Iterative Sample Refinement for Sparsity Promoting Bayesian Hierarchical Models

Jasper M. Everink, Technical University of Denmark, jmev@dtu.dk

Joint work with Martin S. Andersen and Yiqiu Dong

VILLUM FONDEN

Computational **U**ncertainty **Q**uantification for **I**nverse problems

Gaussian priors

Ingredients:

- Forward operator $A \in \mathbb{R}^{m \times n}$
- Signal $x_0 \in \mathbb{R}^n$
- Noisy measurements $b = Ax_0 + e$
- with noise $e \sim \mathcal{N}(0, \Sigma_e)$

Explicit Bayesian approach:

$$\underbrace{\pi(x|b)}_{\text{posterior}} \propto \underbrace{\pi(b|x)}_{\text{likelihood}} \underbrace{\pi(x)}_{\text{prior}}$$

- likelihood: $\pi(b|x) \propto \exp\left(-\frac{1}{2}\|Ax - b\|^2_{\Sigma_e^{-1}}\right)$
- Gaussian prior: $\pi(x) \propto \exp\left(-\frac{1}{2}\|x - \mu\|^2_{\Sigma_x^{-1}}\right)$

Massive Advantage:
Posterior is a Gaussian $\rightarrow$ efficient sampling

$$x|b = \underset{z \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \frac{1}{2}\|Az - \hat{b}\|^2_{\Sigma_e^{-1}} + \frac{1}{2}\|z - \hat{\mu}\|^2_{\Sigma_x^{-1}} \right\},$$

with $\hat{b} \sim \mathcal{N}(b, \Sigma_e)$ and $\hat{\mu} \sim \mathcal{N}(\mu, \Sigma_x)$.

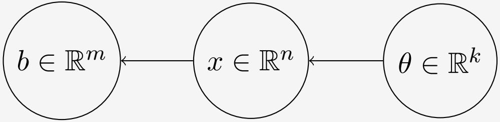
Disadvantage:

- No sparsity
- No constraints

Conditional Gaussian priors

Bayesian hierarchical approach:

$$\underbrace{\pi(x,\theta|b)}_{\text{posterior}} \propto \underbrace{\pi(b|x)}_{\text{likelihood}} \underbrace{\pi(x|\theta)\pi(\theta)}_{\text{prior}}$$



- Cond. Gaussian prior:
  $\pi(x|\theta) \propto \exp\left(-\frac{1}{2}\|x-\mu(\theta)\|^2_{\Sigma_x^{-1}(\theta)}\right)$

Advantage:

- Efficient sampling of $\pi(x|b,\theta)$
- Compressibility (approximately sparse)

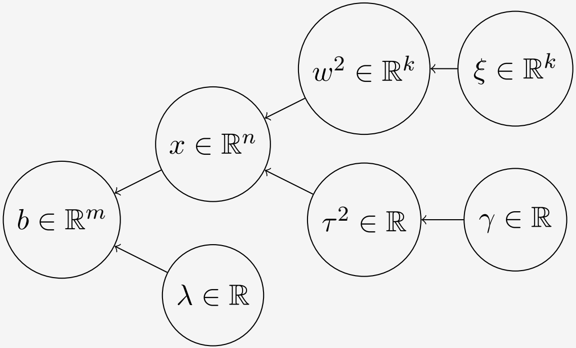
Disadvantage:

- No constraints
- Compressibility requires large hierarchical model ($k \gtrsim n$)

Example of conditional Gaussian model

Horseshoe/shrinkage model:

- $b|x, \lambda \sim \mathcal{N}(Ax, \lambda^{-1}\Sigma_e)$
- $x|\tau^2, w^2 \sim \mathcal{N}(0, \Lambda^{-1}(\tau, w))$
- $\tau^2|\gamma \sim \Gamma^{-1}\left(\frac{\nu}{2}, \frac{\nu}{\gamma}\right)$
- $w_i^2|\xi_i \sim \Gamma^{-1}\left(\frac{\nu}{2}, \frac{\nu}{\xi_i}\right)$
- $\lambda \sim \Gamma(\alpha_\lambda, \beta_\lambda)$
- $\gamma \sim \Gamma^{-1}\left(\frac{1}{2}, \frac{1}{\tau_0^2}\right)$
- $\xi_i \sim \Gamma^{-1}\left(\frac{1}{2}, 1\right)$



Details: *Uribe, F., Dong, Y., & Hansen, P. C. (2023). Horseshoe priors for edge-preserving linear Bayesian inversion. SIAM Journal on Scientific Computing*

