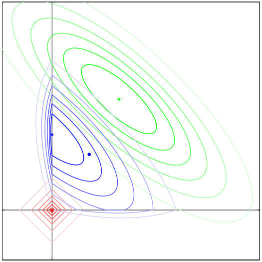
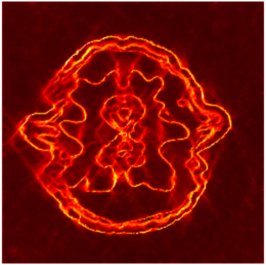
High-Dimensional Bayesian Inversion with Priors Far from Gaussians





Felix Lucka
University College London
f.lucka@ucl.ac.uk

Martin Burger
University of Münster
martin.burger@uni-muenster.de

Centre for Medical Image Computing

SIAM-UQ, April 6, 2016.

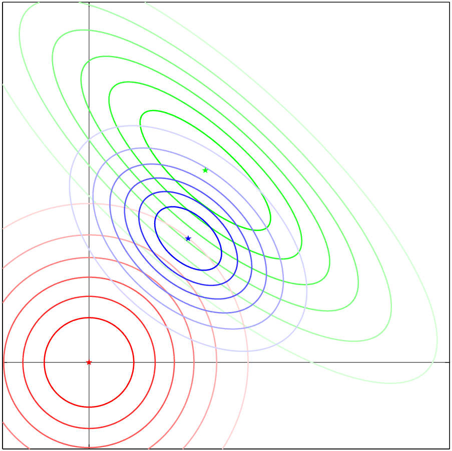
Linear ill-posed inverse problem with additive Gaussian noise:

$$\boxed{f = Au + \varepsilon}$$

$p_{like}(f|u) \propto$
$\exp\left(-\frac{1}{2}\|f - A\,u\|_2^2\right)$

$p_{prior}(u) \propto$
$\exp\left(-\lambda\,\|D^T u\|_2^2\right)$

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



Probabilistic representation allows for a rigorous quantification of the solution's uncertainties.

Bayesian Inference for Inverse Problems

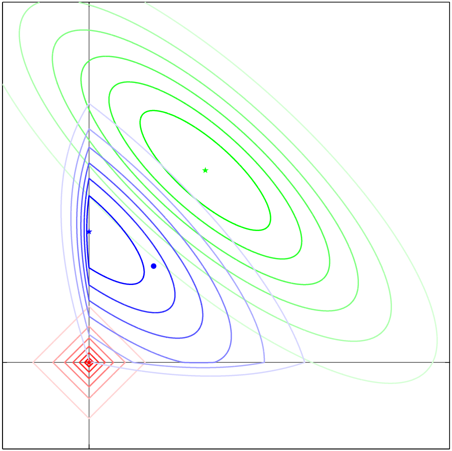
Linear ill-posed inverse problem with additive Gaussian noise:

$$\boxed{f = Au + \varepsilon}$$

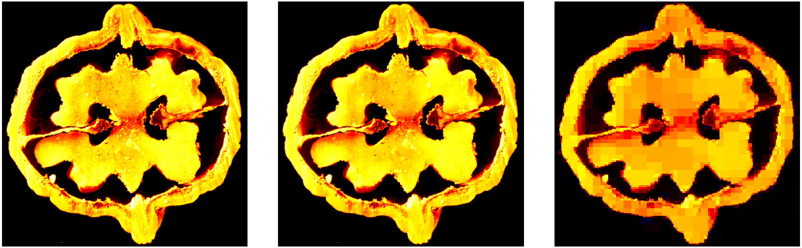
$p_{like}(f|u) \propto$
$\exp\left(-\frac{1}{2}\|f - A\,u\|_2^2\right)$

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

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



Probabilistic representation allows for a rigorous quantification of the solution's uncertainties.

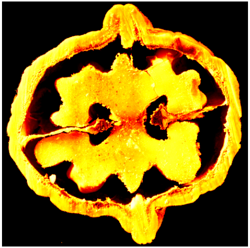


(a) 100% (b) 10% (c) 1%

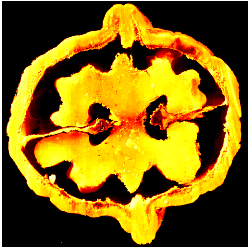
Sparsity a-priori constraints are used in variational regularization, compressed sensing and variable selection:

$$\hat{u}_\lambda \quad = \underset{u}{\operatorname{argmin}} \left\{ \tfrac{1}{2} \| f - A\, u \|_2^2 + \lambda \| D^T u \|_1 \right\}$$

(e.g. *total variation*, *wavelet shrinkage*, *LASSO*,...)



(a) 100%



(b) 10%

(c) 1%

Sparsity a-priori constraints are used in variational regularization, compressed sensing and variable selection:

$$\hat{u}_\lambda \quad = \operatorname*{argmin}_u \left\{ \tfrac{1}{2} \| f - A\, u \|_2^2 + \lambda \| D^T u \|_1 \right\}$$

(e.g. *total variation*, *wavelet shrinkage*, *LASSO*,...)

How about sparsity as a-priori information in the Bayesian approach?

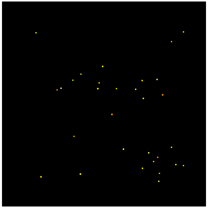
$$\boxed{p_{post}(u|f) \propto \exp\left(-\tfrac{1}{2}\|f - A\,u\|_2^2 - \lambda\,\|D^T u\|_1\right)}$$

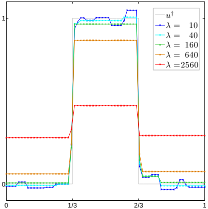
Aims: Bayesian inversion in high dimensions ($n \to \infty$):
MAP vs. CM, characterization of posterior structure.

Priors: Simple ℓ_1, total variation (TV), Besov space priors.

