Computationally efficient UQ for sparse Bayesian learning[a]

UQ for Math Bio @ ICERM

Jan Glaubitz

Joint work with *Y. Marzouk*

May 08, 2025

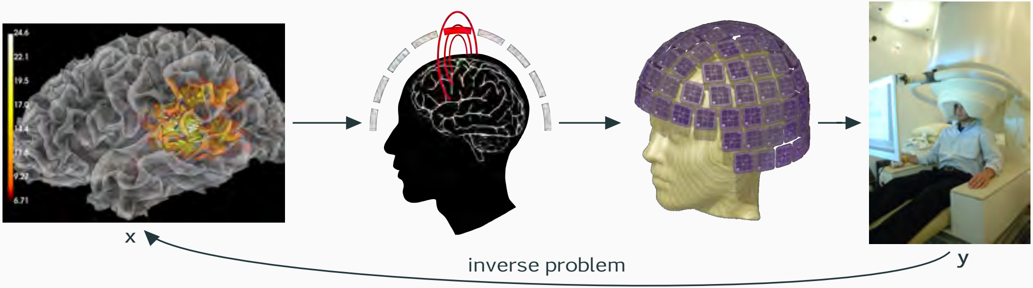
Department of Mathematics
Linköping University
Linköping, Sweden

Uncertainty Quantification Group
Massachusetts Institute of Technology
Cambridge, MA, USA

Associated paper: Glaubitz and Marzouk. "Efficient sampling for sparse Bayesian learning using hierarchical prior normalization." arXiv preprint arXiv:2505.23753 (2025).

[a] This work was supported by the US DOD (ONR MURI) grant #N00014-20-1-2595, the US DOE (SciDAC) grant #DE-SC0012704, and the NAISS grant #2024/22-1207

Motivation: MagnetoEncephaloGraphy (MEG)



Goal. Estimate the brain's electromagnetic activity from its magnetic field[1][2]

Challenge. Resulting inverse problem $\mathbf{y} = F(\mathbf{x}) + \mathbf{e}$ is ill-posed

Belief. $\mathbf{x} \in \mathbb{R}^n$ is sparse $\rightarrow$

$$\min_{\mathbf{x}} \Big\{ \underbrace{\|F(\mathbf{x}) - \mathbf{y}\|}_{\text{data fidelity}} + \underbrace{\lambda \mathcal{R}(\mathbf{x})}_{\text{regularization}} \Big\}, \quad \mathcal{R}(\mathbf{x}) = \|\mathbf{x}\|_1$$

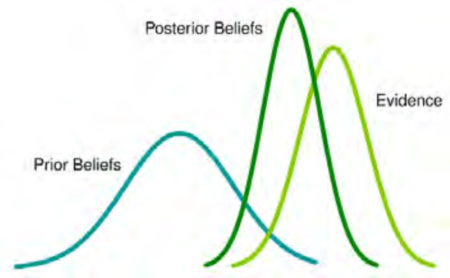
[1] Telenczuk et al., "RAMP: predicting the neural origin of magnetic signals," 2025

[2] Calvetti et al., "A hierarchical Krylov–Bayes iterative inverse solver for MEG with physiological preconditioning," 2015

- Treat the observations $\mathbf{y}$ and parameters $\mathbf{x}$ as random variables
- Goal: Explore the posterior $\mathbf{x}|\mathbf{y}$, given by Bayes' theorem as

$$\pi^{y}(\mathbf{x}) \propto f(\mathbf{x};\mathbf{y})\,\pi^{0}(\mathbf{x})$$

- The likelihood $f(\mathbf{x};\mathbf{y})$ is a probabilistic versions of the data fidelity term
- The prior $\pi^{0}(\mathbf{x})$ is a probabilistic versions of the regularization term