Between Gaussians and Regularization

Efficient sampling from a Gaussian:

$$\operatorname*{argmin}_{z\in\mathbb{R}^n}\left\{\frac{1}{2}\|Az-\hat{b}\|^2_{\Sigma_e^{-1}}+\frac{1}{2}\|z-\hat{\mu}\|^2_{\Sigma_x^{-1}}\right\}.$$

Randomization:

- $\hat{b}\sim\mathcal{N}(b,\Sigma_e)$
- $\hat{\mu}\sim\mathcal{N}(\mu,\Sigma_x)$

Regularized Gaussian distibution:

$$\operatorname*{argmin}_{z\in\mathbb{R}^n}\left\{\frac{1}{2}\|Az-\hat{b}\|^2_{\Sigma_e^{-1}}+\frac{1}{2}\|z-\hat{\mu}\|^2_{\Sigma_x^{-1}}+f(z)\right\}.$$

Sparsity from regularization:

$$\operatorname*{argmin}_{z\in\mathbb{R}^n}\left\{\frac{1}{2}\|Az-b\|^2_{\Sigma_e^{-1}}+f(z)\right\},\ \text{e.g.,}\ \ f(z)=\gamma\|Lz\|_1.$$

Regularized Gaussian distribution

Implicit distribution approach:

$$x|b := \underset{z\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \underbrace{\frac{1}{2}\|Az-\hat{b}\|^2_{\Sigma_e^{-1}}}_{\text{likelihood}} + \underbrace{\frac{1}{2}\|z-\hat{\mu}\|^2_{\Sigma_x^{-1}} + f(z)}_{\text{prior}} \right\},$$

with $\hat{b} \sim \mathcal{N}(b, \Sigma_e)$ and $\hat{\mu} \sim \mathcal{N}(\mu, \Sigma_x)$.

Examples of $f(z)$:

- $\gamma\|z\|_1$ for sparsity
- $\chi_{\mathbb{R}^n_{\geq 0}}(z)$ for constraints
- $\sum_{i=1}^{l} \|L_i z - c_i\|_p,\ p \geq 0$
- $\max_{i=1,\ldots,l} f_i(z),\ f_i \in C^1_{\text{conv}}(\mathbb{R}^n)$

Advantage:

- Efficient sampling
- Flexibility (sparsity and constraints)

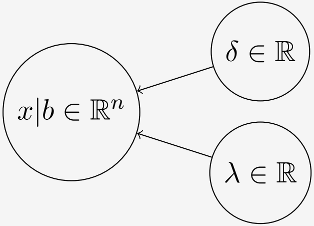
Disadvantage:

- Implicit distribution

Example hierarchical model

Model: assume f is piecewise linear

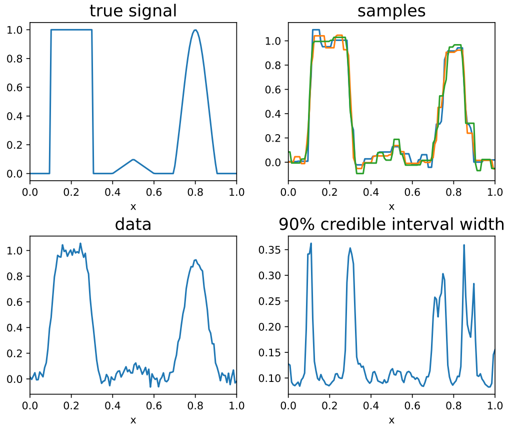
- $x|b, \lambda, \delta := \operatorname{argmin}_{z \in \mathbb{R}^n} \left\{ \frac{\lambda}{2} \|Az - \hat{b}\|^2_{\Sigma_e^{-1}} + \frac{\delta^2}{2} \|z - \hat{\mu}\|^2_{\Sigma_x^{-1}} + \delta f(z) \right\},$
  with $\hat{b} \sim \mathcal{N}(b, \lambda^{-1}\Sigma_e)$ and $\hat{\mu} \sim \mathcal{N}(0, \delta^{-2}\Sigma_x)$
- $\lambda \sim \Gamma(\alpha_\lambda, \beta_\lambda)$
- $\delta \sim \mathrm{MHN}(\alpha_\delta, \beta_\delta, -\gamma_\delta), \quad \pi(\delta) \propto \delta^{\alpha_\delta - 1} \exp\left(-\beta_\delta \delta^2 - \gamma_\delta \delta\right)$



