WESTFÄLISCHE
WILHELMS-UNIVERSITÄT
MÜNSTER

MCMC Sampling for Bayesian Inference using L1-type Priors

(*what I do whenever the ill-posedness of EEG/MEG is just not frustrating enough!*)

AG Imaging Seminar

Felix Lucka

26.06.2012

Sparsity Constraints in Inverse Problems

Current trend in high dimensional inverse problems: Sparsity constraints.

- Total Variation (TV) imaging: Sparsity constraints on the gradient of the unknowns.
- Compressed Sensing: High quality reconstructions from a small amount of data, if a sparse basis/dictionary is a-priori known (e.g., wavelets).

Some nice images here!

Sparsity Constraints in Inverse Problems

Commonly applied formulation and analysis by means of variational regularization, mostly by incorporating L1-type norms:

$$\hat{u}_\alpha = \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \|m - A\,u\|_2^2 + \alpha\, |D\,u|_1 \right\}$$

assuming additive Gaussian i.i.d. noise $\sim \mathcal{N}(0, \sigma^2)$

Notation:

- $m \in \mathbb{R}^k$: The noisy measurement data given
- $u \in \mathbb{R}^n$: The unknowns to recover w.r.t. the chosen discretization
- $A \in \mathbb{R}^{k \times n}$: Discretization of the forward operator w.r.t. the domains of u and m.
- $D \in \mathbb{R}^{l \times n}$: Discrete formulation of the mapping onto the (potentially) sparse quantity.

Sparsity Constraints in Inverse Problems

Sparsity constraints relying on L1-type norms can also be formulated in the Bayesian framework.

- Likelihood model:

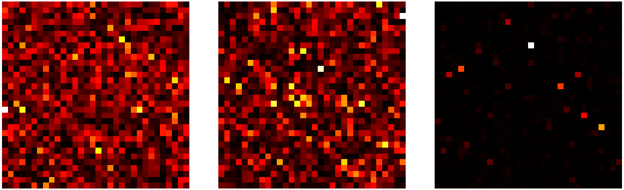
$$M = A\,u + \mathcal{E} \quad \overset{\mathcal{E} \sim \mathcal{N}(0, \sigma^2 I_k)}{\Longrightarrow} \quad p_{li}(m|u) \propto \exp\left(-\frac{1}{2\,\sigma^2}\|m - A\,u\|_2^2\right)$$

- Prior model:

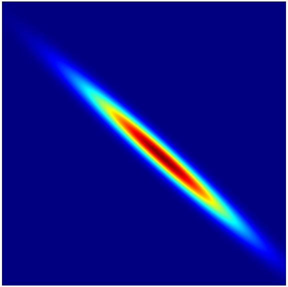
$$p_{pr}(u) \propto \exp\left(-\lambda\,|D\,u|_1\right)$$

- Resulting posterior:

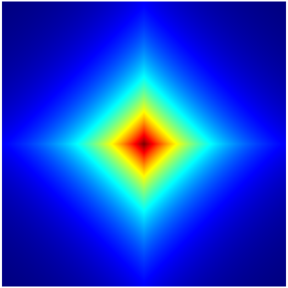
$$p_{post}(u|m) \propto \exp\left(-\frac{1}{2\,\sigma^2}\|m - A\,u\|_2^2 - \lambda\,|D\,u|_1\right)$$



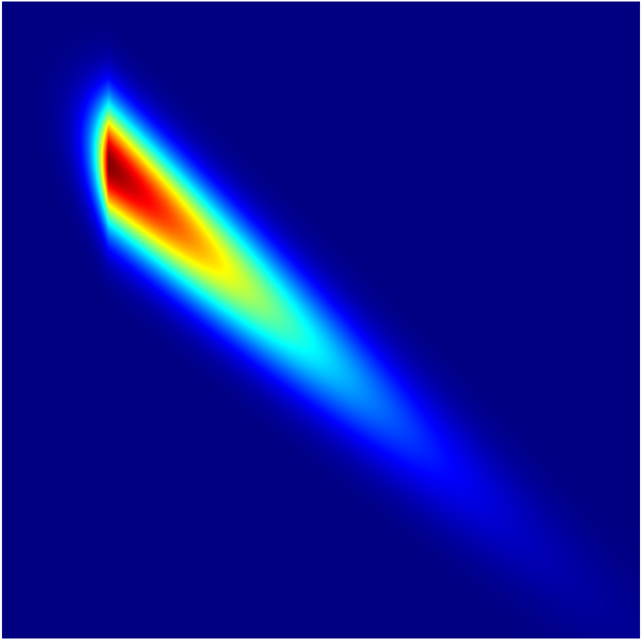
(a) $\exp\left(-\frac{1}{2}\|u\|_2^2\right)$ (b) $\exp\left(-|u|_1\right)$ (c) $1/(1+u^2)$