Starting points:

- **Lassas, Siltanen, 2004**. *Can one use total variation prior for edge-preserving Bayesian inversion?*, Inverse Problems, 20.
- **Lassas, Saksman, Siltanen, 2009**. *Discretization invariant Bayesian inversion and Besov space priors*, *Inverse Problems and Imaging*, 3(1).
- **Kolehmainen, Lassas, Niinimäki, Siltanen, 2012**. *Sparsity-promoting Bayesian inversion*, Inverse Problems, 28(2).





Efficient MCMC Techniques for ℓ_1 Priors

Task: Monte Carlo integration by samples from

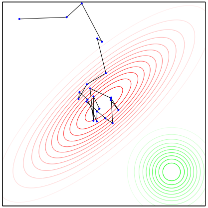
$$\boxed{p_{post}(u|f) \propto \exp\left(-\tfrac{1}{2}\|f - A\,u\|_2^2 - \lambda\,\|D^T u\|_1\right)}$$

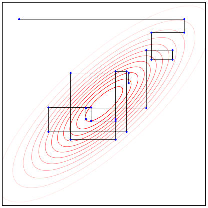
Problem: Standard Markov chain Monte Carlo (MCMC) sampler (Metropolis-Hastings) inefficient for large n or λ.

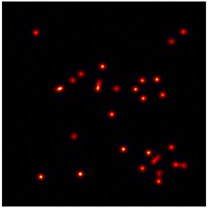
Contributions:

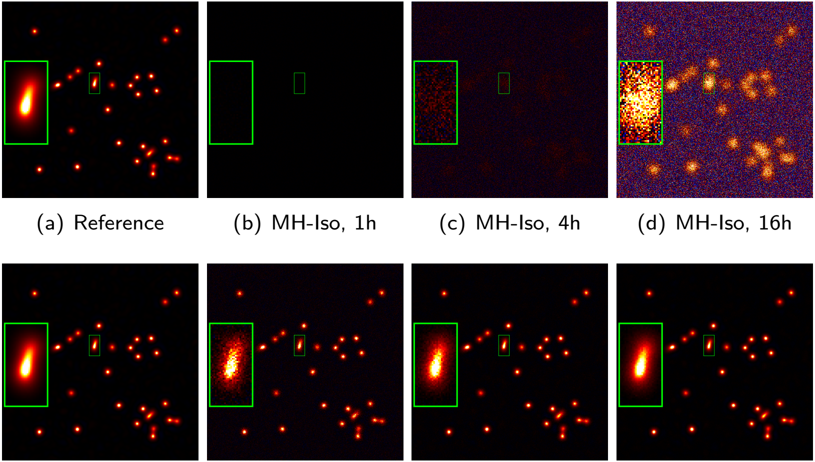
- Development of explicit single component Gibbs sampler.
- Tedious implementation for different scenarios.
- Still efficient in high dimensions ($n > 10^6$).
- Detailed evaluation and comparison to MH.

L, 2012. *Fast Markov chain Monte Carlo sampling for sparse Bayesian inference in high-dimensional inverse problems using L1-type priors*, Inverse Problems, 28(12):125012.







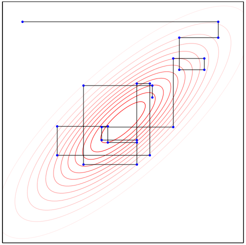


(a) Reference (b) MH-Iso, 1h (c) MH-Iso, 4h (d) MH-Iso, 16h

(e) Reference (f) SC Gibbs, 1h (g) SC Gibbs, 4h (h) SC Gibbs, 16h

Deconvolution, simple ℓ_1 prior, $n = 513 \times 513 = 263\,169$.

$$\boxed{p_{prior}(u) \propto \exp\left(-\lambda \|D^T u\|_1\right)}$$

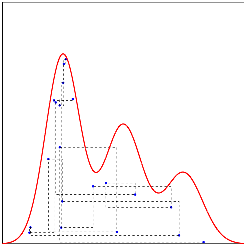


Limitations:

- D must be diagonalizable (synthesis priors):
- ℓ_p^q-prior: $\exp\left(-\lambda \|D^T u\|_p^q\right)$? TV in 2D/3D?
- Non-negativity or other hard-constraints?

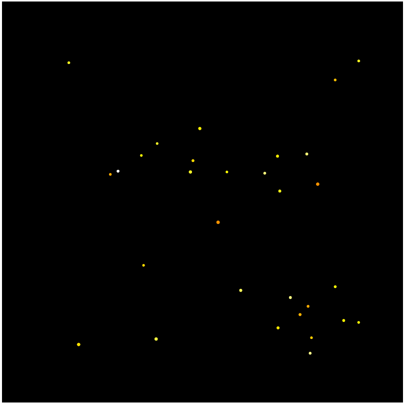
Contributions:

- Replace explicit by generalized slice sampling.
- Implementation & evaluation for most common priors.

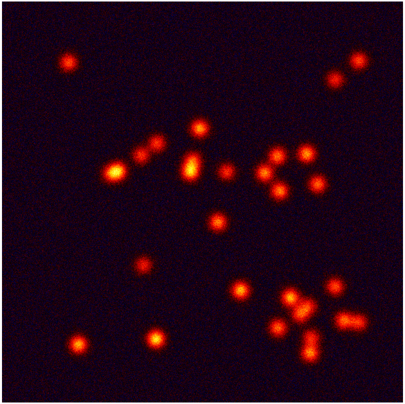


Neal, 2003. *Slice Sampling*, *Annals of Statistics* 31(3).

L, 2016. *Fast Gibbs sampling for high-dimensional Bayesian inversion*, *submitted*, arXiv:1602.08595.

