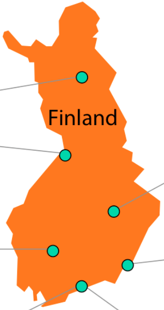
Sparsity-promoting Bayesian inversion

Samuli Siltanen

Department of Mathematics and Statistics
University of Helsinki

Inverse Problems and Optimal Control for PDEs
University of Warwick, May 23–27, 2011

Finnish Centre of Excellence in Inverse Problems Research



OULUN YLIOPISTO UNIVERSITY of OULU

UNIVERSITY OF EASTERN FINLAND

TAMPERE UNIVERSITY OF TECHNOLOGY

Open your mind. LUT.
Lappeenranta University of Technology

Aalto University

UNIVERSITY OF HELSINKI

http://wiki.helsinki.fi/display/inverse/Home

This is a joint work with

Ville Kolehmainen, University of Eastern Finland

Matti Lassas, University of Helsinki

Kati Niinimäki, University of Eastern Finland

Eero Saksman, University of Helsinki

Outline

Outline

Our starting point is the *continuum model* for an indirect measurement

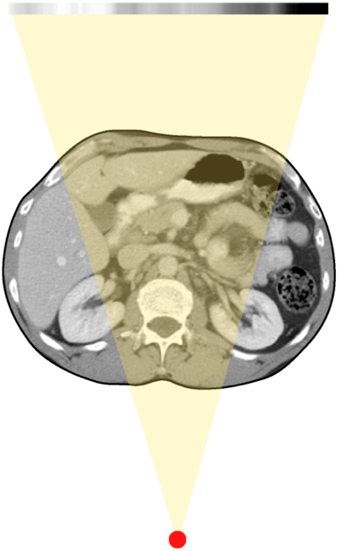
Consider a quantity U that can be observed indirectly by

$$M = AU + \mathcal{E},$$

where A is a smoothing linear operator, $\mathcal{E}$ is white noise, and U and M are functions.

In the Bayesian framework, $U = U(x, \omega)$, $M = M(z, \omega)$ and $\mathcal{E} = \mathcal{E}(z, \omega)$ are taken to be random functions.

X-ray tomography: Continuum model



$$M = AU + \mathcal{E}$$

A measurement device gives only a finite number of data, leading to *practical measurement model*

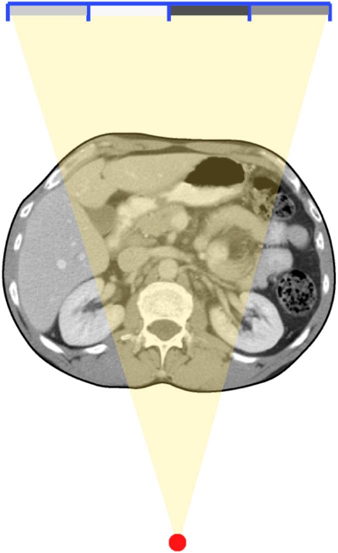
The measurement is a realization of the random variable

$$M_k = P_k M = A_k U + \mathcal{E}_k,$$

where $A_k = P_k A$ and $\mathcal{E}_k = P_k \mathcal{E}$. Here P_k is a linear orthogonal projection with k-dimensional range.

The given data is a realization $\widehat{m}_k = M_k(z, \omega_0)$ of the measurement $M_k(z, \omega)$, where $\omega_0 \in \Omega$ is a specific element in the probability space.

X-ray tomography: Practical measurement model



$$M_k = A_k U + \mathcal{E}_k$$

The inverse problem

This study concentrates on the inverse problem

$$\text{given a realization } \widehat{m}_k, \text{ estimate } U,$$

using estimates and confidence intervals related to a Bayesian posterior probability distribution.

Numerical solution of the inverse problem is based on the discrete *computational model*

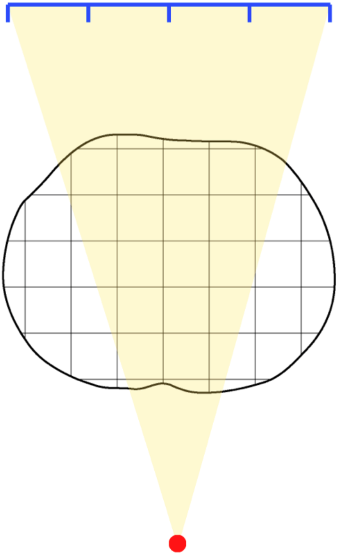
We need to discretize the unknown function U. Assume that U is *a priori* known to take values in a function space Y.

Choose a linear projection $T_n : Y \to Y$ with n-dimensional range Y_n, and define a random variable $U_n := T_n U$ taking values in Y_n. This leads to the *computational model*

$$M_{kn} = A_k U_n + \mathcal{E}_k.$$

Note that realizations of M_{kn} can be simulated by computer but cannot be measured in reality.

X-ray tomography: Computational model

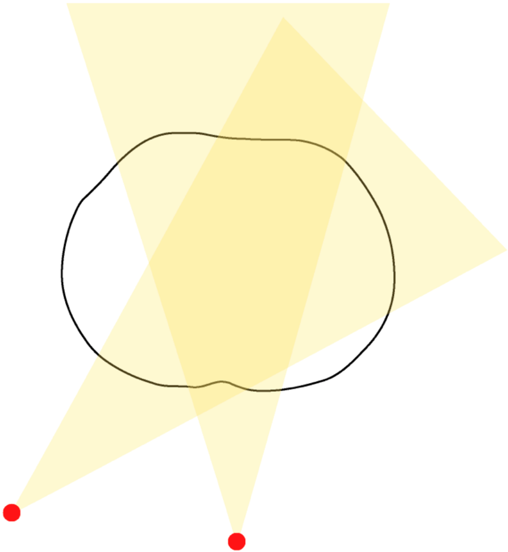


$$M_{kn} = A_k U_n + \mathcal{E}_k$$

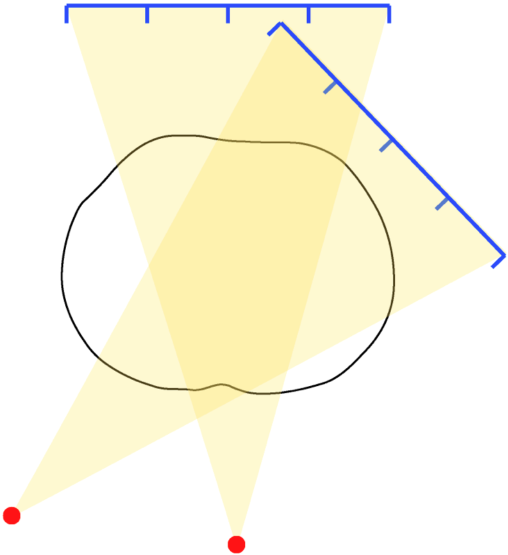
The computational model $M_{kn} = A_k U_n + \mathcal{E}_k$ involves two *independent* discretizations:

P_k is related to the measurement device, and
T_n is related to the finite representation of U.