Likelihood:
$\exp\left(-\frac{1}{2\,\sigma^2}\|m - A\,u\|_2^2\right)$



Prior: $\exp\left(-\lambda\,|u|_1\right)$
(λ via discrepancy principle)



Posterior: $\exp\left(-\frac{1}{2\,\sigma^2}\|m - A\,u\|_2^2 - \lambda\,|u|_1\right)$

Sparsity Constraints in Inverse Problems

Direct correspondence to variational regularization by maximum a-posteriori-estimation (MAP) inference strategy:

$$
\begin{aligned}
\hat{u}_{\text{MAP}} &:= \underset{u \in \mathbb{R}^n}{\operatorname{argmax}} \; p_{post}(u|m) \\
&= \underset{u \in \mathbb{R}^n}{\operatorname{argmax}} \left\{ \exp\left(-\frac{1}{2\,\sigma^2} \|m - A\,u\|_2^2 - \lambda\, |D\,u|_1 \right) \right\} \\
&= \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \|m - A\,u\|_2^2 + 2\,\sigma^2 \lambda\, |D\,u|_1 \right\}
\end{aligned}
$$

$\Longrightarrow$ Properties of MAP estimate (e.g., *discretization invariance*) are well understood.

Sparsity Constraints in Inverse Problems

But there is more to Bayesian inference:

- Conditional mean-estimates (CM)
- Confidence intervals estimates
- Conditional covariance estimates
- Histogram estimates
- Generalized Bayes estimators
- Marginalization
- Model selection or averaging
- Experiment design

Influence of sparsity constraints on these quantities: Less well understood.

M. Lassas and S. Siltanen, 2004.
Can one use total variation prior for edge-preserving Bayesian inversion?

M. Lassas, E. Saksman, and S. Siltanen, 2009.
Discretization invariant Bayesian inversion and Besov space priors.

V. Kolehmainen, M. Lassas, K. Niinimäki, and S. Siltanen, 2012.
Sparsity-promoting Bayesian inversion.

Sparsity Constraints in Inverse Problems

Key issue

Examining L1-type priors might help to further understand the relation between variational regularization theory and Bayesian inference!

Key problem

Bayesian inference relies on computing integrals w.r.t. high-dim. posterior p_{post}.
Standard Monte Carlo integration techniques break down for ill-posed, high-dimensional problems with sparsity constraints.

This talks summarizes partial results from:

F. Lucka, 2012.
Fast MCMC sampling for sparse Bayesian inference in high-dimensional inverse problems using L1-type priors
submitted to Inverse Problems; arXiv:1206.0262v1

Monte Carlo Integration in a Nutshell

$$\mathbb{E}\left[f(x)\right] = \int_{\mathbb{R}^n} f(x)\, p(x)\ \mathrm{d}\, x$$

- *Traditional Gauss-type quadrature:*
  Construct suitable grid $\{x_i\}_i$, w.r.t $\omega(x) := p(x)$ and approximate by $\sum_{i=1}^{K} \omega_i f(x_i)$.
  $\Longrightarrow$ Grid construction and evaluation infeasible in high dimensions.

- *Monte Carlo integration idea:*
  Generate suitable grid $\{x_i\}_i$, w.r.t $p(x)$ by drawing $x_i \sim p(x)$ and approximate by $\frac{1}{K} \sum_{i=1}^{K} f(x_i)$. By the *Law of large numbers*:

$$\frac{1}{K} \sum_{i=1}^{K} f(x_i) \overset{K \rightarrow \infty}{\longrightarrow} \mathbb{E}_{p(x)}\left[f(x)\right] = \int_{\mathbb{R}^n} f(x)\, p(x)\ \mathrm{d}\, x$$

in L1 with rate $O(K^{-1/2})$ (independent of n).

Markov Chain Monte Carlo

Not able to draw independent samples?
$\rightsquigarrow$ With $\{x_i\}_i$ being an ergodic Markov chain, it still works!

Markov chain Monte Carlo (MCMC) methods are algorithms to construct such a chain:

- Huge number of MCMC methods exists.
- No "universal" method.
- Most methods rely on one two basic schemes:
  - Metropolis-Hastings (MH) Sampling [Metropolis et al., 1953; Hastings, 1970]
  - Gibbs Sampling [Geman & Geman, 1984]
- Posteriors from inverse problems seem to be "special".