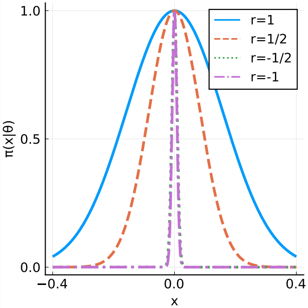


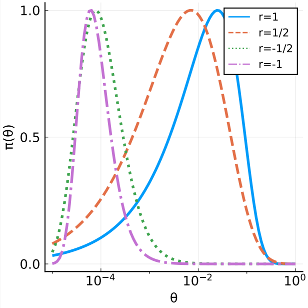
[3] Stuart, "Inverse problems: a Bayesian perspective," 2010

[4] Calvetti & Somersalo, "Bayesian Scientific Computing," 2023

Sparse Bayesian learning (SBL) priors:[a][b]

$$
\begin{aligned}
&\text{(joint prior)} && \pi^0(\mathrm{x}, \boldsymbol{\theta}) = \pi^0(\mathrm{x}|\boldsymbol{\theta})\,\pi^0(\boldsymbol{\theta}),\\
&\text{(cond. prior)} && \pi^0(\mathrm{x}|\boldsymbol{\theta}) = \prod_i \mathcal{N}(x_i|0, \theta_i),\\
&\text{(hyper-prior)} && \pi^0(\boldsymbol{\theta}) = \prod_i \mathcal{GG}(\theta_i|r, \beta, \vartheta)
\end{aligned}
$$





[a] Tipping, "Sparse Bayesian learning and the relevance vector machine," 2001

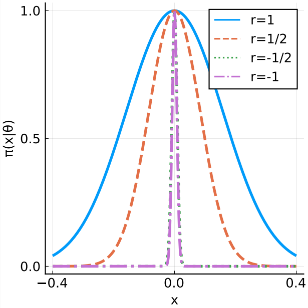
[b] Calvetti et al., "Sparse reconstructions from few noisy data," 2020

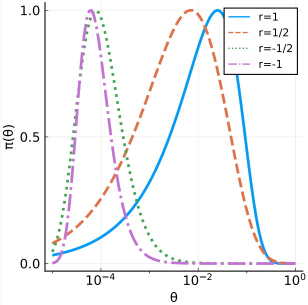
Sparse Bayesian learning (SBL) priors:[a][b]

$$\begin{aligned}
&\text{(joint prior)} && \pi^0(\mathbf{x},\boldsymbol{\theta}) = \pi^0(\mathbf{x}|\boldsymbol{\theta})\,\pi^0(\boldsymbol{\theta}),\\
&\text{(cond. prior)} && \pi^0(\mathbf{x}|\boldsymbol{\theta}) = \prod_i \mathcal{N}(x_i|0,\theta_i),\\
&\text{(hyper-prior)} && \pi^0(\boldsymbol{\theta}) = \prod_i \mathcal{GG}(\theta_i|r,\beta,\vartheta)
\end{aligned}$$

Intuition:

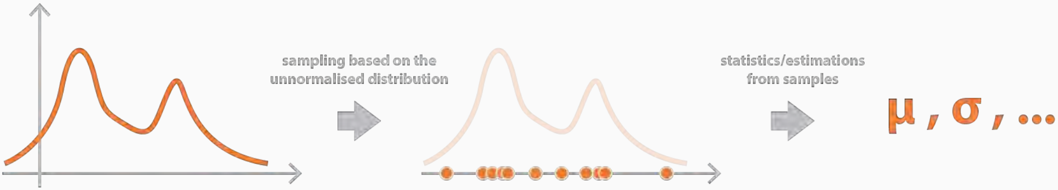
- $\mathcal{GG}$ makes $\theta_i \approx 0$ likely, while allowing for outliers
- If $\theta_i \approx 0$, then $\mathbf{x}_i^2 \approx 0$ is most likely
- If θ_i is outlier ($\theta_i \gg 0$), $x_i^2 \gg 0$ becomes more likely





[a] Tipping, "Sparse Bayesian learning and the relevance vector machine," 2001
[b] Calvetti et al., "Sparse reconstructions from few noisy data," 2020