The numbers k and n are independent

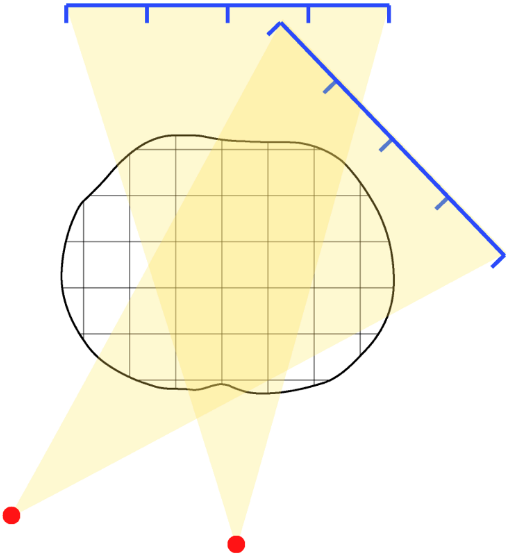


The numbers k and n are independent



$k = 8$

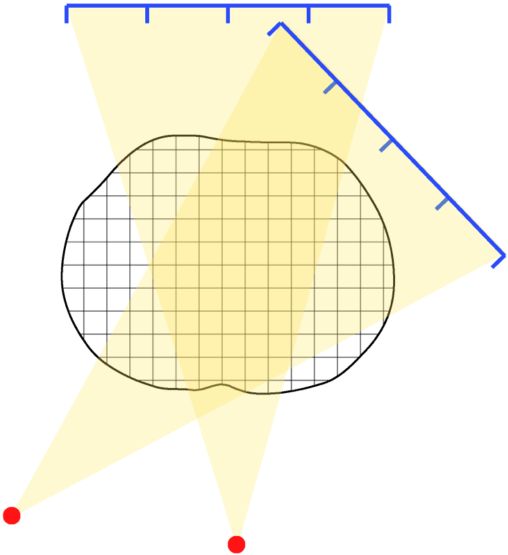
The numbers k and n are independent



$k = 8$

$n = 48$

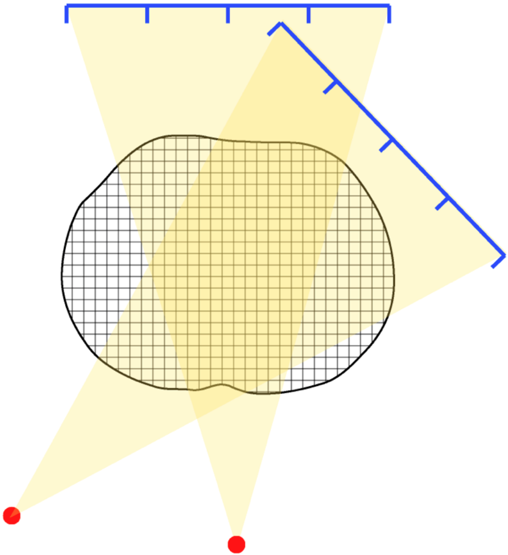
The numbers k and n are independent



$k = 8$

$n = 156$

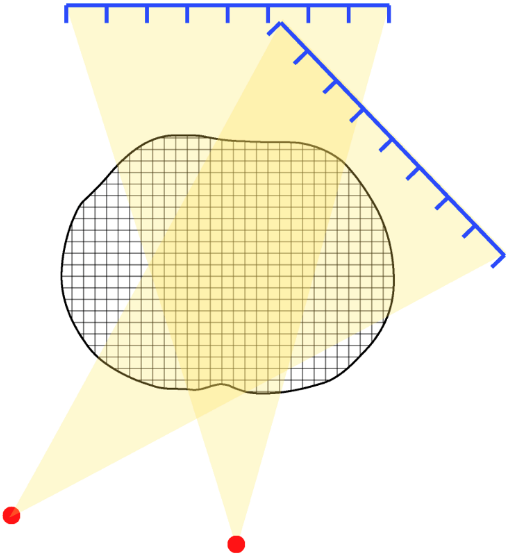
The numbers k and n are independent



$k = 8$

$n = 440$

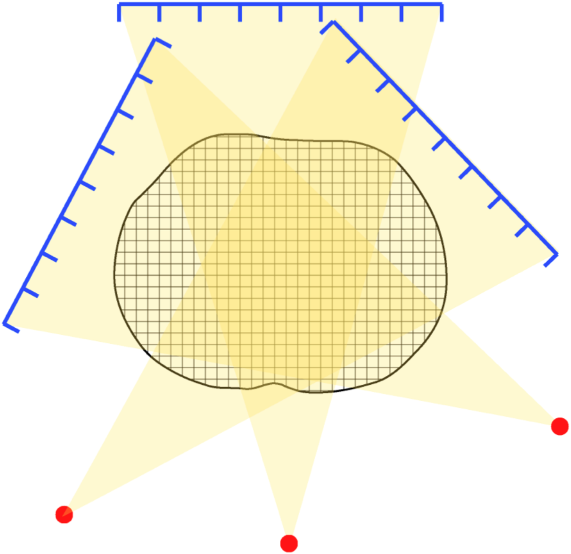
The numbers k and n are independent



$k = 16$

$n = 440$

The numbers k and n are independent



$k = 24$

$n = 440$

Bayesian estimates are drawn from the posterior distribution related to the computational model

The posterior distribution is defined by

$$\pi_{kn}(u_n \,|\, m_{kn}) = C\,\pi(u_n)\,\pi(m_{kn} \,|\, u_n).$$

Data is given as a realization of the practical measurement model $M_k = A_k U + \mathcal{E}_k$: $\widehat{m}_k = M_k(\omega_0)$.

The conditional mean estimate $u_{kn}^{\mathcal{CM}}$ is defined by

$$u_{kn}^{\mathcal{CM}} := \int_{\mathbb{R}^n} u_n\,\pi_{kn}(u_n \,|\, \widehat{m}_k)\,d\mu(u_n),$$

and the maximum a posteriori estimate $u_{kn}^{\mathcal{MAP}}$ is defined by

$$\pi_{kn}(u_{kn}^{\mathcal{MAP}} | \,\widehat{m}_k) = \max_{u_n}\{\pi_{kn}(u_n | \,\widehat{m}_k)\}.$$

We wish to construct discretization-invariant Bayesian inversion strategies

We achieve discretization-invariance by constructing a well-defined infinite-dimensional Bayesian framework, which can be projected to any finite dimension.