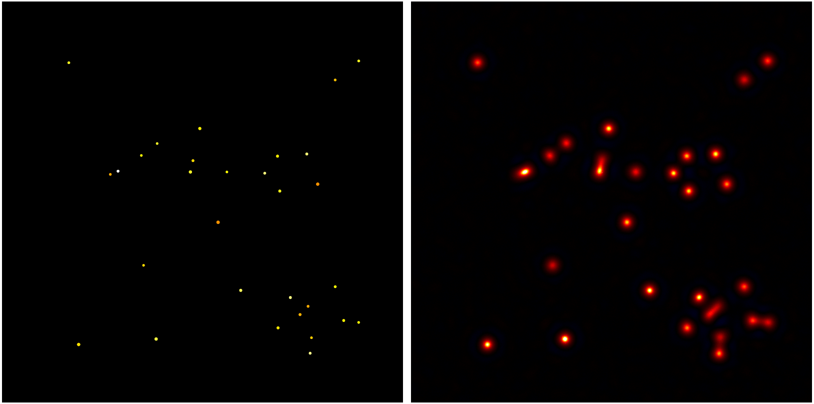


(a) Unknown function $\tilde{u}$



(b) Data f

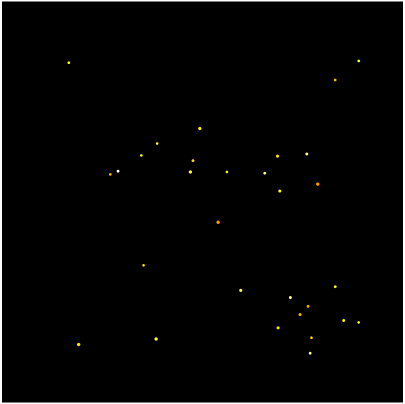
Deconvolution, simple ℓ_1 prior, $n = 1023 \times 1023 = 1\,046\,529$.



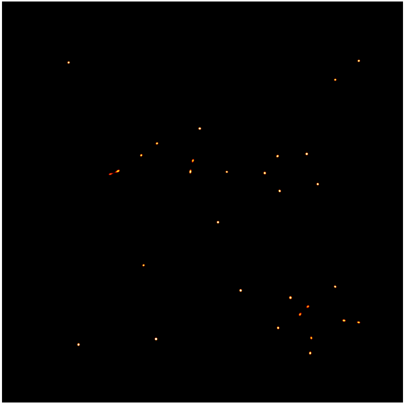
(a) Unknown function $\tilde{u}$ (b) CM estimate by our Gibbs sampler

Deconvolution, simple ℓ_1 prior, $n = 1023 \times 1023 = 1\,046\,529$.

Image Deblurring Example in 2D



(a) Unknown function $\tilde{u}$

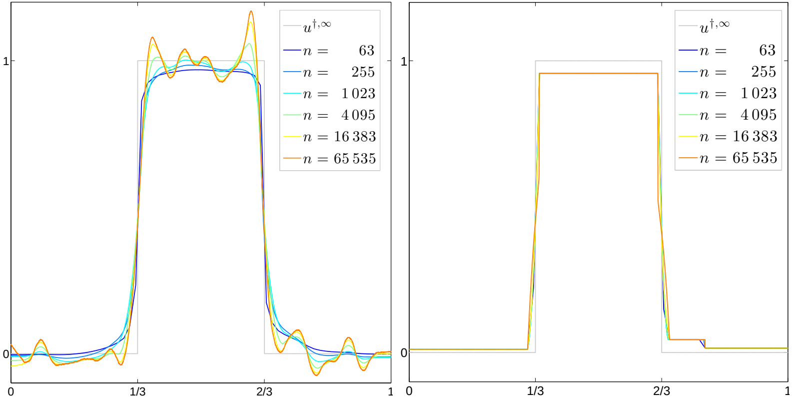


(b) MAP estimate by ADMM

Deconvolution, simple ℓ_1 prior, $n = 1023 \times 1023 = 1\,046\,529$.

"Can one use total variation prior for edge-preserving Bayesian inversion?"

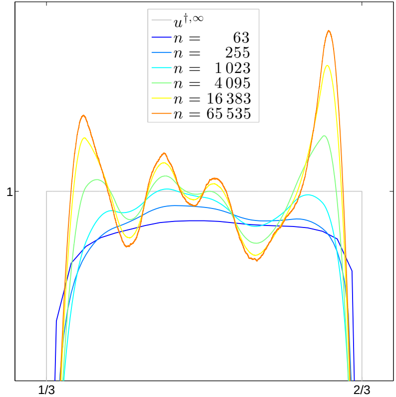
- For $\lambda_n = const.$ and $n \longrightarrow \infty$ the TV prior diverges.
- CM diverges.
- MAP converges to edge-preserving limit.



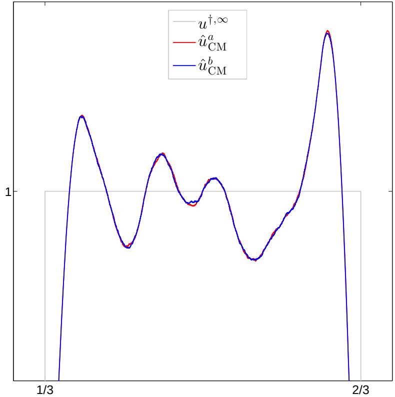
(a) CM by our Gibbs Sampler (b) MAP by ADMM

"Can one use total variation prior for edge-preserving Bayesian inversion?"

- For $\lambda_n = const.$ and $n \longrightarrow \infty$ the TV prior diverges.
- CM diverges.
- MAP converges to edge-preserving limit.