Bayesian inference: MCMC sampling



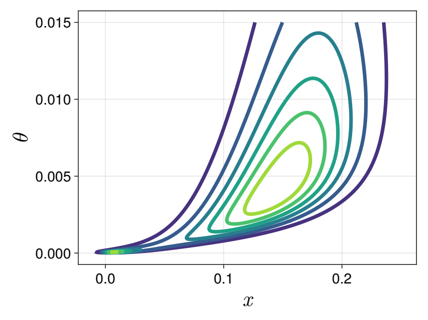
Goal. Probe posterior $\pi^{y}(\mathbf{x}, \theta) \propto f(\mathbf{x}; \mathbf{y}) \pi^{0}(\mathbf{x}, \theta)$ using MCMC

Challenges for SBL.[a]

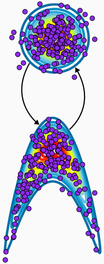
(i) multi-modal

(ii) strong correlation between $\mathbf{x}$ and θ

(iii) high-dimensional



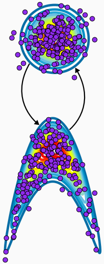
[a] Calvetti & Somersalo, "Computationally efficient sampling methods for sparsity promoting hierarchical Bayesian models," 2024



Idea. Transform the SBL prior $\pi^0(\mathbf{x}, \boldsymbol{\theta})$ into a standard normal one $\phi^0(\mathbf{u}, \boldsymbol{\tau})$

Need. $S : (\mathbf{x}, \boldsymbol{\theta}) \mapsto (\mathbf{u}, \boldsymbol{\tau})$ s. t. $(\mathbf{u}, \boldsymbol{\tau})$ follows ϕ^0 whenever $(\mathbf{x}, \boldsymbol{\theta})$ follows π^0

Our approach: Hierarchical prior normalization

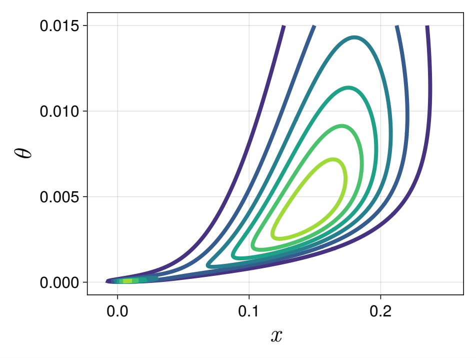


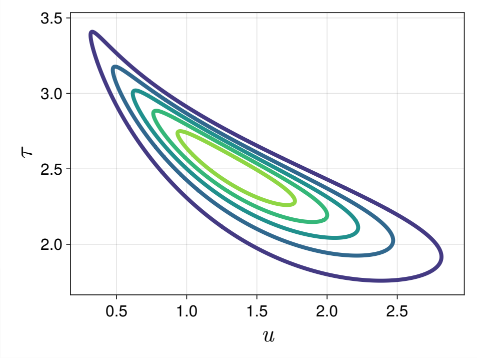
Idea. Transform the SBL prior $\pi^0(\mathbf{x}, \boldsymbol{\theta})$ into a standard normal one $\phi^0(\mathbf{u}, \boldsymbol{\tau})$

Need. $S : (\mathbf{x}, \boldsymbol{\theta}) \mapsto (\mathbf{u}, \boldsymbol{\tau})$ s. t. $(\mathbf{u}, \boldsymbol{\tau})$ follows ϕ^0 whenever $(\mathbf{x}, \boldsymbol{\theta})$ follows π^0

Get. Simpler "prior-normalized" posterior

$$\phi^y(\mathbf{u}, \boldsymbol{\tau}) = (S_\# \pi^y)(\mathbf{u}, \boldsymbol{\tau}) = \frac{1}{Z} f(S^{-1}(\mathbf{u}, \boldsymbol{\tau}); \mathbf{y})\, \phi^0(\mathbf{u}, \boldsymbol{\tau})$$





Product-like form.