One advantage of discretization-invariance is that the prior distributions $\pi(u_n)$ represent the same prior information at any dimension n. This is useful for delayed acceptance MCMC algorithms, for example.

Discretization-invariant Bayesian estimates behave well as $k \to \infty$ and $n \to \infty$

Typically, a fixed measurement device is given, and k is constant. In that case, these limits should exist:

$$\lim_{n\to\infty} u_{kn}^{\mathcal{CM}} \qquad \text{and} \qquad \lim_{n\to\infty} u_{kn}^{\mathcal{MAP}}.$$

On the other hand, we might have the opportunity to update our device to a more accurate one, increasing k independently of n. If we keep n fixed, the following limits should exist:

$$\lim_{k\to\infty} u_{kn}^{\mathcal{CM}} \qquad \text{and} \qquad \lim_{k\to\infty} u_{kn}^{\mathcal{MAP}}.$$

History of infinite-dimensional Bayesian inversion

1970 Franklin Linear inverse problems, Hilbert space
1984 Mandelbaum Gaussian posterior means, Hilbert space
1989 Lehtinen, Päivärinta and Somersalo
Bayesian linear inverse problems with Gaussian priors in Polish spaces

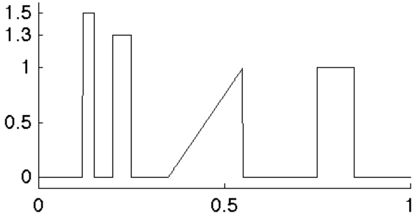
1991 Fitzpatrick Hypothesis testing, linear, Banach space
1995 Luschgy Extension of Mandelbaum to Banach space
2002 Lasanen Discretization study, Gaussian, Hilbert space
2004 Lassas and S Total variation prior not discretization-invariant
2005 Piiroinen Nonlinear inverse problems, Suslin space
2008 Neubauer & Pikkarainen Convergence rates, Hilbert space
2009 Lassas, Saksman & S Discretization-invariance, non-Gaussian
2009 Helin & Lassas Hierarchical Bayesian priors in function spaces
2009 Cotter, Dashti, Robinson & Stuart Fluid mechanics
2010 Stuart Review, analysis of non-white noise
2011 Kolehmainen, Lassas, Niinimäki & S Deconvolution

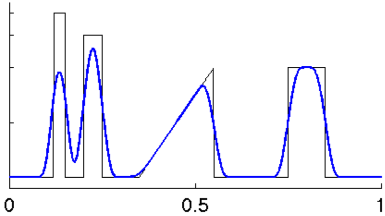
Outline

Continuum model for 1D deconvolution

Denote by $\mathbb{T}$ the circle resulting from identifying the endpoints of the interval $[0, 1]$. Take $\mathcal{A}$ to be the periodic convolution operator with a smooth kernel Ψ:

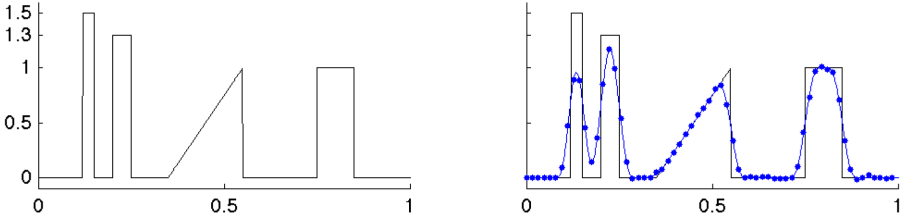
$$(\mathcal{A}U)(x) = \int_{\mathbb{T}} \Psi(x - y)U(y)d\sigma(y)$$





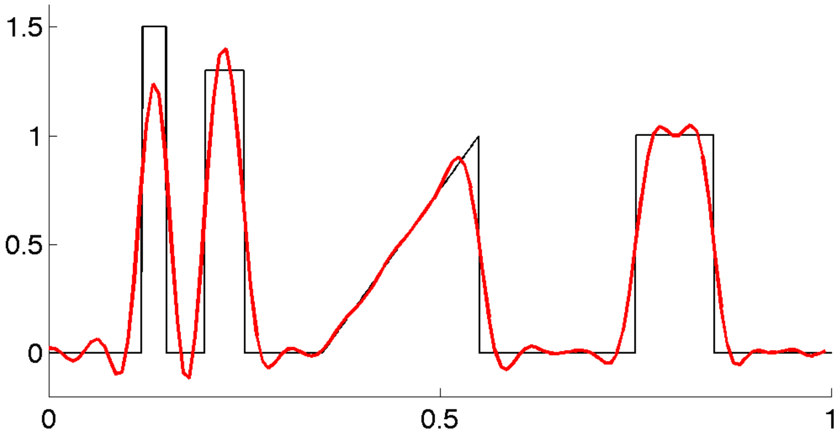
Finite data for 1D deconvolution

We simulate measurement data $\widehat{m}_{63} \in \mathbb{R}^{63}$ with noise level $\sigma = 0.01 \cdot \max U(x)$ as $\widehat{m}_{63}(\nu) = (\mathcal{A}U)((\nu - 1)/K) + \sigma \cdot r$, where $1 \leq \nu \leq 63$ and $r \sim \mathcal{N}(0, 1)$.



Estimate using a Gaussian smoothness prior

Dimension $n = 128$. Here the MAP and CM estimates coincide.



Outline

Bayesian inversion using total variation prior

Theorem (Lassas and S 2004)

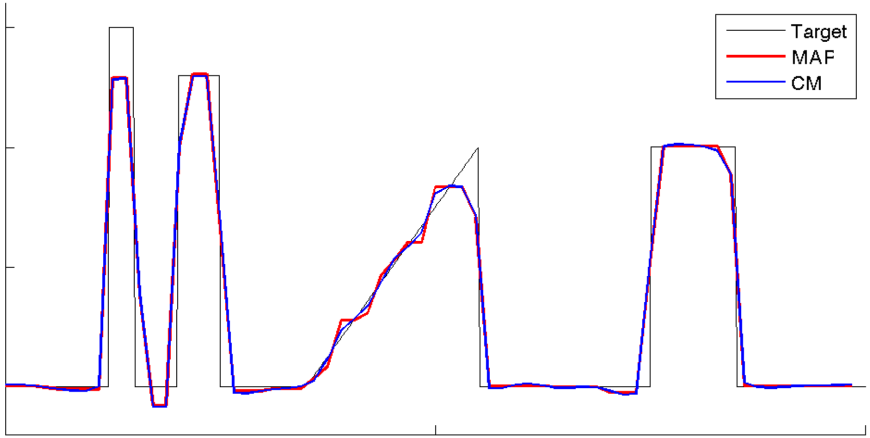
Total variation prior is not discretization-invariant.