In this talk: Comparison between the most basic variants of MH and Gibbs sampling for high-dimensional posteriors from inverse problem scenarios.

Felix Lucka (*felix.lucka@uni-muenster.de*)

Symmetric, Random-Walk Metropolis-Hastings Sampling

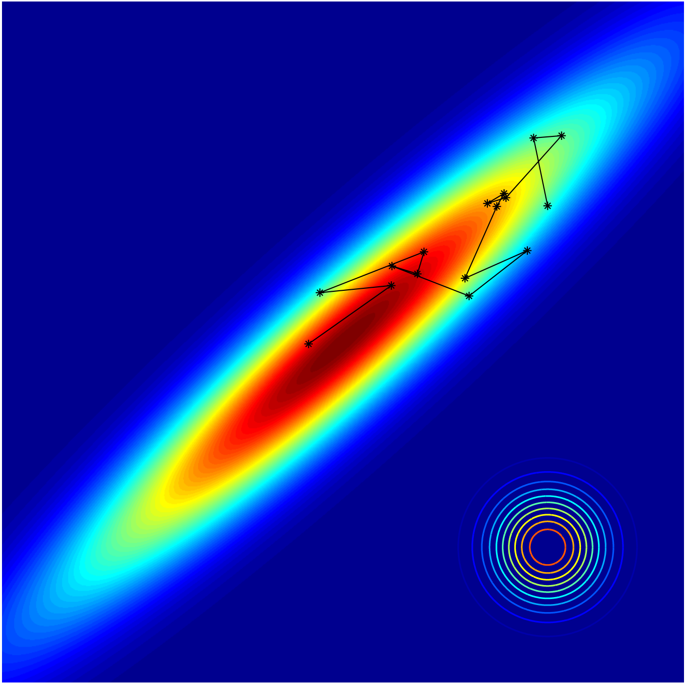
Given: Density $p(x), x \in \mathbb{R}^n$ to sample from.
Let $p_{pro}(z)$ be a symmetric density in $\mathbb{R}^n$ and $x_0 \in \mathbb{R}^n$ an initial state. Define burn-in size K_0 and sample size K.
For $i = 1,\ldots,K_0 + K$ do:

1. Draw z from $p_{pro}(z)$ and set $y = x_{i-1} + z$
2. Compute the acceptance ratio $r = \dfrac{p(y)}{p(x_{i-1})}$
3. Draw $\theta \in [0, 1]$ from a uniform probability density.
4. If $r \geqslant \theta$, set $x_i = y$, else set $x_i = x_{i-1}$.

Return $x_{K_0+1}, \ldots, x_K$.

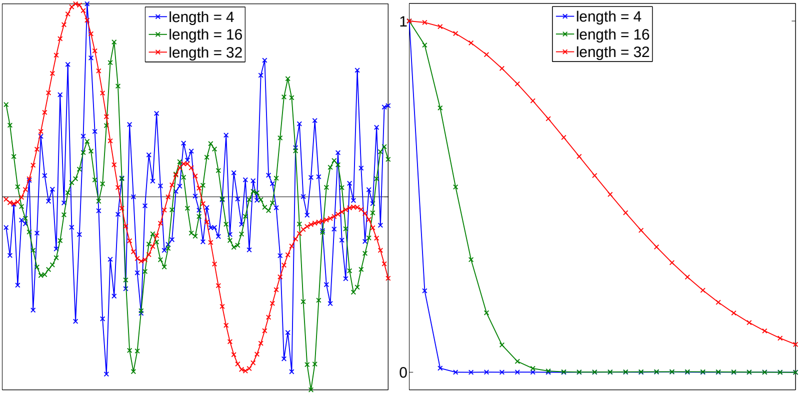
- Requires one evaluation of $p(x)$ and one sample from p_{pro} per step, no "real" knowledge about p is needed, not even normalization.
  $\rightsquigarrow$ "Black box" sampling algorithm.
- Most widely used.
- Good performance requires careful tuning of p_{pro}.
- Basis for very sophisticated sampling algorithms.
- Simulated annealing for the global optimization works in the same way.