$$\pi^0(\mathbf{x},\boldsymbol{\theta}) = \prod_{i=1}^{n} \mathcal{N}(x_i|0,\theta_i)\mathcal{GG}(\theta_i|r,\beta,\vartheta) \implies S(\mathbf{x},\boldsymbol{\theta}) = \begin{bmatrix} s_1(x_1,\theta_1) \\ \vdots \\ s_n(x_n,\theta_n) \end{bmatrix}$$

with $s_i : (x_i, \theta_i) \mapsto (u_i, \tau_i)$ transforming $\mathcal{N}(x_i|0,\theta_i)\mathcal{GG}(\theta_i|r,\beta,\vartheta)$ into $\mathcal{N}(\mathbf{0}_2, I_2)$

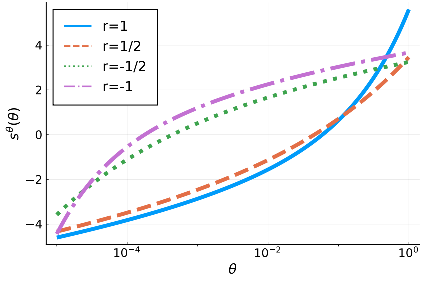
Product-like form.

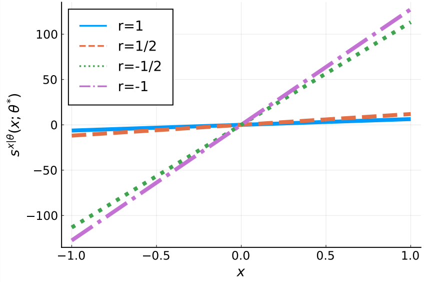
$$\pi^0(\mathbf{x},\boldsymbol{\theta}) = \prod_{i=1}^{n} \mathcal{N}(x_i|0,\theta_i)\mathcal{GG}(\theta_i|r,\beta,\vartheta) \implies S(\mathbf{x},\boldsymbol{\theta}) = \begin{bmatrix} s_1(x_1,\theta_1) \\ \vdots \\ s_n(x_n,\theta_n) \end{bmatrix}$$

with $s_i : (x_i,\theta_i) \mapsto (u_i,\tau_i)$ transforming $\mathcal{N}(x_i|0,\theta_i)\mathcal{GG}(\theta_i|r,\beta,\vartheta)$ into $\mathcal{N}(\mathbf{0}_2,I_2)$

Knothe–Rosenblatt (KR) rearrangements.

$$s(x,\theta) = \begin{bmatrix} s^{\theta}(\theta) \\ s^{x|\theta}(x;\theta) \end{bmatrix}, \quad \begin{cases} s^{\theta} = \left(\Psi^0\right)^{-1} \circ \mathcal{P}^{\theta}, \\ s^{x|\theta} = x/\sqrt{\theta}. \end{cases}$$