Sketch of proof: Apply a variant of the central limit theorem to the independent, identically distributed random consecutive differences.

New numerical experiments are reported in
Kolehmainen, Lassas, Niinimäki and S (submitted).

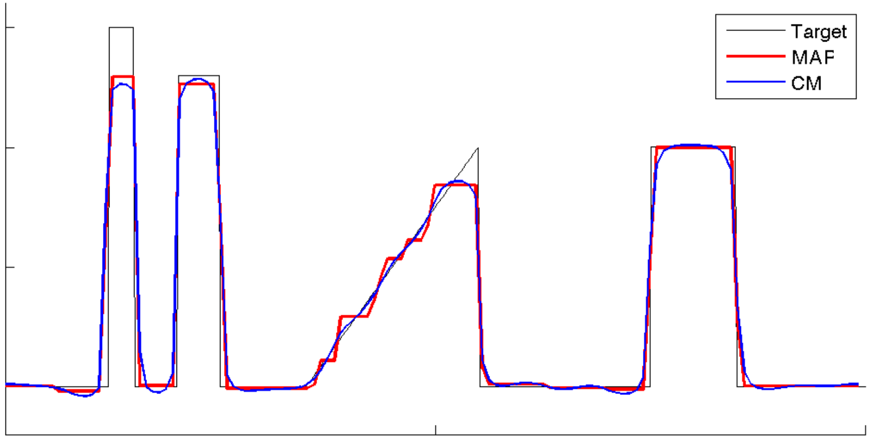
Total variation, fixed parameter

$n = 64$



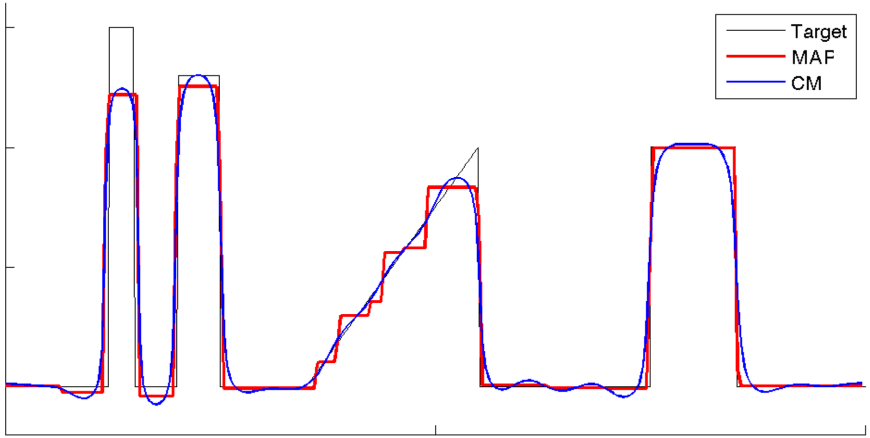
Total variation, fixed parameter

$n = 128$



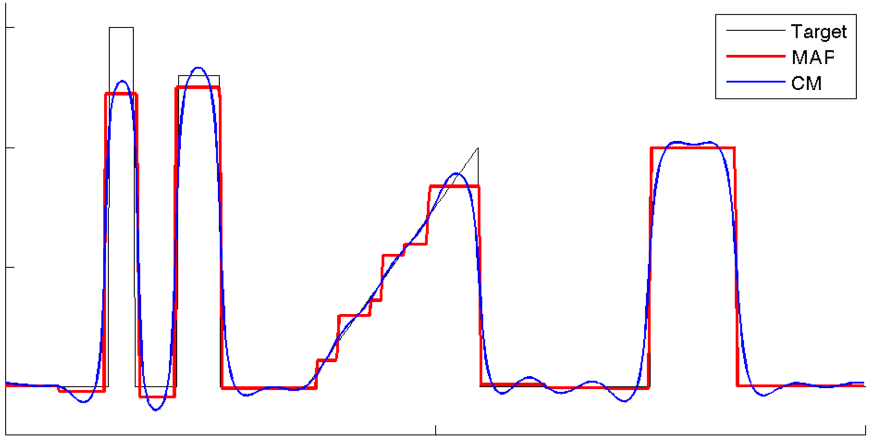
Total variation, fixed parameter

$n = 256$



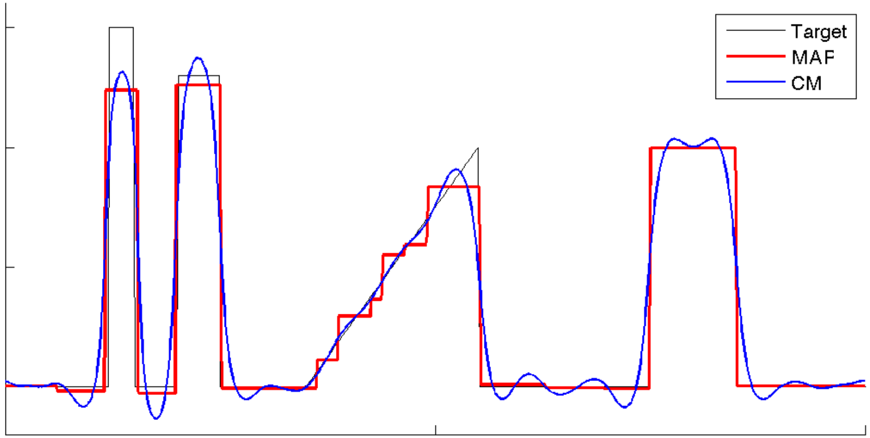
Total variation, fixed parameter

$n = 512$