- $\lambda | x \sim \Gamma\left(\frac{m}{2} + \alpha_\lambda, \frac{1}{2}\|Ax - b\|^2_{\Sigma_e^{-1}} + \beta_\lambda\right)$
- $\delta | x \sim \mathrm{MHN}\left(\frac{n - \dim(\partial f(x))}{2} + \alpha_\delta, \frac{1}{2}\|x - \mu\|^2_{\Sigma_x^{-1}} + \beta_\delta, -f(z) - \gamma_\delta\right)$

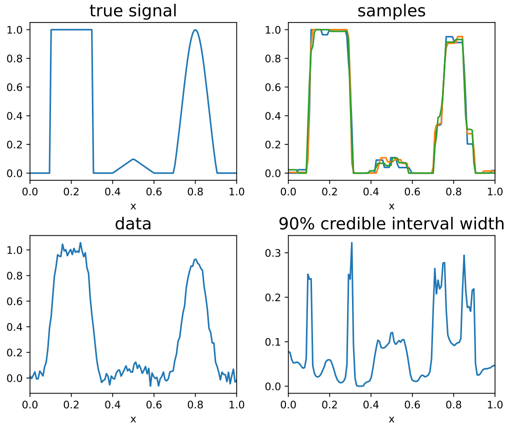
Computational example, Gaussian deblurring in $\mathbb{R}^{128}$

- $x|b, \lambda, \delta := \operatorname{argmin}_{z \in \mathbb{R}^n} \left\{ \frac{\lambda}{2} \|Az - \hat{b}\|_2^2 + \frac{\delta^2}{2} \|L(z - \hat{\mu})\|_2^2 + 20\delta \|Lz\|_1 \right\},$
- $\lambda \sim \Gamma(1, 10^{-4})$, $\delta \sim \mathrm{MHN}(1, 10^{-4}, -10^{-4})$



Computational example, Gaussian deblurring in $\mathbb{R}^{128}$

- $x|b,\lambda,\delta := \text{argmin}_{z\in[0,1]^n}\left\{\frac{\lambda}{2}\|Az-\hat{b}\|_2^2+\frac{\delta^2}{2}\|L(z-\hat{\mu})\|_2^2+20\delta\|Lz\|_1\right\}$,
- $\lambda \sim \Gamma(1, 10^{-4})$, $\delta \sim \text{MHN}(1, 10^{-4}, -10^{-4})$



The convergence issue

Conditional Gaussian:
Convergence of the Markov chain with inaccurate solutions*:

$$x_{n+1} = \underset{z \in x_n + \mathcal{K}_{n+1}}{\operatorname{argmin}} \left\{ \frac{1}{2} \| Az - \hat{b} \|^2_{\Sigma_e^{-1}} + \frac{1}{2} \| z - \hat{\mu} \|^2_{\Sigma_x^{-1}} \right\},$$

due to access to an explicit density.

*Details: *Féron, O., Orieux, F., & Giovannelli, J. F. (2015). Gradient scan Gibbs sampler: An efficient algorithm for high-dimensional Gaussian distributions. IEEE Journal of Selected Topics in Signal Processing*

Because we do not know the implicit density of

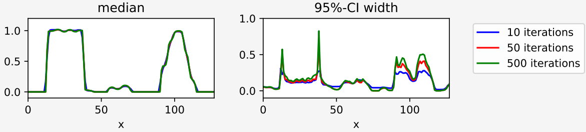
$$\underset{z \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \frac{1}{2} \| Az - \hat{b} \|^2_{\Sigma_e^{-1}} + \frac{1}{2} \| z - \hat{\mu} \|^2_{\Sigma_x^{-1}} + f(z) \right\},$$

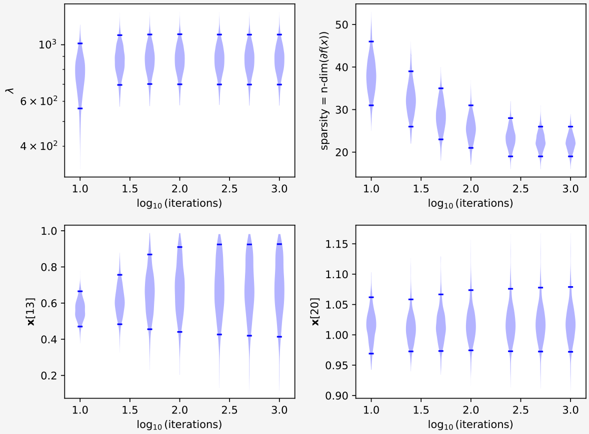
We do not have convergence guarantees when solving the problem inaccurately, unless...