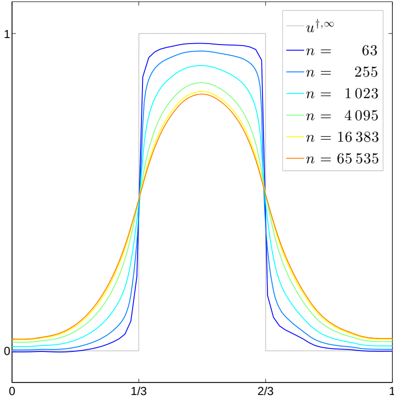
(a) Zoom into CM estimates



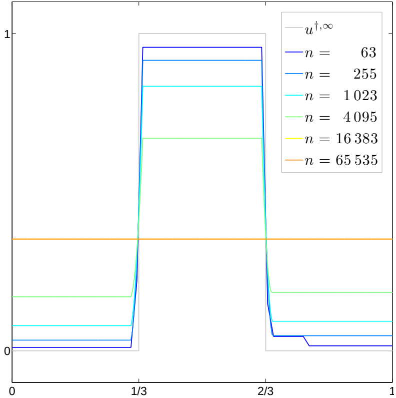
(b) MCMC convergence check

"Can one use total variation prior for edge-preserving Bayesian inversion?"

- For $\lambda_n \propto \sqrt{n+1}$ and $n \longrightarrow \infty$ the TV prior converges to a smoothness prior.
- CM converges to smooth limit.
- MAP converges to constant.

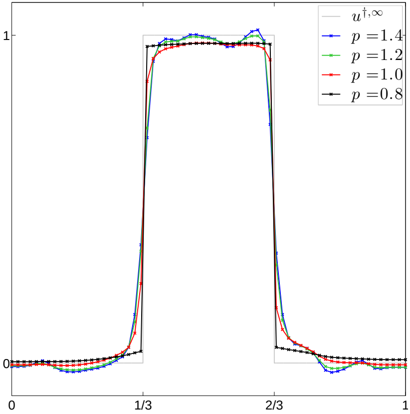


(a) CM by our Gibbs Sampler

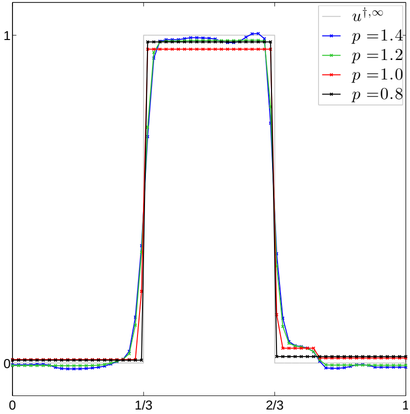


(b) MAP by ADMM

$$p_{post}(u) \propto \exp\left(-\frac{1}{2}\|f - A\,u\|^2_{\Sigma_\varepsilon^{-1}} - \lambda\,\|D^T u\|_p^p\right)$$



(a) CM (Gibbs-MCMC)



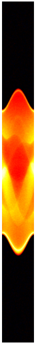
(b) MAP (Simulated Annealing)

For images dimensions > 1: No theory yet...but we can compute it.

Test scenario:

- CT using only 45 projection angles and 500 measurement pixel.

real solution



data f

colormap

For images dimensions > 1: No theory yet...but we can compute it.

MAP, $n = \ 64^2$, $\lambda = 500$

CM, $n = \ 64^2$, $\lambda = 500$

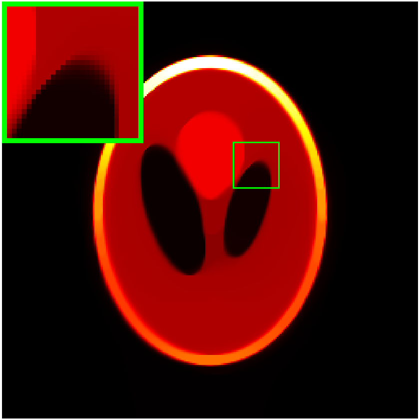
For images dimensions > 1: No theory yet...but we can compute it.

MAP, $n = 128^2$, $\lambda = 500$

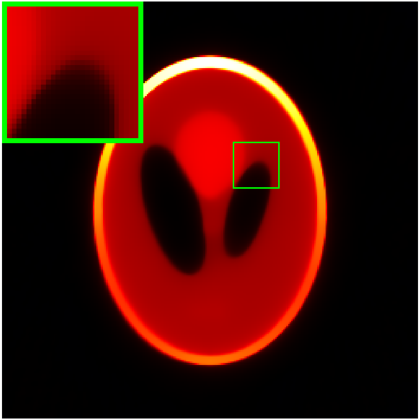
CM, $n = 128^2$, $\lambda = 500$

Need for New Theoretical Predictions

For images dimensions > 1: No theory yet...but we can compute it.



MAP, $n = 256^2$, $\lambda = 500$



CM, $n = 256^2$, $\lambda = 500$

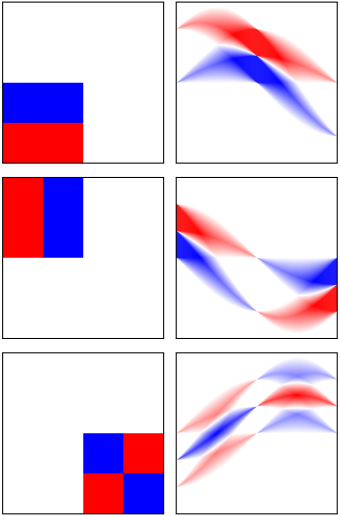
cf. Louchet, 2008, Louchet & Moisan, 2013 for the denoising case, $A = I$.