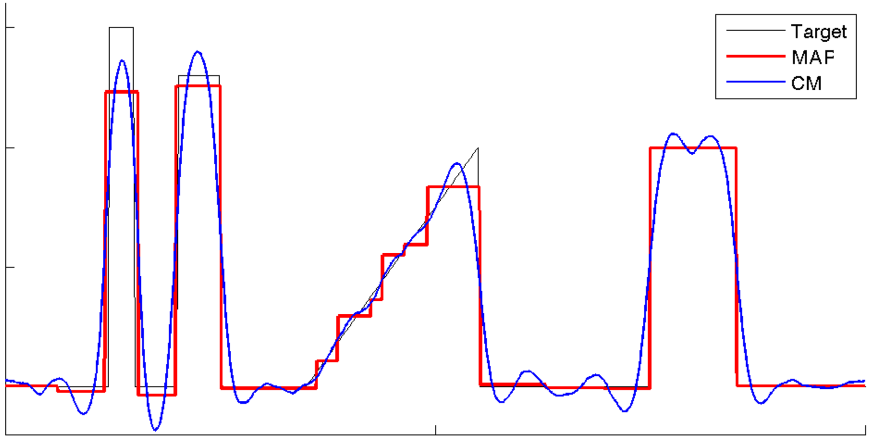
Total variation, fixed parameter

$n = 1024$



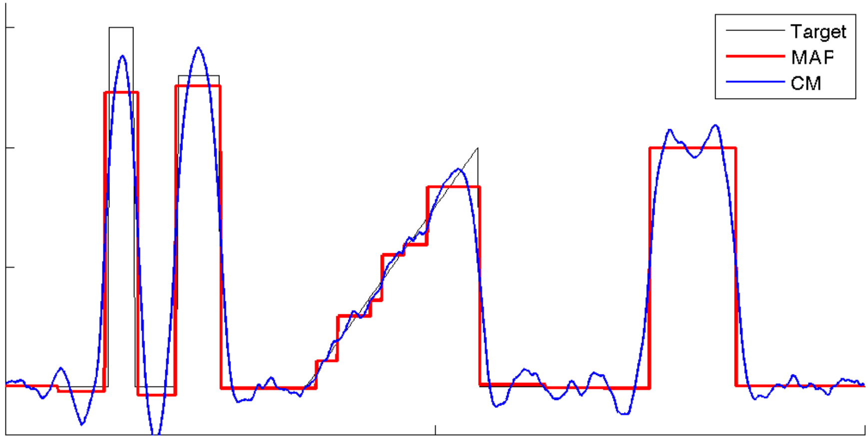
Total variation, fixed parameter

$n = 2048$



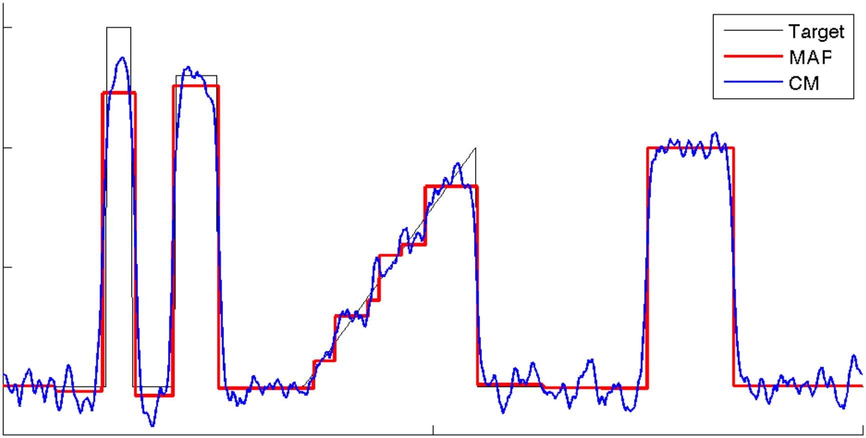
Total variation, fixed parameter

$n = 4096$



Total variation, fixed parameter

$n = 8192$

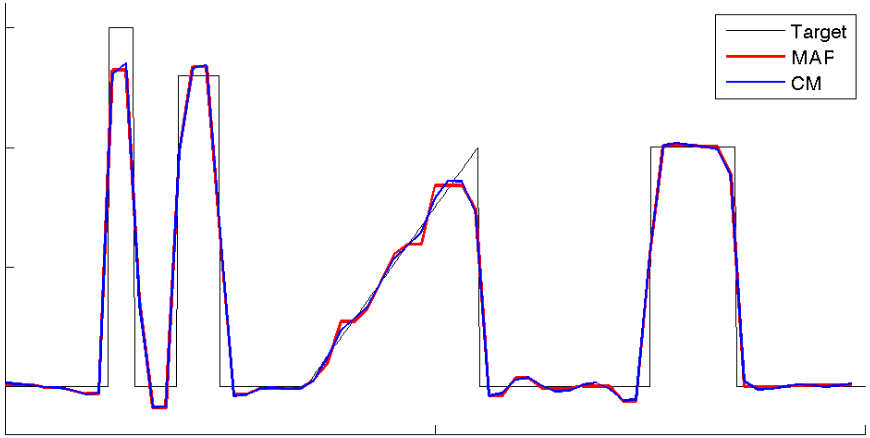


With fixed parameter, the MAP estimates converge,
but CM estimates diverge.

Let's see what happens with a parameter scaled as $\sqrt{n}$.

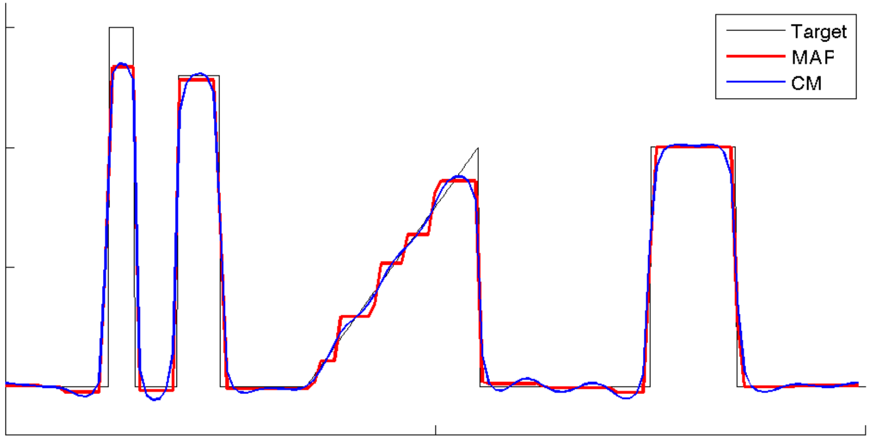
Total variation, parameter $\sim \sqrt{n}$