In this talk:
$p_{pro} = \mathcal{N}(0, \kappa^2 \; I_n)$



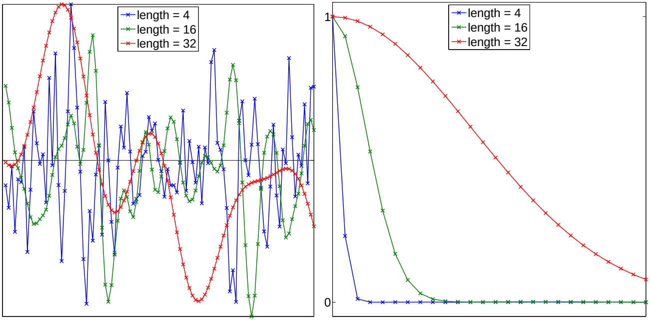
Evaluate performance of a sampler via its autocorrelation function (acf):
Let $g : \mathbb{R}^n \to \mathbb{R}^1$, and $g_i := g(u_i)$, $i = 1, \ldots, K$, then

$$R(\tau) := \frac{1}{(K-\tau)\hat{\varrho}} \sum_{i=1}^{K-\tau} (g_i - \hat{\mu})(g_{i+\tau} - \hat{\mu}) \qquad (\text{“}\mathit{lag}\text{-}\tau\ \mathit{ac}\ \text{w.r.t. } g\text{”})$$



(a) Stochastic processes... (b) ...and their autocorrelation functions

- A rapid decay of $R(\tau)$ means that samples get uncorrelated fast.
- Temporal acf (tacf): acf rescaled by computation time per sample, t_s: $R^*(t) := R(t/t_s)$ for all $t = i \cdot t_s, i \in \{0, \ldots, K-1\}$.
- Use $g(u) := \langle \nu_1, u \rangle$, where ν_1 is the largest eigenvector of the covariance matrix of $p(x)$ to test the "worst case".

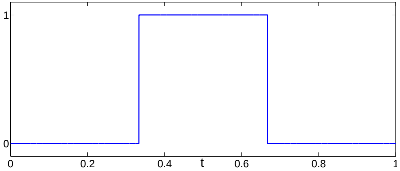


(c) Stochastic processes... (d) ...and their autocorrelation functions

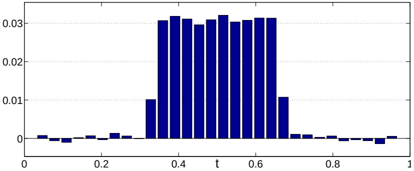
Total Variation Deblurring Example in 1D (from Lassas & Siltanen, 2004)

- Model of a charge coupled device (CCD) in 1D.
- Unknown light intensity $\tilde{u} : [0,1] \to \mathbb{R}^+$, indicator on $[\frac{1}{3}, \frac{2}{3}]$.
- Integrated into $k = 30$ CCD pixels $[\frac{1}{k+2}, \frac{k+1}{k+2}] \subset [0,1]$.
- Noise is added.
- $\tilde{u}$ is reconstructed on a regular, n-dim. grid.
- D is the forward finite difference operator with NB cond.

$$p_{post}(u|m) \propto \exp\left(-\frac{1}{2\,\sigma^2}\|m - A\,u\|_2^2 - \lambda\,|D\,u|_1\right)$$



(e) The unknown function $\tilde{u}(t)$



(f) The measurement data m

Total Variation Deblurring Example in 1D (from Lassas & Siltanen, 2004)

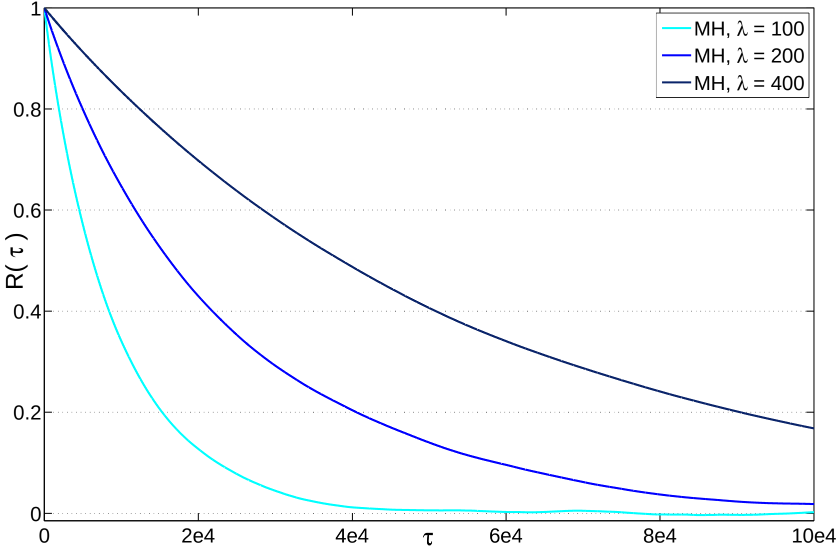


Figure: Autocorrelation plots $R(\tau)$ for MH Sampler and $n = 63$.

