS sampler, has fewer steps.
- If we are only interested in marginal distributions, surrogate distribution strategy is promising to produce more efficient algorithms.

Outline

Imperial College London

Model Review and New Data

Recall:

- **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**:

$$M_i^\epsilon \sim \mathrm{N}(M_0, \sigma_\epsilon^2); \ x_i \sim \mathrm{N}(x_0, R_x^2); \ c_i \sim \mathrm{N}(c_0, R_c^2).$$

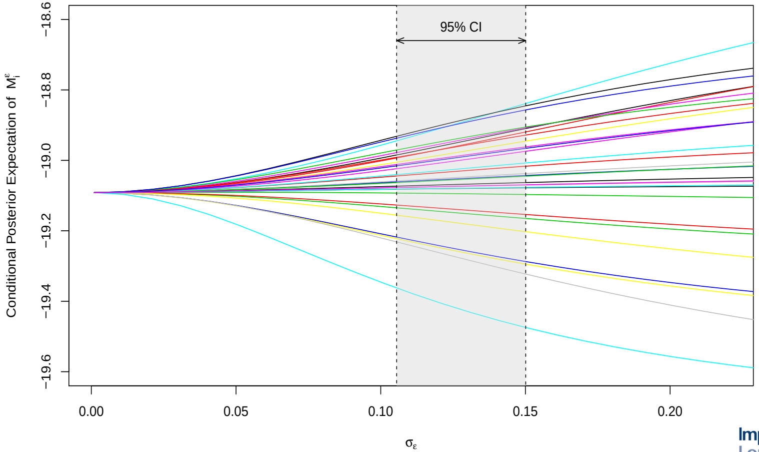
σ_ϵ small $\Longrightarrow$ "Standardizable candle"

Data: A "JLA" sample of 740 SNIa in Betoule, et al. (2014).

Imperial College London

Shrinkage Estimation

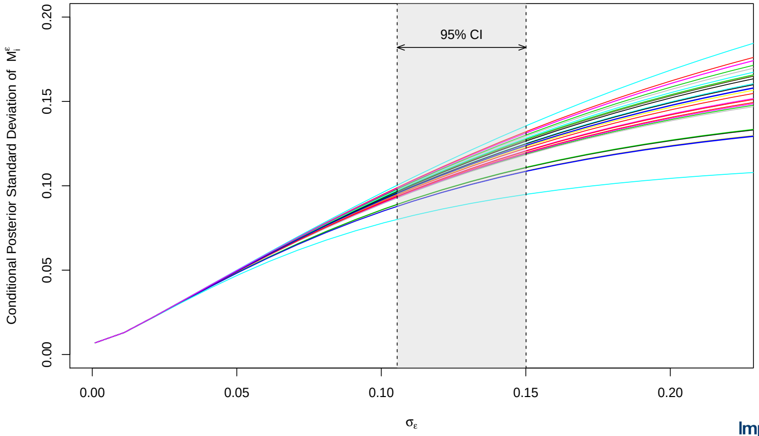
Low mean squared error estimates of M_i^ϵ



Imperial College London

Shrinkage Error

Reduced standard deviations

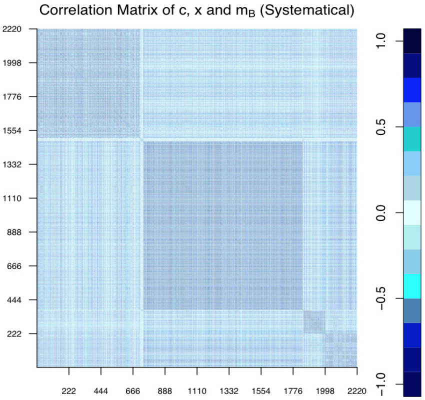


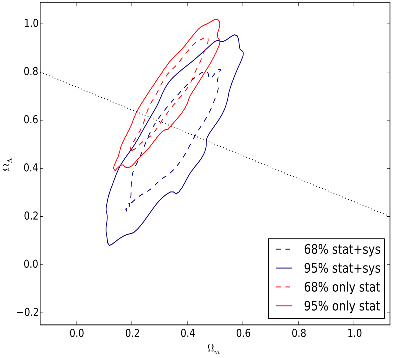
Systematic Errors

- Systematic errors: seven sources of uncertainties.
- Blocks: different surveys.