$n = 64$



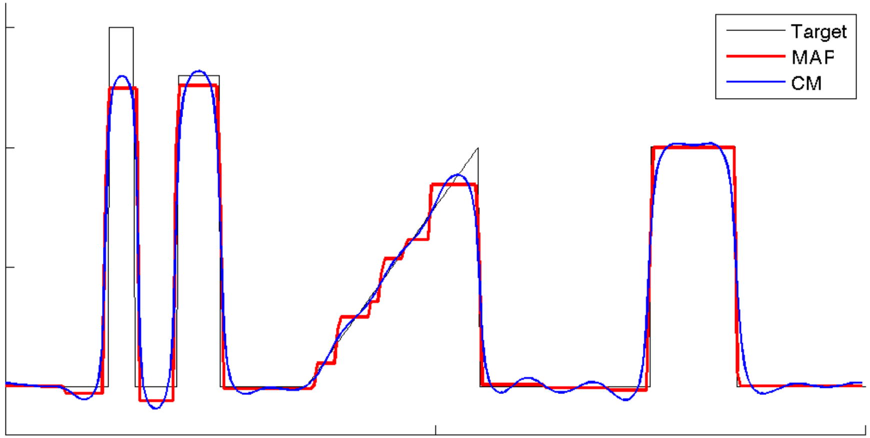
Total variation, parameter $\sim \sqrt{n}$

$n = 128$



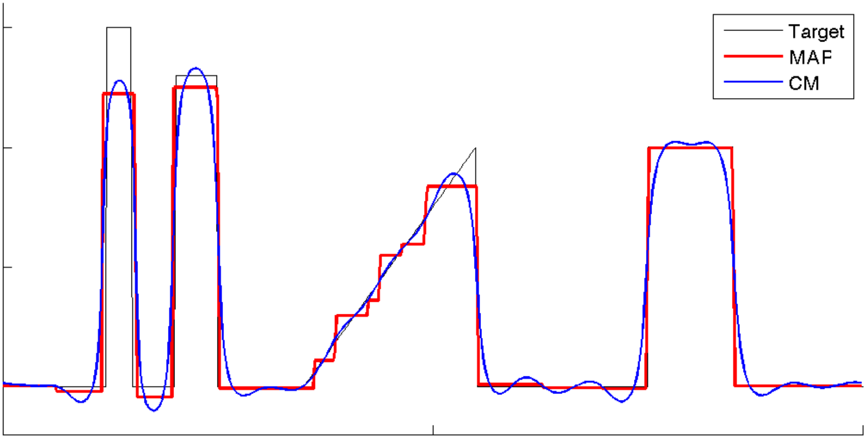
Total variation, parameter $\sim \sqrt{n}$

$n = 256$



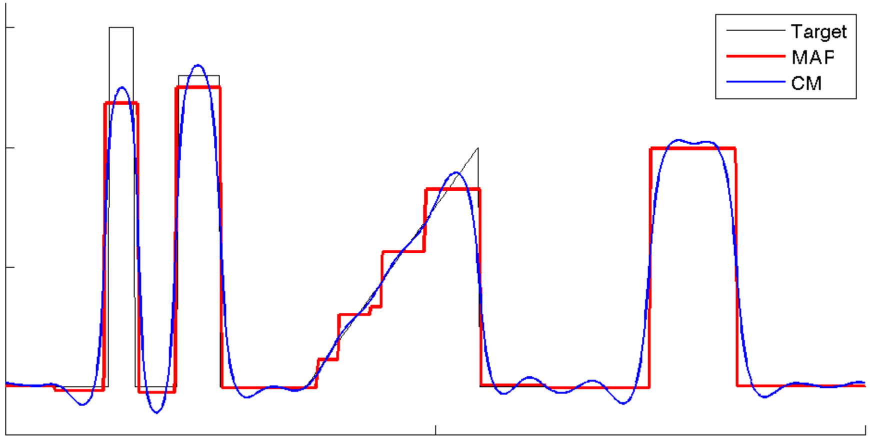
Total variation, parameter $\sim \sqrt{n}$

$n = 512$



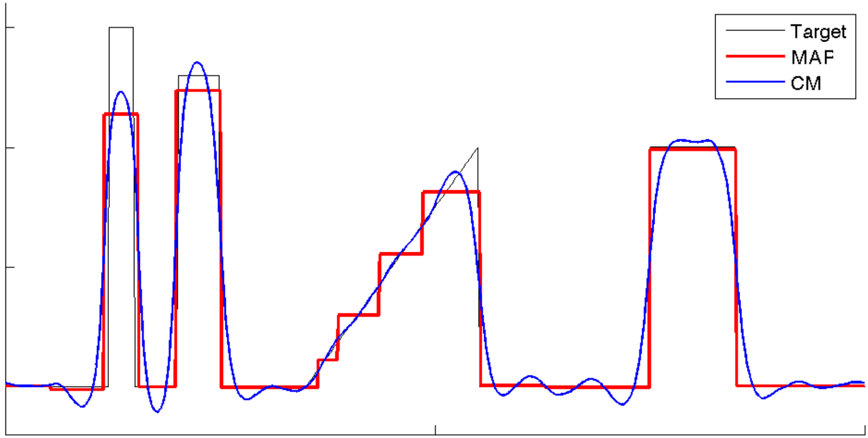
Total variation, parameter $\sim \sqrt{n}$

$n = 1024$



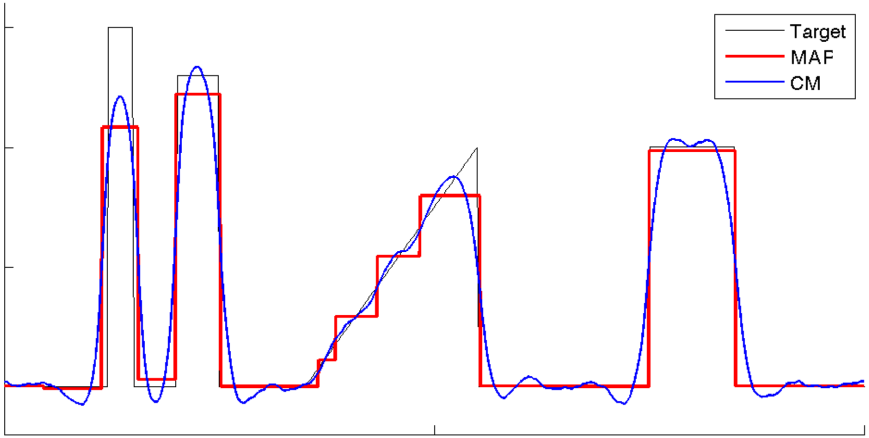
Total variation, parameter $\sim \sqrt{n}$