Total Variation Deblurring Example in 1D (from Lassas & Siltanen, 2004)

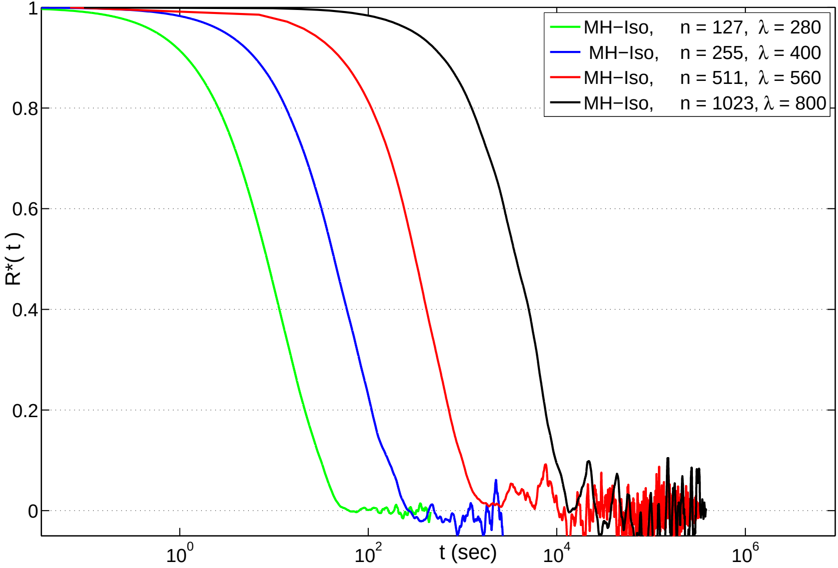


Figure: Temporal autocorrelation plots $R^*(t)$ for MH Sampler.

Total Variation Deblurring Example in 1D (from Lassas & Siltanen, 2004)

Results:

- Efficiency of MH samplers dramatically decreases when λ or n increase.
- Even for moderate n, most inference procedures become infeasible.

What else can we do?

- More sophisticated variants of MH sampling?
- Sample surrogate hyperparameter models?
- Try out the other basic scheme: Gibbs sampling.

Total Variation Deblurring Example in 1D (from Lassas & Siltanen, 2004)

Results:

- Efficiency of MH samplers dramatically decreases when λ or n increase.
- Even for moderate n, most inference procedures become infeasible.

What else can we do?

- More sophisticated variants of MH sampling?
- Sample surrogate hyperparameter models?
- Try out the other basic scheme: Gibbs sampling.

Single Component Gibbs Sampling

Given: Density $p(x), x \in \mathbb{R}^n$ to sample from.
Let $x_0 \in \mathbb{R}^n$ be an initial state. Define burn-in size K_0 and sample size K.
For $i = 1,\ldots,K_0 + K$ do:

1 Set $x_i := x_{i-1}$.

2 For $j = 1,\ldots,n$ do:

 (i) Draw s randomly from $\{1, \ldots, n\}$ (random scan).

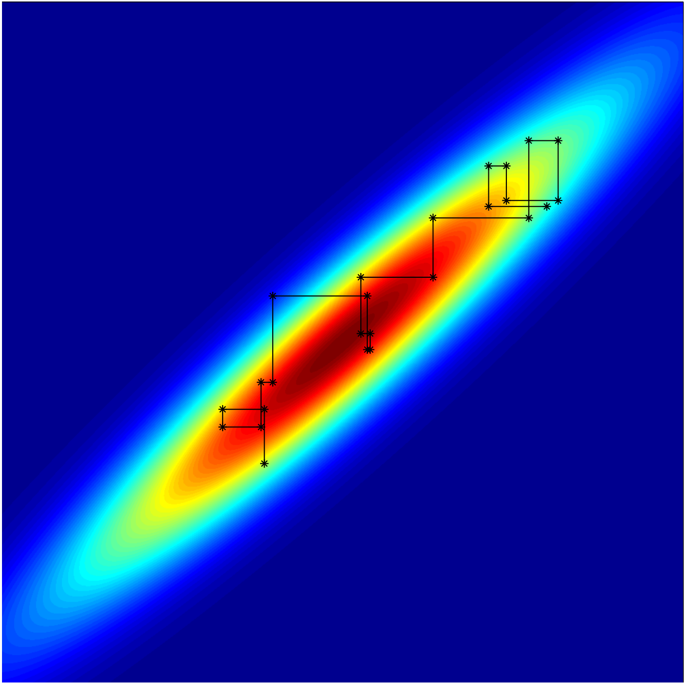
 (ii) Draw $(x_i)_s$ from the conditional, 1-dim density $p(\cdot|(x_i)_{[-s]})$.