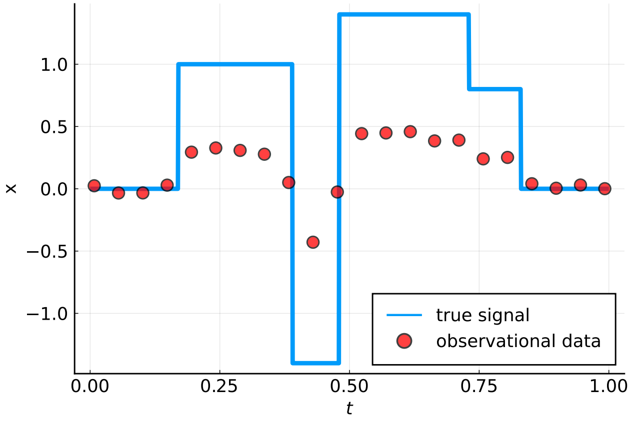




Example 1: Signal deblurring

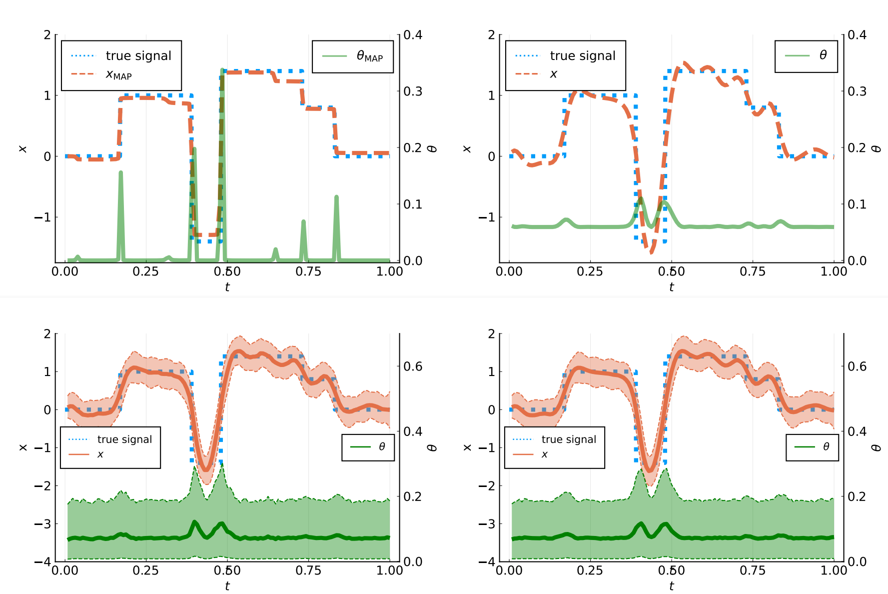
Goal. Recover $n = 128$ nodal values from $m = 22$ noisy blurry observations:

$$\mathbf{y} = F\mathbf{x} + \mathbf{e} \quad \text{where} \quad [L\mathbf{x}]_i = x_{i+1} - x_i \text{ is sparse}$$



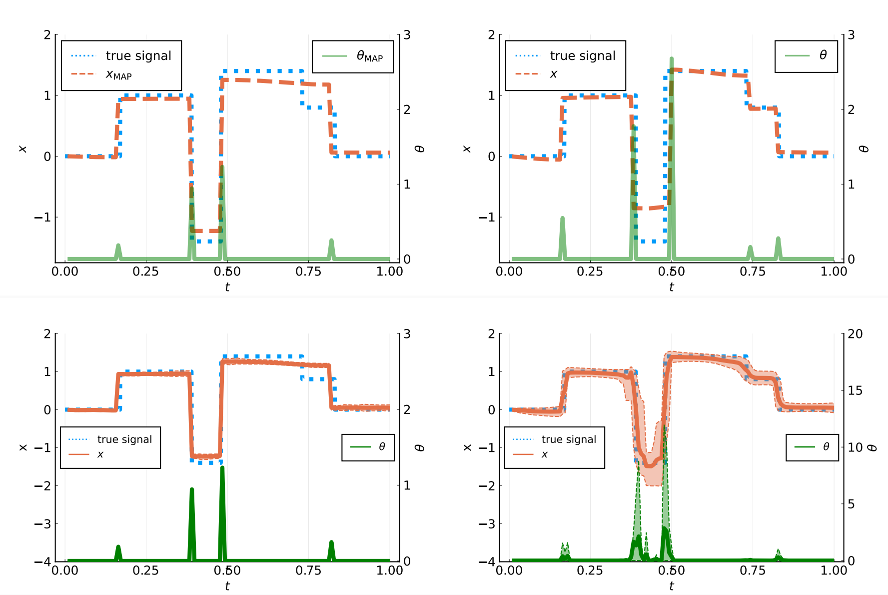
Compare. AM sampler applied to original and prior-normalized posterior