$n = 2048$



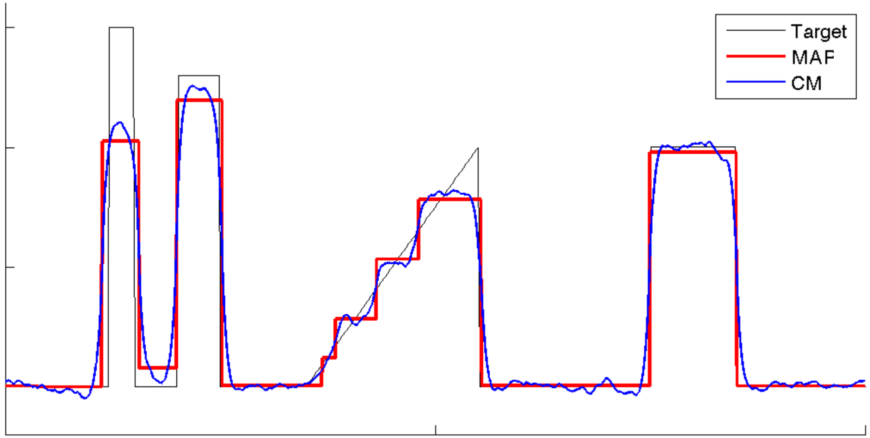
Total variation, parameter $\sim \sqrt{n}$

$n = 4096$

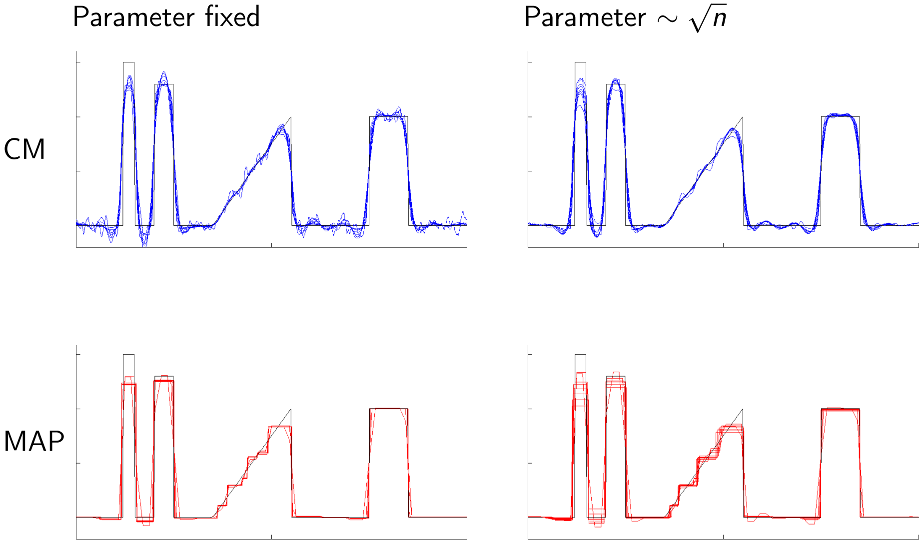


Total variation, parameter $\sim \sqrt{n}$

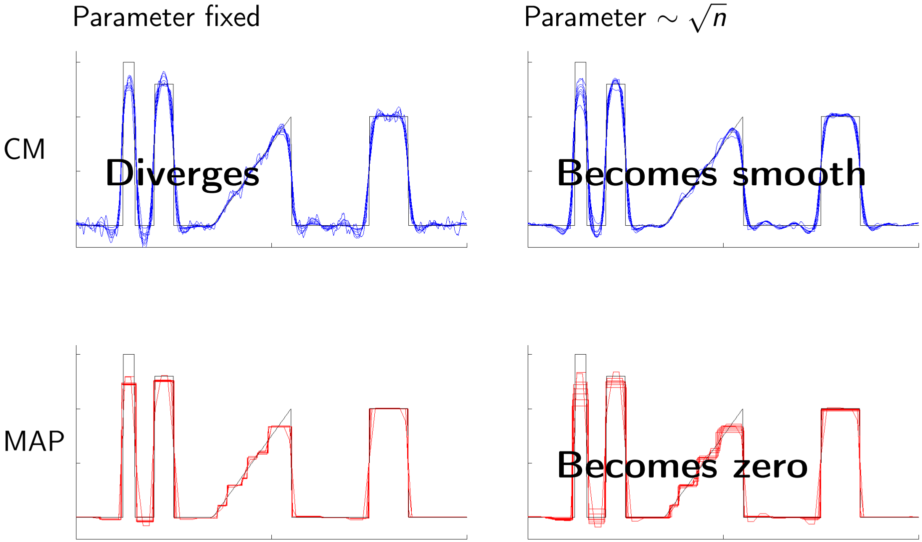
$n = 8192$



Discretization dilemma with total variation prior



Discretization dilemma with total variation prior



Outline

Following Daubechies (1992) we construct a wavelet basis for 1-periodic functions

Let ψ^C and ϕ^C be compactly supported mother wavelet and scaling function, respectively, suitable for orthonormal multiresolution analysis in $\mathbb{R}$. Set

$$\begin{aligned}\phi_{j,k}(x) &= \sum_{\ell\in\mathbb{Z}} 2^{j/2}\phi^C(2^j(x+\ell)-k),\\ \psi_{j,k}(x) &= \sum_{\ell\in\mathbb{Z}} 2^{j/2}\psi^C(2^j(x+\ell)-k).\end{aligned}$$

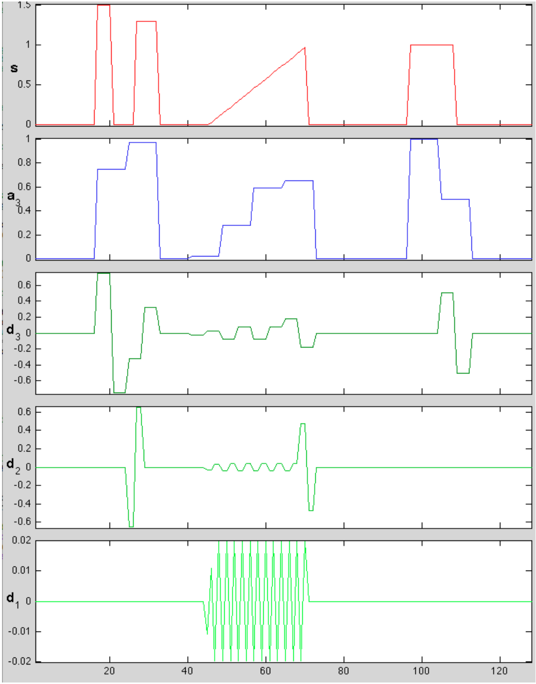
Define $V_j := \overline{\text{span}\{\phi_{j,k} \mid k\in\mathbb{Z}\}}$ and $W_j := \overline{\text{span}\{\psi_{j,k} \mid k\in\mathbb{Z}\}}$. Then V_j are spaces of constant functions for $j\le 0$, and we have a ladder $V_0\subset V_1\subset V_2\subset\cdots$ of multiresolution spaces satisfying $\overline{\cup_{j\ge 0}V_j} = L^2(\mathbb{T})$.