Return $x_{K_0+1}, \ldots, x_K$.

In order to be fast one needs to be able

1. to compute the 1-dim distributions fast and explicit.
2. to sample from 1-dim distributions fast, robust and exact.

Point 2. turned out to be rather nasty, involved and time consuming to implement $\rightsquigarrow$ Details can be found in the paper.



Total Variation Deblurring Example in 1D (from Lassas & Siltanen, 2004)

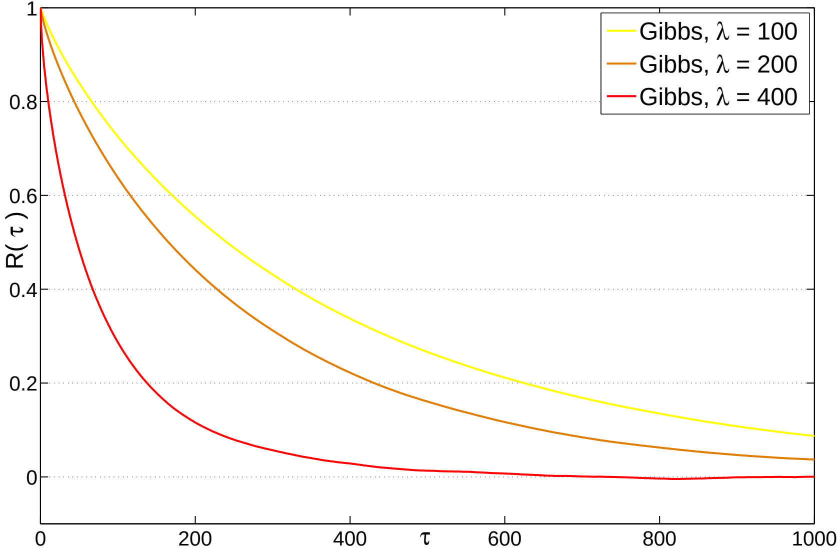


Figure: Autocorrelation plots $R(\tau)$ for Gibbs Sampler and $n = 63$.

Total Variation Deblurring Example in 1D (from Lassas & Siltanen, 2004)

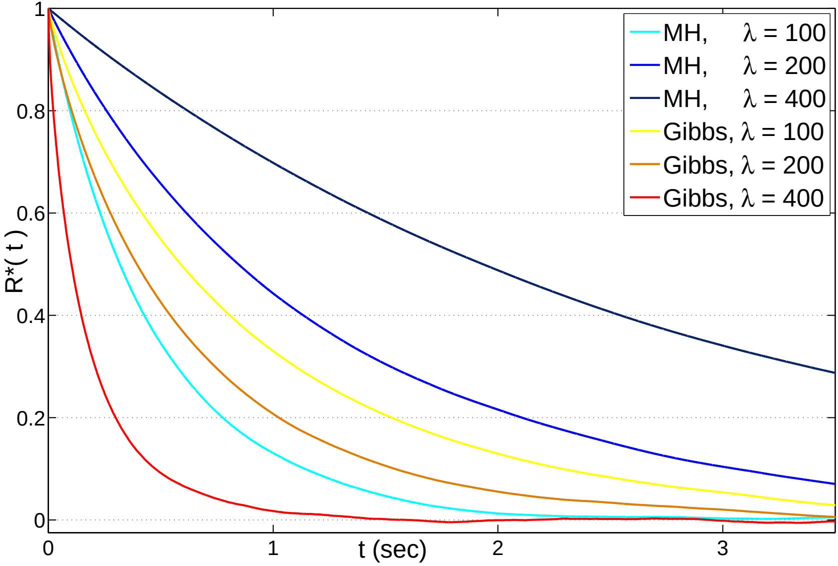


Figure: Temporal autocorrelation plots $R^*(t)$ for $n = 63$.