Setup. $r = 1$, $J = 6$ chains, 10^7 samples, MAP initialization

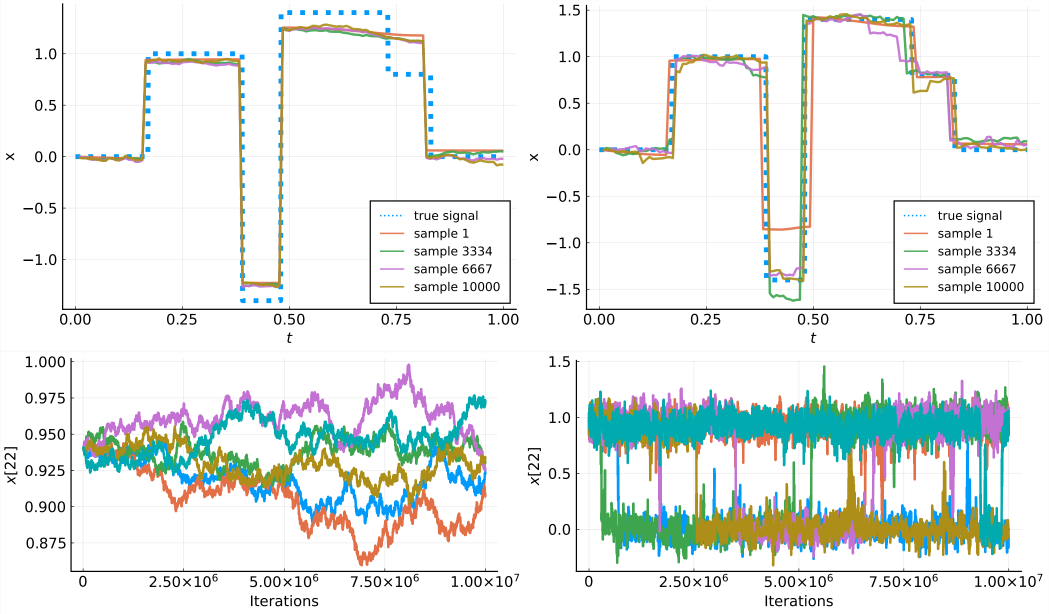


Statistics (MAP initialization, $r = -1$)

Setup. $r = -1$, $J = 6$ chains, 10^7 samples, MAP initialization

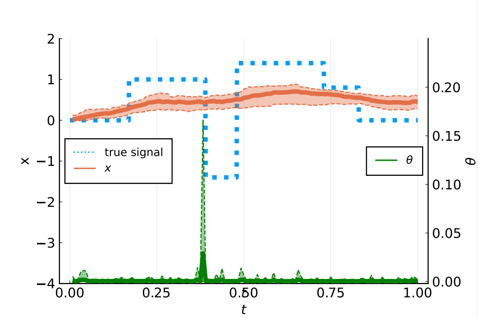


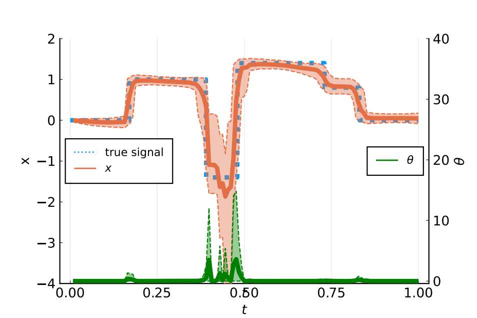
Samples & traces (MAP initialization, $r = -1$)

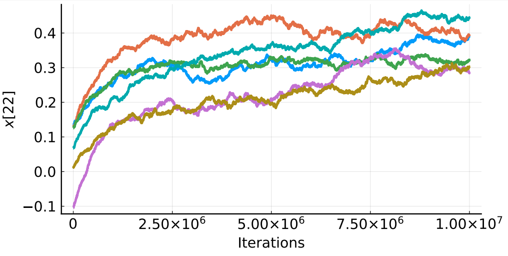


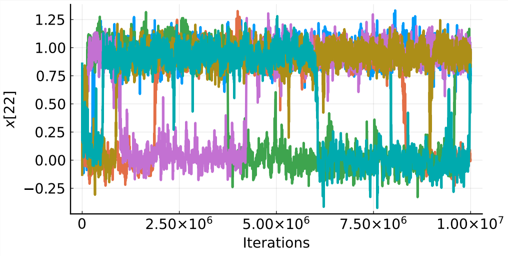
Statistics & samples (random initialization, $r = -1$)