The convergence issue, is it an issue?

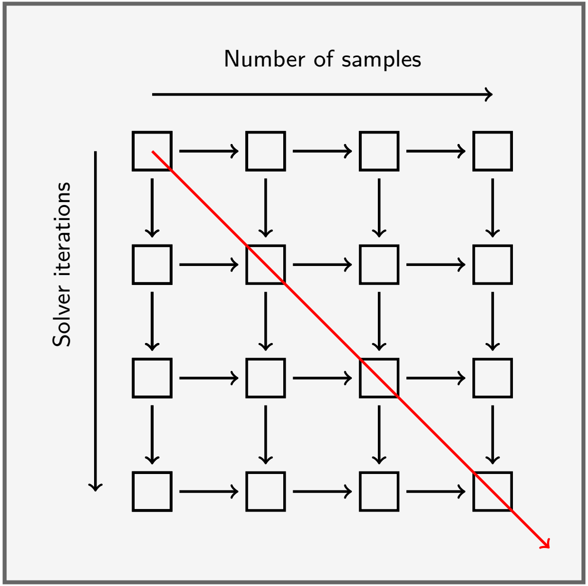
How many iterations?

Solver: ADMM





Iterative Sample Refinement: the idea



Iterative Sample Refinement: technical

(Randomize-then-Optimize)-within-Gibbs:

- For $i = 1, \ldots, \infty$:
    - Sample $\lambda_i \sim \lambda \,|\, x_{i-1}, b$
    - Sample $\delta_i \sim \delta \,|\, x_{i-1}, b$
    - Sample $\hat{b}_i \sim \mathcal{N}(b, \lambda_i^{-1}\Sigma_e)$
    - Sample $\hat{\mu}_i \sim \mathcal{N}(\mu, \delta_i^{-1}\Sigma_x)$
    - Solve for x_i

Iterative Sample Refinement: technical

Randomize-then-(Optimize-within-Gibbs):

- For $i = 1, \ldots,$: (pre-randomization)
  - Sample random functions λ_i, δ_i, $\hat{b}_i$, $\hat{\mu}_i$
- For $i = 1, \ldots, \infty$
  - Compute $\lambda_i = \lambda_i(x_{i-1})$
  - Compute $\delta_i = \delta_i(x_{i-1})$
  - Compute $\hat{b}_i = \hat{b}_i(\lambda_i)$
  - Compute $\hat{\mu}_i = \hat{\mu}_i(\delta_i)$
  - Solve for x_i

Replace random sampling by deterministic function, e.g.,

replace $\lambda_i \sim \lambda \,|\, x_{i-1}, b$ by $x \mapsto \lambda_i \,|\, x, b$.

Examples:

$$\lambda_i(x) = \text{CDF}^{-1}_{\Gamma(\alpha(x),\beta(x))}(u_i),\ u_i \sim \text{Unif}(0,1)$$

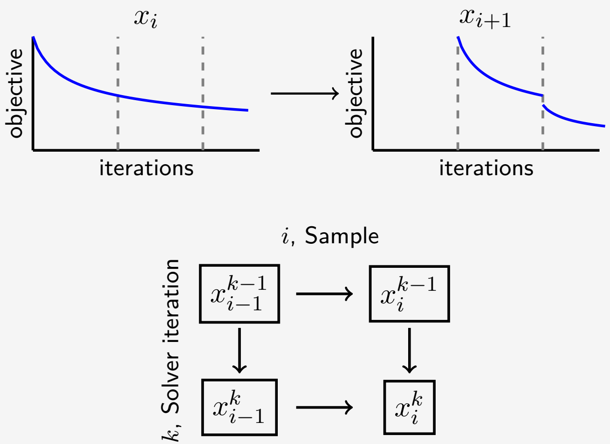
$$\hat{b}_i(\lambda) = b + \sqrt{\lambda}\epsilon_i,\ \epsilon \sim \mathcal{N}(0, \Sigma_e).$$

Turns a Gibbs sampler into a sequence of optimization problems.