Total Variation Deblurring Example in 1D (from Lassas & Siltanen, 2004)

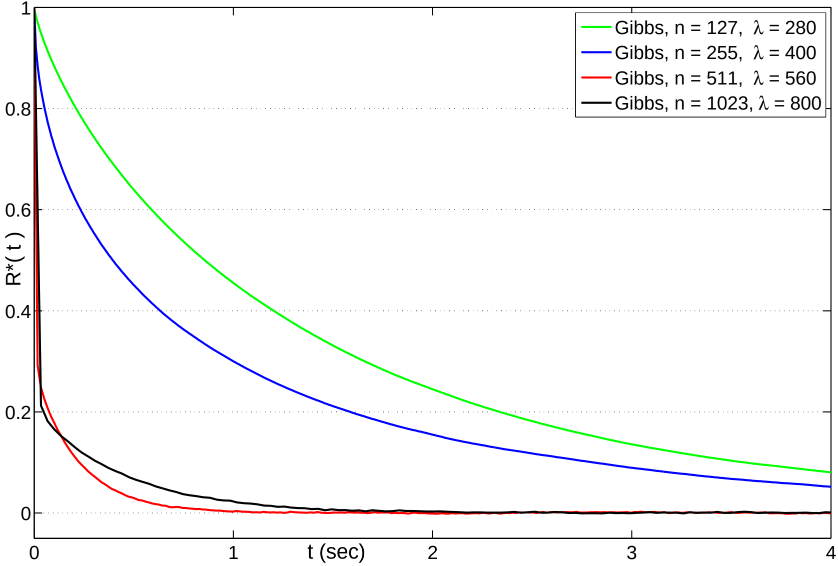


Figure: Temporal autocorrelation plots $R^*(t)$ for Gibbs Sampler

Total Variation Deblurring Example in 1D (from Lassas & Siltanen, 2004)

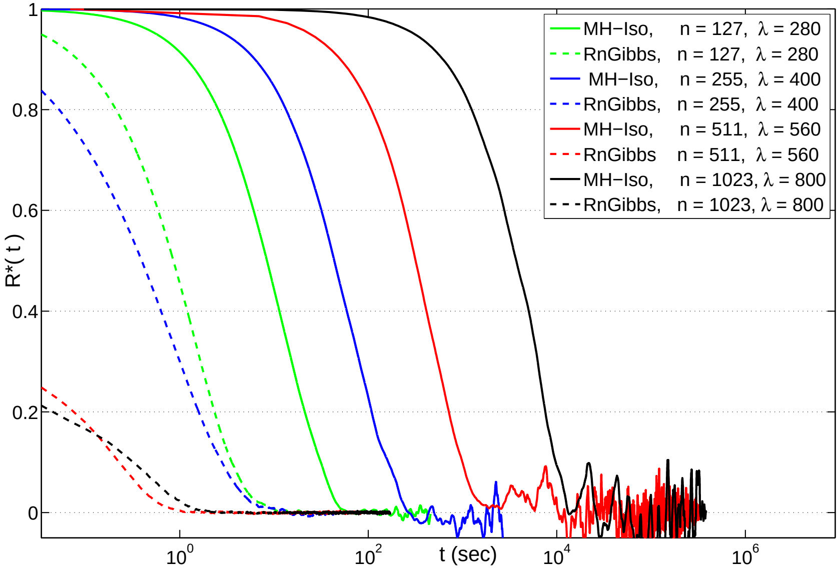


Figure: Temporal autocorrelation plots $R^*(t)$.

Total Variation Deblurring Example in 1D (from Lassas & Siltanen, 2004)

New sampler can be used to address theoretical questions:

- Lassas & Siltanen, 2004: For $\lambda_n \propto \sqrt{n+1}$, the TV prior converges to a smoothness prior in the limit $n \longrightarrow \infty$.
- MH sampling to compute CM estimate for $n = 63, 255, 1023, 4095$.
- Even after a month of computation time only partly satisfying results.

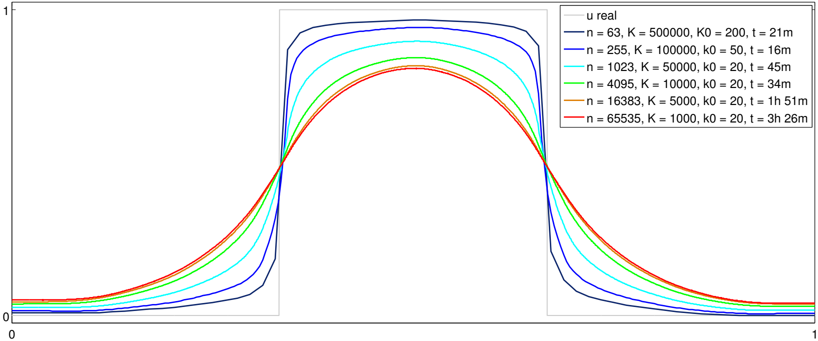
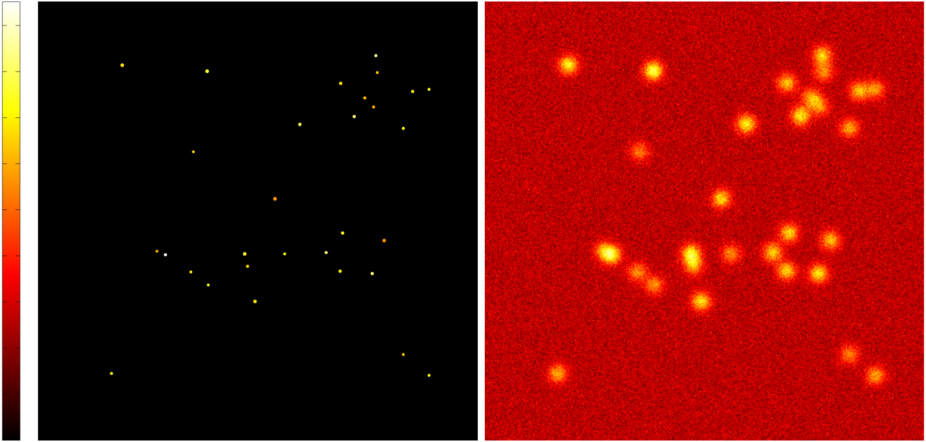