Effect on cosmological parameters:

$$\hat{C}_{\text{stat}} \text{ vs } \hat{C}_{\text{stat}} + \hat{C}_{\text{sys}}.$$

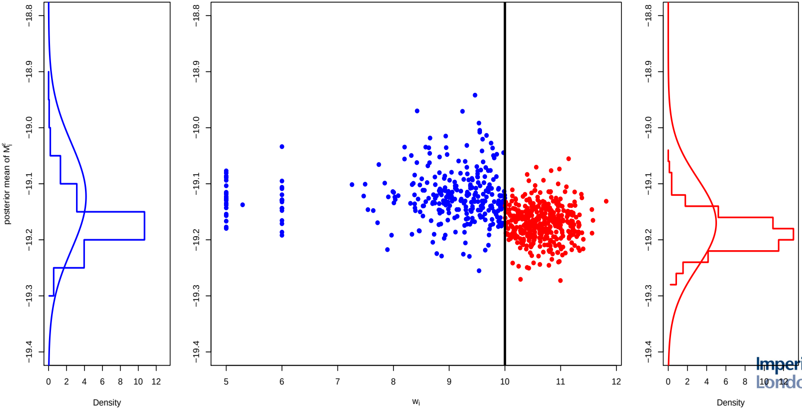




Adjusting for Galaxy Mass: Method I

Method I: Divide M_i^ϵ by $w_i = \log_{10}(M_{\text{galaxy}}/M_\odot)$;

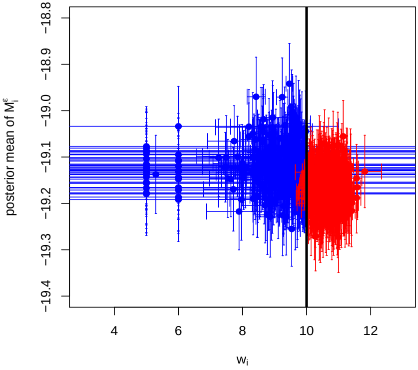
$$\begin{cases} M_i^\epsilon \sim \mathrm{N}(M_{01}, \sigma_{\epsilon 1}^2), \text{ if } w_i < 10, \\ M_i^\epsilon \sim \mathrm{N}(M_{02}, \sigma_{\epsilon 2}^2), \text{ if } w_i > 10. \end{cases}$$



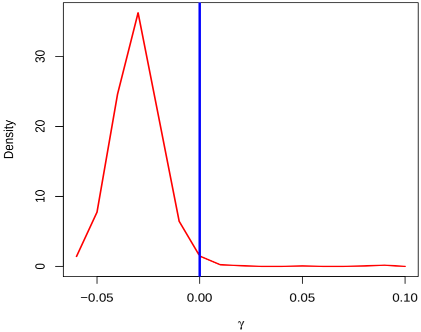
Imperial College London

Adjusting for Galaxy Mass: Method II

Much scatter in both M_i and w_i.



Treat w_i as covariate like x_i and c_i,
$\hat{w}_i \sim \mathrm{N}(w_i, \hat{\sigma}_w^2)$:
$m_{Bi} = \mu_i + M_i^\epsilon - \alpha x_i + \beta c_i + \gamma w_i$.



Imperial College London

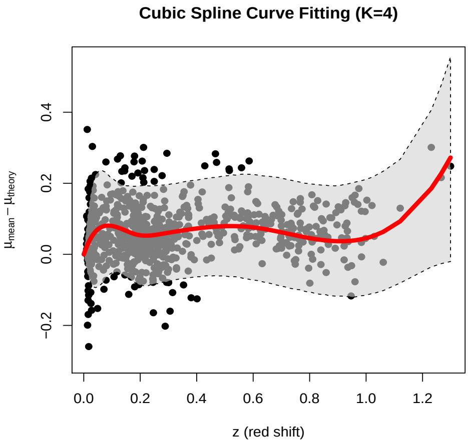
Model Checking

Model setting:

- Fix $(\Omega_m, \Omega_\Lambda)$;
- $m_{Bi} = \tilde{\mu}_i + M_i^\epsilon - \alpha x_i + \beta c_i$;
- $\tilde{\mu}_i = \mu(z_i, \Omega_m, \Omega_\Lambda, H_0) + t(z_i)$, $t(z_i)$—cubic spline

Results:

- Red line—posterior mean;
- Gray band—95% region;
- Black dots—$(\hat{m}_{Bi} - M_0 + \alpha \hat{x}_i - \beta \hat{c}_i) - \tilde{\mu}_i$.



Imperial College London

Outline

Imperial College London

Conclusion

- **Summary**
  - Combining strategy and surrogate distribution samplers are useful to produce more efficiency in convergence.
  - The hierarchical Gaussian model reflects the underlying physical understanding of supernova cosmology.
- **Future Work**
  - More numerical examples to illustrate the algorithms.
  - Complete the theory of surrogate distribution strategy.
  - Embed this hierarchical model into a model for the full time-series of the supernova explosion, using Gaussian process to impute apparent magnitudes over time.

Imperial College London