Setup. $r = -1$, $J = 6$ chains, 10^7 samples, random initialization







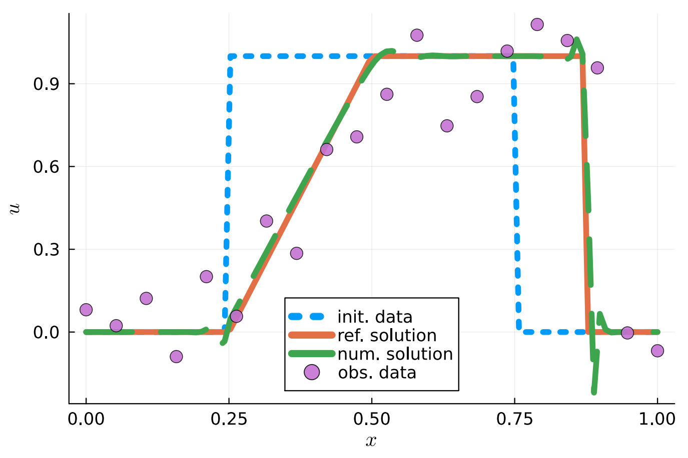


Example 2: Inverse Burgers' equation

Burgers' equation. $\partial_t u(x,t) + \partial_x u(x,t)^2 = 0$ with $u(x,0) = u_0(x)$

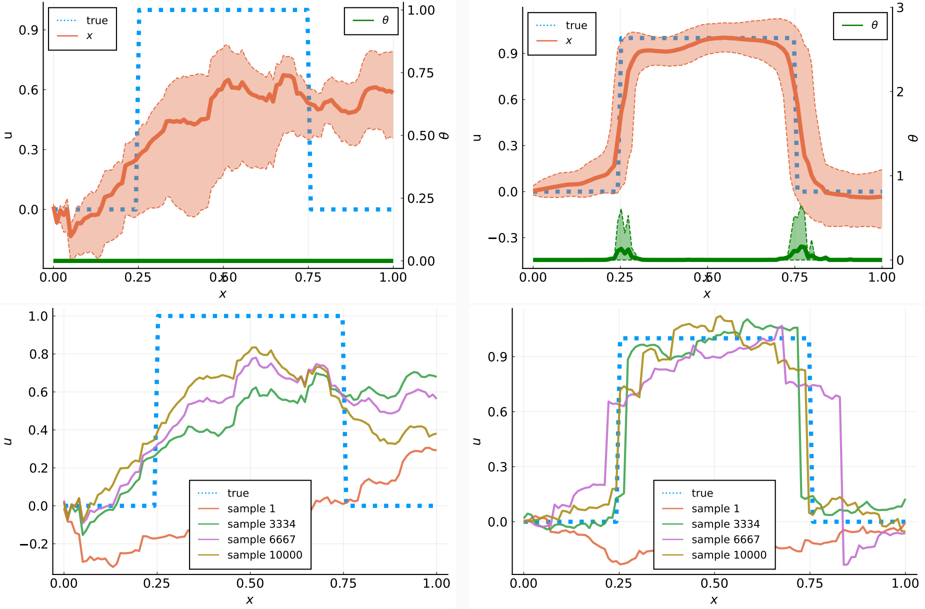
Goal. Recover u_0 at $n = 100$ points from $m = 20$ noisy solution observations:

$$\mathbf{b} = \mathcal{F}(\mathbf{u}) + \mathbf{e} \quad \text{where} \quad [L\mathbf{u}]_i = u_{i+1} - u_i \text{ is sparse}$$