Figure: CM estimate computed for $n = 63, 255, 1023, 4095, 16383, 65535$ using Gibbs sampler on a comparable CPU.

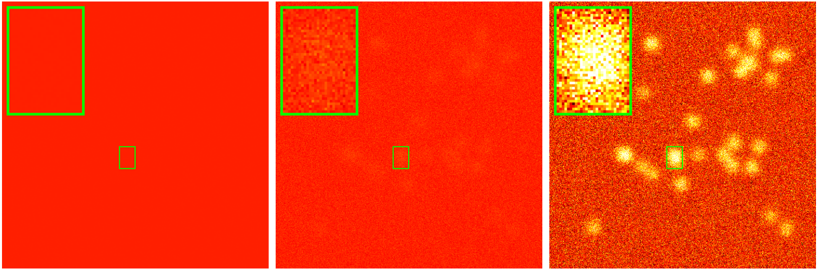
Image Deblurring Example in 2D



Unknown function $\tilde{u}$ Measurement data m

- Gaussian blurring kernel
- Relative noise level of 10%
- Reconstruction using $n = 511 \times 511 = 261\,121$.

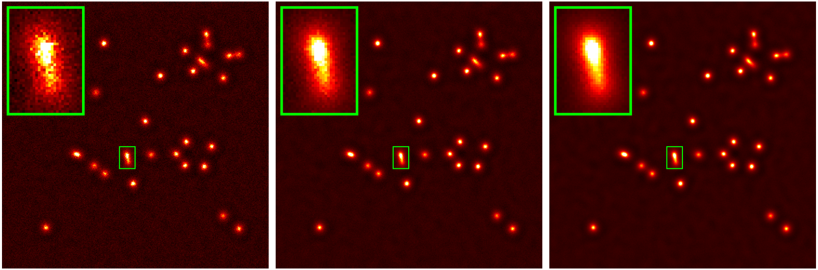
Image Deblurring Example in 2D



(a) 1h comp. time (b) 5h comp. time (c) 20h comp. time

Figure: CM estimates by MH sampler

Image Deblurring Example in 2D



(a) 1h comp. time (b) 5h comp. time (c) 20h comp. time

Figure: CM estimates by Gibbs sampler

Conclusions & Outlook

- MH is a "black-box sampler". It may fail dramatically in specific scenarios.
- But this is not a general feature of MCMC!
- Gibbs sampler incorporate more posterior-specific information into the sampling and perform way better.
- Promising results in dimensions larger than any previously reported use for L1-type inverse problems ($n > 1\,000\,000$ still works...).

$\Longrightarrow$ Results challenge common beliefs about MCMC in general.

Work to do:

- Real applications: Sparse tomography using Besov space priors like in [Kolehmainen, Lassas, Niinimäki, Siltanen, 2012]
- Tackle theoretical questions, e.g., of how stair-casing in TV can be seen from a Bayesian perspective.
- Comparison to more sophisticated variants of MH and Gibbs schemes.
- Generalization to arbitrary D in $|Du|_1$.

Thank you for your attention!

Full results and all details in:

F. Lucka , 2012.
Fast MCMC sampling for sparse Bayesian inference in high-dimensional inverse problems using L1-type priors
submitted to Inverse Problems; arXiv:1206.0262v1

- More sampling methods.
- 2D deblurring with $n = 511^2 = 261\,121$.
- Nasty details of the Gibbs sampler!
- Implementation and code.