An ℓ_1-type, wavelet-based prior:

$$\boxed{p_{prior}(u) \propto \exp\left(-\lambda \|WV^T u\|_1\right)}$$

motivated by:

- **M. Lassas, E. Saksman, S. Siltanen, 2009**. *Discretization invariant Bayesian inversion and Besov space priors*, *Inverse Probl Imaging*, 3(1).
- **V. Kolehmainen, M. Lassas, K. Niinimäki, S. Siltanen, 2012**. *Sparsity-promoting Bayesian inversion*, *Inverse Probl*, 28(2).
- **K. Hämäläinen, A. Kallonen, V. Kolehmainen, M. Lassas, K. Niinimäki, S. Siltanen, 2013**. *Sparse Tomography*, *SIAM J Sci Comput*, 35(3).

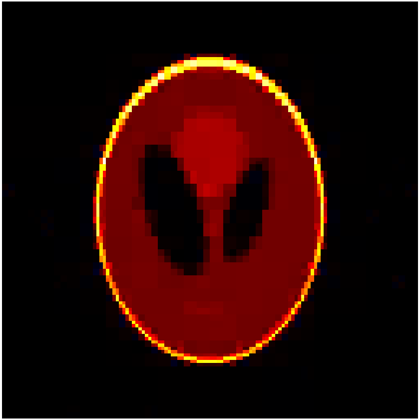


Reconstructions for $\lambda = 2e4$, $n = 64 \times 64 = 4.096$

MAP estimate (by ADMM)

CM estimate (by our Gibbs sampler)

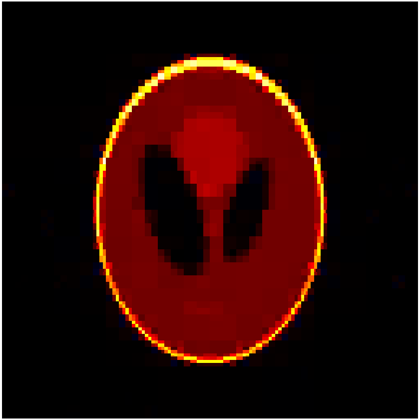
Reconstructions for $\lambda = 2e4$, $n = 128 \times 128 = 16.384$



MAP estimate (by ADMM)

CM estimate (by our Gibbs sampler)

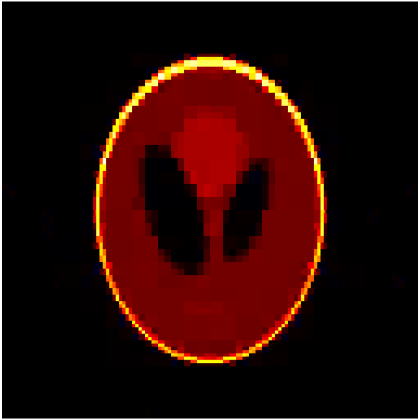
Reconstructions for $\lambda = 2e4$, $n = 256 \times 256 = 65.536$



MAP estimate (by ADMM)

CM estimate (by our Gibbs sampler)

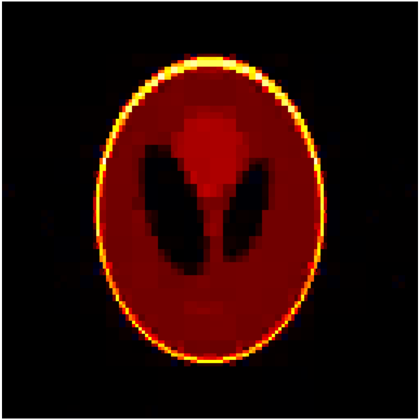
Reconstructions for $\lambda = 2e4$, $n = 512{\times}512 = 262.144$



MAP estimate (by ADMM)

CM estimate (by our Gibbs sampler)

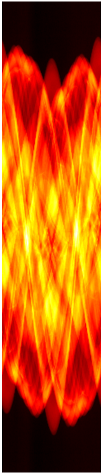
Reconstructions for $\lambda = 2e4$, $n = 1024 \times 1024 = 1.048.576$



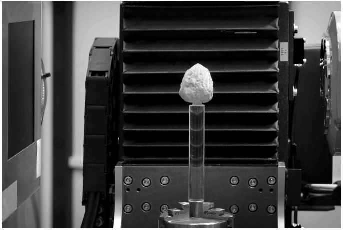
MAP estimate (by ADMM)

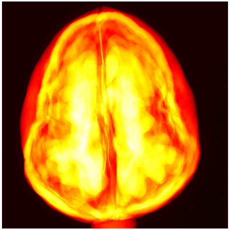
CM estimate (by our Gibbs sampler)

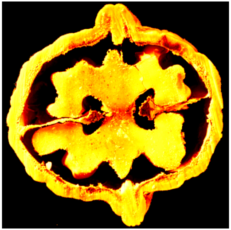
Application to Experimental Data: Walnut-CT



- Cooperation with Samuli Siltanen, Esa Niemi et al.
- Implementation of MCMC methods for Fanbeam-CT.
- Besov and TV prior; non-negativity constraints.
- Stochastic noise modeling.
- Bayesian perspective on limited angle CT.

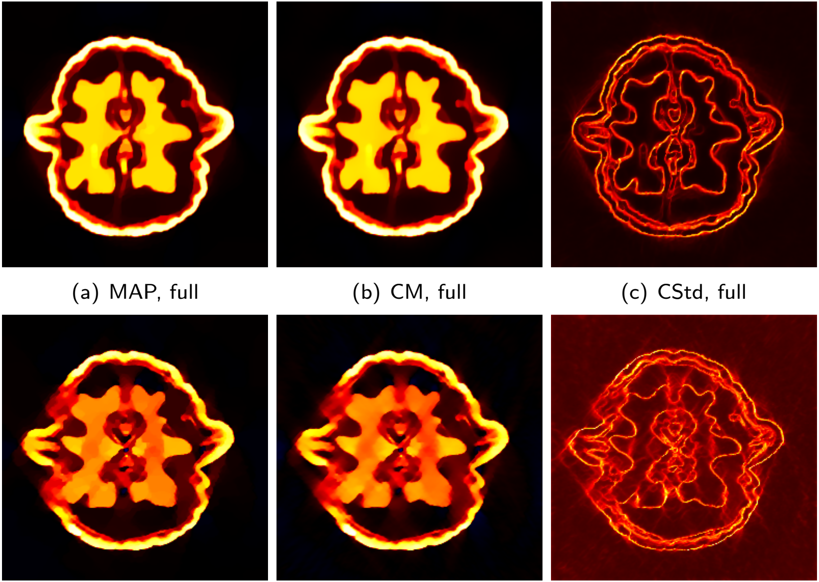






Use the data set for your own work:
http://www.fips.fi/dataset.php (documentation: arXiv:1502.04064)