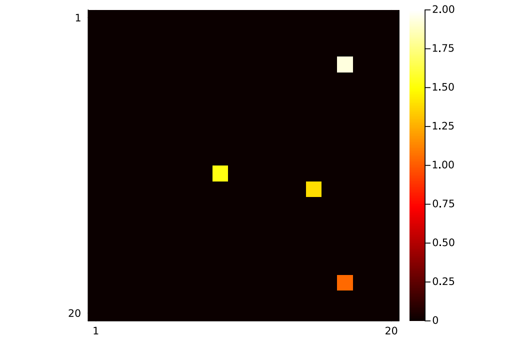
Statistics & samples

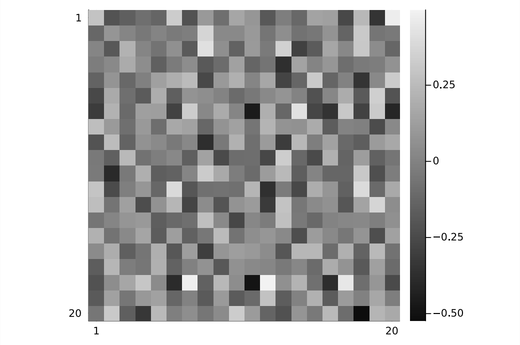
Setup. $r = -1$, $J = 1$ chain, 10^7 samples, random initialization



Goal. Recover 20×20 impulse image from its noisy DCT:

$$\mathbf{y} = F\mathbf{x} + \mathbf{e} \quad \text{where} \quad \mathbf{x} \text{ is sparse}$$



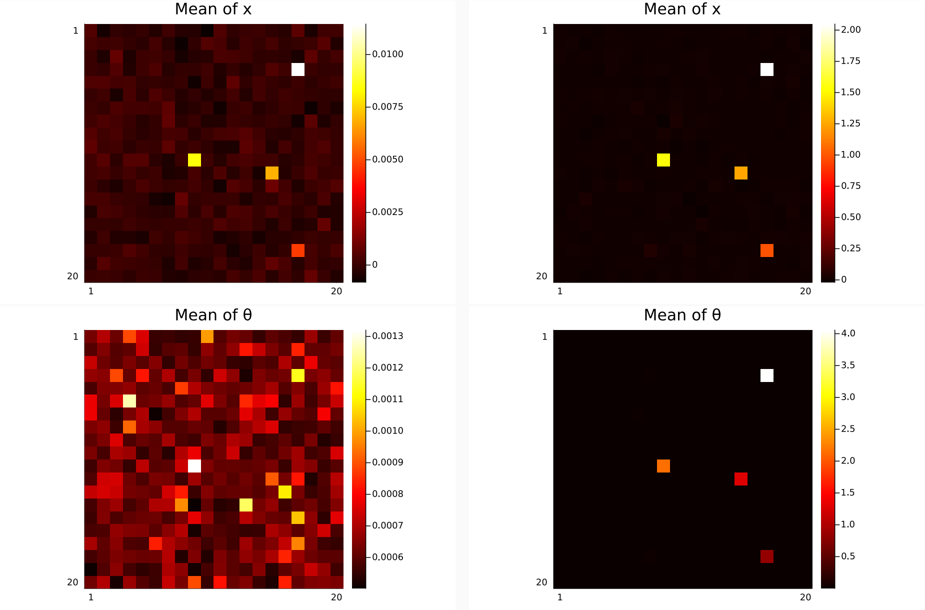


Setup. $r = -1$, $J = 4$ chains, random initialization

Compare. Original posterior (Gibbs sampler, 10^5 samples, ≈ 20m runtime) vs prior-normalized posterior (ESS, 10^4, ≈ 3m runtime)

Statistics

Original (left) vs prior-normalized (right) posterior



Takeaways.

- We propose hierarchical prior-normalization
- Yields significant improvements
- Can be done analytically

Outlook.

- Generally applies to scale-mixtures-of-normal priors
- Combine with geometry-exploiting samplers
- Group/joint-sparsity

The End

Thank You!

jan.glaubitz@liu.se

For more details:

Glaubitz and Marzouk. “Efficient sampling for sparse Bayesian learning using hierarchical prior normalization.” arXiv preprint arXiv:2505.23753 (2025)