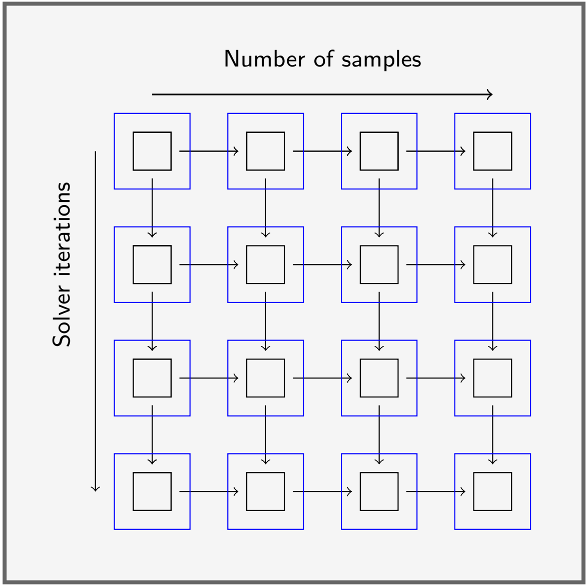
Iterative Sample Refinement: technical

Iterative sample refinement

- For $i = 1, \ldots, i_{\max}$: (pre-randomization)
  - Sample functions λ_i, δ_i, $\hat{b}_i$, $\hat{\mu}_i$
- For $k = 1, \ldots, k_{\max}$
  - For $i = 1, \ldots, i_{\max}$
    - Compute $\lambda_i = \lambda_i(x_{i-1}^{k-1})$
    - Compute $\delta_i = \delta_i(x_{i-1}^{k-1})$
    - Compute $\hat{b}_i = \hat{b}_i(\lambda_i)$
    - Compute $\hat{\mu}_i = \hat{\mu}_i(\delta_i)$
    - Improve x_i^k from x_i^{k-1}

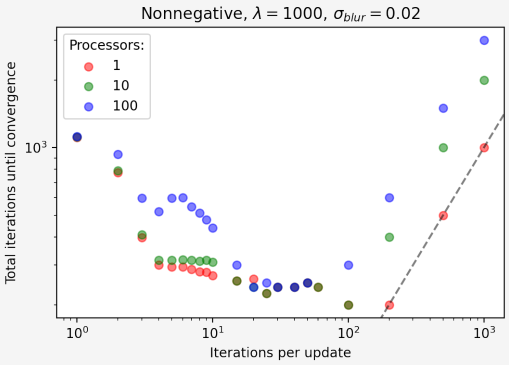


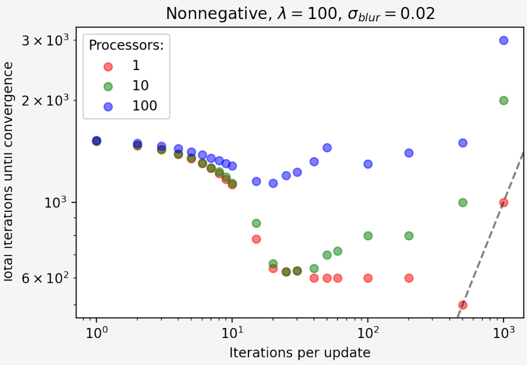
Iterative Sample Refinement: parallelization



Iterative Sample Refinement: examples

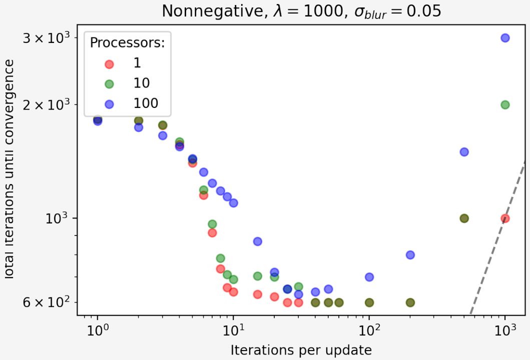
1D Gaussian deblurring with nonnegativity constraints, solved using FISTA.

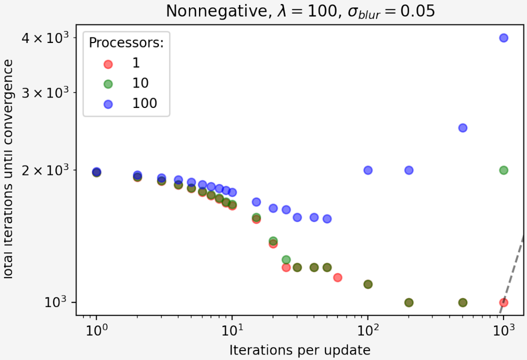




Iterative Sample Refinement: examples

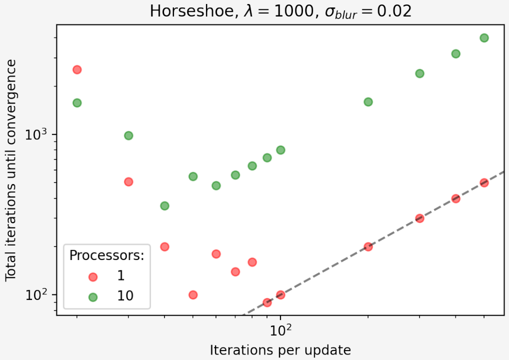
1D Gaussian deblurring with nonnegativity constraints, solved using FISTA.