Walnut-CT with TV Prior: Full vs. Limited Angle



(a) MAP, full

(b) CM, full

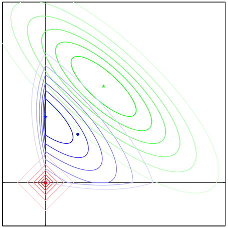
(c) CStd, full

(d) MAP, limited

(e) CM, limited

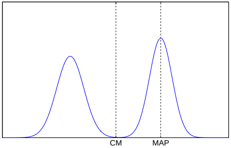
(f) CStd, limited

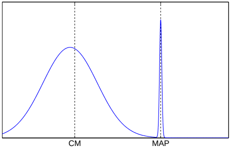
$$\hat{u}_{\text{MAP}} := \underset{u \in \mathbb{R}^n}{\operatorname{argmax}} \{ \, p_{post}(u|f) \} \quad \text{OR} \quad \hat{u}_{\text{CM}} := \int u \; p_{post}(u|f) \, \mathrm{d}u$$



Observations...

- Gaussian priors: MAP = CM. Funny coincidence?
- For reasonable non-Gaussian priors, MAP are sparser, sharper, look and perform better...
- If the CM looks good, it looks like the MAP.
- UQ wrt the CM (= variance) might not be interesting.
- Gribonval, Marchart, Louchet and Moisan, 2011-2013: CM are MAP for different priors.





...are in contradiction with classical Bayes cost formalism which discriminates MAP (= variational regularization) and advocates CM.

- An estimator is a random variable, as it relies on f and u.
- How does it perform on average? Which estimator is "best"?
- $\rightsquigarrow$ Define a cost function $\Psi(u, v)$.
- Bayes cost is the expected cost:

$$BC(\hat{u}) = \iint \Psi(u, \hat{u}(f))\, p_{like}(f|u)\,\mathrm{d}f\; p_{prior}(u)\,\mathrm{d}u$$

- Bayes estimator $\hat{u}_{BC}$ for given Ψ minimizes Bayes cost. Turns out:

$$\hat{u}_{BC}(f) = \operatorname*{argmin}_{\hat{u}} \left\{ \int \Psi(u, \hat{u}(f))\, p_{post}(u|f)\,\mathrm{d}u \right\}$$

The Classical Bayes Cost Argument

- CM is Bayes estimator for $\Psi(u, \hat{u}) = \|u - \hat{u}\|_2^2$ (MSE).
- Also the minimum variance estimator.
- The mean value is the intuitive "average", the "center of mass".
- MAP is asymptotic Bayes estimator of

$$\Psi_\epsilon(u, \hat{u}) = \begin{cases} 0, & \text{if} \quad \|u - \hat{u}\|_\infty \leqslant \epsilon \\ 1 & \text{otherwise,} \end{cases}$$

  for $\epsilon \to 0$ (uniform cost). $\Longrightarrow$ Not a proper Bayes estimator.

MAP and CM seem fundamentally different $\Longrightarrow$ one should decide!

- *"A real Bayesian would not use the MAP estimate"*
- People feel "ashamed" when they have to compute MAP estimates (even when their results are good).

A False Conclusion

"A real Bayesian would not use the MAP estimate as it is not a proper Bayes estimator".

"MAP estimate can be seen as an asymptotic Bayes estimator of

$$\Psi_\epsilon(u, \hat{u}) = \begin{cases} 0, & \text{if} \quad \|u - \hat{u}\|_\infty < \epsilon \\ 1 & \text{otherwise,} \end{cases}$$

for $\epsilon \to 0$. $\implies$ It is not a proper Bayes estimator."

"A real Bayesian would not use the MAP estimate as it is not a proper Bayes estimator".

"MAP estimate can be seen as an asymptotic Bayes estimator of

$$\Psi_\epsilon(u, \hat{u}) = \begin{cases} 0, & \text{if} \quad \|u - \hat{u}\|_\infty < \epsilon \\ 1 & \text{otherwise,} \end{cases}$$

for $\epsilon \rightarrow 0$. ???$\Longrightarrow$??? It is not a proper Bayes estimator."

A False Conclusion

"A real Bayesian would not use the MAP estimate as it is not a proper Bayes estimator".

"MAP estimate can be seen as an asymptotic Bayes estimator of

$$\Psi_\epsilon(u, \hat{u}) = \begin{cases} 0, & \text{if} \quad \|u - \hat{u}\|_\infty < \epsilon \\ 1 & \text{otherwise,} \end{cases}$$

for $\epsilon \to 0$. ???$\Longrightarrow$??? It is not a proper Bayes estimator."

"MAP estimator is asymptotic Bayes estimator for some degenerate Ψ"
$\not\Rightarrow$ "MAP can't be Bayes estimator for some proper Ψ" !!!!

A False Conclusion

“A real Bayesian would not use the MAP estimate as it is not a proper Bayes estimator”.

”MAP estimate can be seen as an asymptotic Bayes estimator of

$$\Psi_\epsilon(u, \hat{u}) = \begin{cases} 0, & \text{if} \quad \|u - \hat{u}\|_\infty < \epsilon \\ 1 & \text{otherwise,} \end{cases}$$

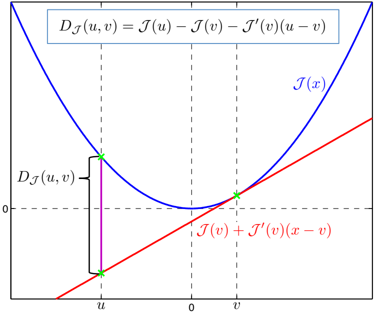
for $\epsilon \rightarrow 0$. ???$\Longrightarrow$??? It is not a proper Bayes estimator.”

”MAP estimator is asymptotic Bayes estimator for some degenerate Ψ”
$\not\Rightarrow$ “MAP can't be Bayes estimator for some proper Ψ” !!!!

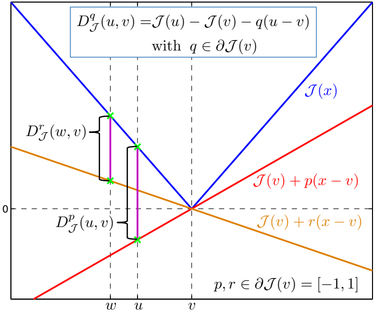
We need new cost functions!

For a proper, convex functional $\mathcal{J} : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{\infty\}$, the Bregman distance $D_{\mathcal{J}}^p(u, v)$ between $u, v \in \mathbb{R}^n$ for a subgradient $p \in \partial\mathcal{J}(v)$ is defined as

$$D_{\mathcal{J}}^p(u, v) = \mathcal{J}(u) - \mathcal{J}(v) - \langle p, u - v \rangle, \qquad p \in \partial\mathcal{J}(v)$$



(g) $\mathcal{J}(x) = x^2$



(h) $\mathcal{J}(x) = |x|$

Basically, $D_{\mathcal{J}}(u, v)$ measures the difference between $\mathcal{J}$ and its linearization in v at another point u.

$$\boxed{p_{post}(u|f) \propto \exp\left(-\tfrac{1}{2}\|f - A\,u\|_2^2 - \lambda\,\mathcal{J}(u)\right)}$$

with $\mathcal{J}$ proper, convex (prior is log-concave).

Definition:

(a) $\Psi_{\mathrm{LS}}(u, \hat{u}) := \|A(\hat{u} - u)\|_2^2 + \beta\|L(\hat{u} - u)\|_2^2$

(b) $\Psi_{\mathrm{Brg}}(u, \hat{u}) := \|A(\hat{u} - u)\|_2^2 + 2\lambda D_{\mathcal{J}}(\hat{u}, u)$

for a regular L, $\beta > 0$.

Properties:

- Proper, convex cost functions
- For $\mathcal{J}(u) = \beta/\lambda\|Lu\|_2^2$ (Gaussian case!) we have $\lambda D_{\mathcal{J}}(\hat{u}, u) = \beta\|L(\hat{u} - u)\|_2^2$, and $\Psi_{\mathrm{LS}}(u, \hat{u}) = \Psi_{\mathrm{Brg}}(u, \hat{u})$!

Theorems:

(I) The CM estimate is the Bayes estimator for $\Psi_{\mathrm{LS}}(u, \hat{u})$

(II) The MAP estimate is the Bayes estimator for $\Psi_{\mathrm{Brg}}(u, \hat{u})$

$$p_{post}(u|f) \propto \exp\left(-\tfrac{1}{2}\|f - A\,u\|_2^2 - \lambda\,\mathcal{J}(u)\right)$$

Definition:

(a) $\Psi_{\mathrm{LS}}(u,\hat{u}) := \|A(\hat{u}-u)\|_2^2 + \beta\|L(\hat{u}-u)\|_2^2$

(b) $\Psi_{\mathrm{Brg}}(u,\hat{u}) := \|A(\hat{u}-u)\|_2^2 + 2\lambda D_{\mathcal{J}}(\hat{u},u)$

for a regular L, $\beta > 0$.

Theorems:

(I) The CM estimate is the Bayes estimator for $\Psi_{\mathrm{LS}}(u,\hat{u})$

(II) The MAP estimate is the Bayes estimator for $\Psi_{\mathrm{Brg}}(u,\hat{u})$

Non-Gaussian case:

- $\mathrm{dom}(\mathcal{J})$ usually defines a (subset of a) Banach space for $n \to \infty$.
- In such a space: No natural Hilbert space norm as limit of $\|Lu\|^2$.
- Hilbert space norm not meaningful measure, e.g. for functions in BV.
- Only choice: $L = 0 \Longrightarrow \Psi_{\mathrm{LS}}$ only measures in output space, bad for ill-posed inverse problems!

Average optimality condition for the CM estimate:

$$A^*(A\hat{u}_{\text{CM}} - f) + \lambda \hat{p}_{\text{CM}} = 0, \qquad \hat{p}_{\text{CM}} = \int \mathcal{J}'(u) p_{post}(u|f) \mathrm{d}u$$

$$A^*(A\hat{u}_{\text{MAP}} - f) + \lambda \hat{p}_{\text{MAP}} = 0, \qquad \hat{p}_{\text{MAP}} = \mathcal{J}'(\hat{u}_{\text{MAP}})$$

Difference: $\mathcal{J}'(\mathbb{E}_{(u|f)}[u]) \neq \mathbb{E}_{(u|f)}[\mathcal{J}'(u)]$ (except for Gaussian prior).

"The posterior is well centered around the CM but not around the MAP estimate."