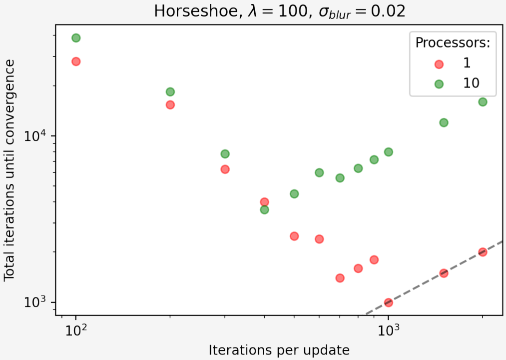




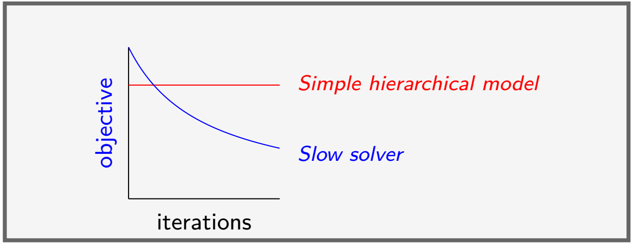
Iterative Sample Refinement: examples

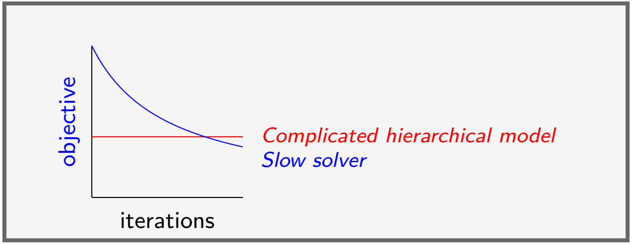
1D Gaussian deblurring with horseshoe model solved using PCGLS.

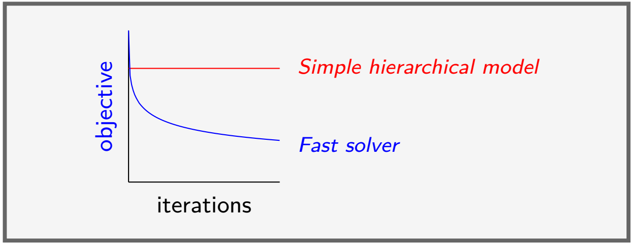


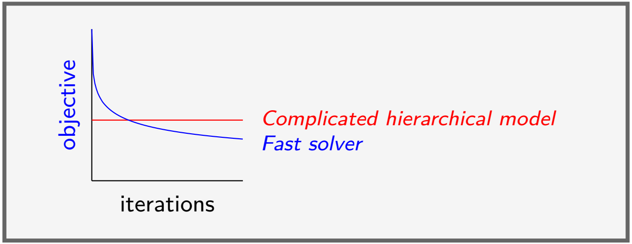


Iterative Sample Refinement: is it beneficial?









Open questions

- Does it scale well to large scale problems?
- What is a good performance measure?
- Are there "easy" ways of tuning the setup?

Summary

Iterative sample refinement:
A trade-off between more and/or better samples in Randomise-then-Optimise(-within-Gibbs) samplers.

Advantages:

- Guaranteed convergence.
- Can be parallelized.

Disadvantages:

- Difficult to tune and analyse.
- Only beneficial for simple hierarchical models

DTU

Iterative Sample Refinement for Sparsity Promoting Bayesian Hierarchical Models

Jasper M. Everink, Technical University of Denmark, jmev@dtu.dk

Joint work with Martin S. Andersen and Yiqiu Dong

Papers (on regularized Gaussian distribution):

- *Bayesian inference with projected densities, SIAM/ASA JUQ*
- *Sparse Bayesian inference with regularized Gaussian distributions, Inverse Problems*

CUQI VILLUM FONDEN

Computational **U**ncertainty **Q**uantification for **I**nverse problems