$\Longrightarrow$ Use optimality condition to rewrite posterior in terms of $\hat{u}_{\text{MAP}}$:

$$p_{post}(u|f) \propto \exp\left(-\frac{1}{2}\|A(u - \hat{u}_{\text{MAP}})\|_2^2 - \lambda D_{\mathcal{J}}^{\hat{p}_{\text{MAP}}}(u, \hat{u}_{\text{MAP}})\right)$$

Posterior energy is sum of two convex functionals both minimized by $\hat{u}_{\text{MAP}}$.

Two new inequalities,

$$\mathbb{E}_{(u|f)} \|L(\hat{u}_{\text{CM}} - u)\|_2^2 \leqslant \mathbb{E}_{(u|f)} \|L(\hat{u}_{\text{MAP}} - u)\|_2^2$$
$$\mathbb{E}_{(u|f)} D_{\mathcal{J}}(\hat{u}_{\text{MAP}}, u) \leqslant \mathbb{E}_{(u|f)} D_{\mathcal{J}}(\hat{u}_{\text{CM}}, u)$$

indicate that the use of anisotropic priors calls for different uncertainty measures than variance or mean square risks.

References:

- **M. Burger, F.L., 2014**. *Maximum a posteriori estimates in linear inverse problems with log-concave priors are proper Bayes estimators*, *Inverse Problems, 30(11):114004.*
- **T. Helin, M. Burger, 2015**. *Maximum a posteriori probability estimates in infinite-dimensional Bayesian inverse problems*, *Inverse Problems, 31(8):085009.*

Bayesian Modeling:

- Modeling sparsity with ℓ_1 priors can fail: Sometimes, only the MAP is sparse, nothing else.
- Alternatives include hierarchical Bayesian models and spike-and-slab priors.

Bayesian Computation:

- Elementary MCMC samplers may perform very differently.
- Contrary to common beliefs sample-based Bayesian inversion in high dimensions ($n > 10^6$) is feasible if tailored samplers are developed.
- Reason for the efficiency of the Gibbs samplers is unclear.

Bayesian Estimation / Uncertainty Quantification

- MAP estimates are proper Bayes estimators, minimizing a cost function potentially better suited to asymptotic Banach space structure.
- But: Everything beyond "MAP or CM?" is far more interesting and can really complement variational approaches.
- However: Extracting information from posterior samples is a non-trivial (future research) topic.
- The anisotropic structure of the priors calls for different uncertainty measures than variance or mean square risks.
- Bregman distances are interesting tools for Bayesian inversion.

L, 2016. *Fast Gibbs sampling for high-dimensional Bayesian inversion.* submitted, arXiv:1602.08595

L., 2014. *Bayesian Inversion in Biomedical Imaging* PhD Thesis, University of Münster.

M. Burger, L., 2014. *Maximum a posteriori estimates in linear inverse problems with log-concave priors are proper Bayes estimators* *Inverse Problems*, 30(11):114004.

L., 2012. *Fast Markov chain Monte Carlo sampling for sparse Bayesian inference in high-dimensional inverse problems using L1-type priors.* *Inverse Problems*, 28(12):125012.

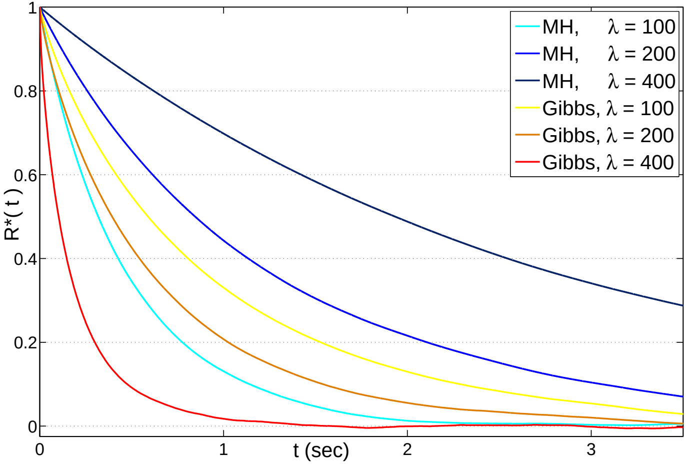
Thank you for your attention!

L, 2016. *Fast Gibbs sampling for high-dimensional Bayesian inversion.* submitted, arXiv:1602.08595

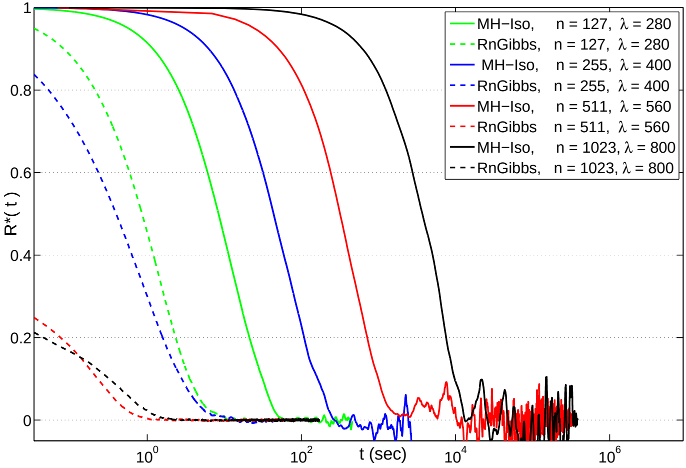
L., 2014. *Bayesian Inversion in Biomedical Imaging* PhD Thesis, University of Münster.

M. Burger, L., 2014. *Maximum a posteriori estimates in linear inverse problems with log-concave priors are proper Bayes estimators* *Inverse Problems*, 30(11):114004.

L., 2012. *Fast Markov chain Monte Carlo sampling for sparse Bayesian inference in high-dimensional inverse problems using L1-type priors.* *Inverse Problems*, 28(12):125012.



Temporal autocorrelation $R^*(t)$ for 1D TV-deblurring, $n = 63$.



Temporal autocorrelation $R^*(t)$ for 1D TV-deblurring.