Wavelet decomposition divides a function into details at different scales

We can write

$$U(x) = c_0 + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} w_{j,k}\, \psi_{j,k}(x),$$

where $c_0 = \langle U, \phi_{0,0} \rangle$ and $w_{j,k} = \langle U, \psi_{j,k} \rangle$.



low-pass signal

$j = 0$

$j = 1$

$j = 2$

Besov space norms can be expressed in terms of wavelet coefficients

The function U belongs to $B^s_{pq}(\mathbb{T})$ if and only if

$$\|U\|_{B^s_{pq}(\mathbb{T})} := \left(|c_0|^q + \sum_{j=0}^{\infty} 2^{jq(s+\frac{1}{2}-\frac{1}{p})} \left(\sum_{k=0}^{2^j-1} |w_{j,k}|^p \right)^{\frac{q}{p}} \right)^{\frac{1}{q}} < \infty.$$

The sparsity-promoting choice $p = 1 = q$ and $s = 1$ leads to

$$\|U\|_{B^1_{11}(\mathbb{T})} = |c_0| + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} 2^{j/2} |w_{j,k}|.$$

Outline

Besov space priors take a simple form in terms of wavelet coefficients

Let $N > 0$ and set $n = 2^N$. Take the projection T_n to be

$$T_n U(x) = c_0 + \sum_{j=0}^{N} \sum_{k=0}^{2^j-1} w_{j,k}\, \psi_{j,k}(x).$$

The choice $p = 1 = q$ and $s = 1$ leads to the prior

$$\pi(u_n) = C \exp(-\alpha_n \left[|c_0| + \sum_{j=0}^{N} \sum_{k=0}^{2^j-1} 2^{j/2} |w_{j,k}| \right]).$$

Bayesian inversion using Besov space priors

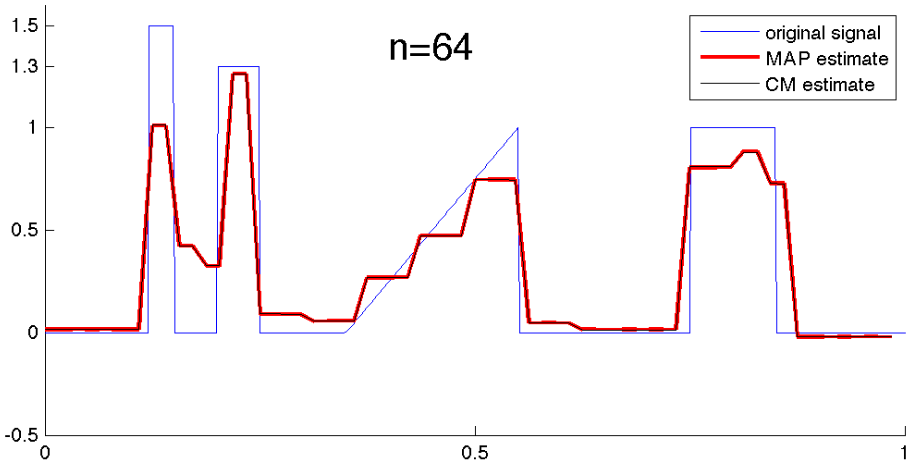
Theorem (Lassas, Saksman and S 2009)

Besov space priors are discretization-invariant.

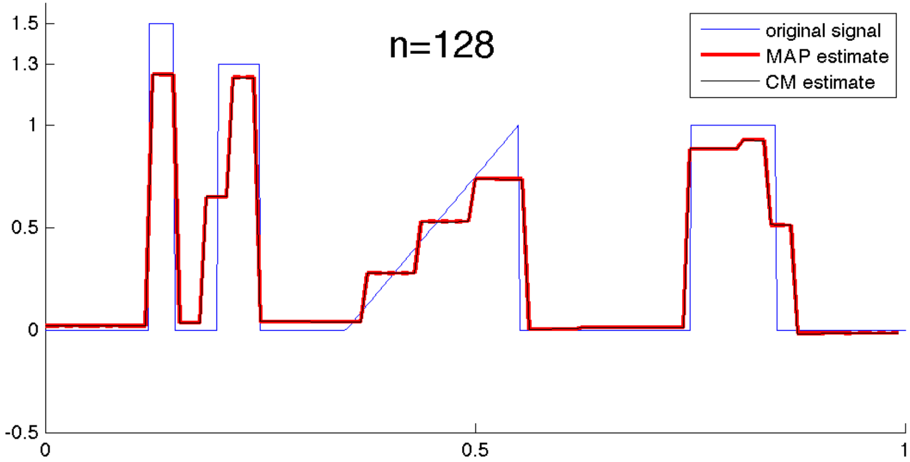
Sketch of proof: Construction of well-defined Bayesian inversion theory in infinite-dimensional Besov spaces that allow wavelet bases. Discretizations are achieved by truncating the wavelet expansion.

Numerical experiments are reported in
Kolehmainen, Lassas, Niinimäki and S (submitted).

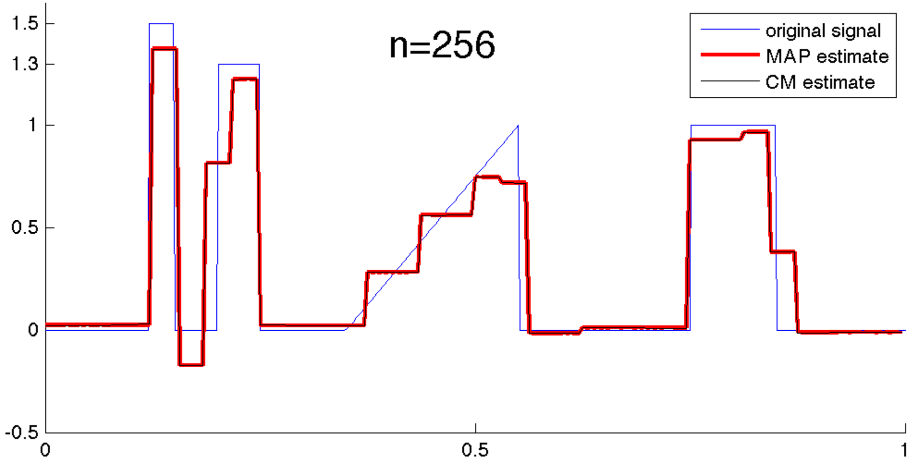
Deterministic Besov space regularization was first introduced in
Daubechies, Defrise and De Mol 2004.



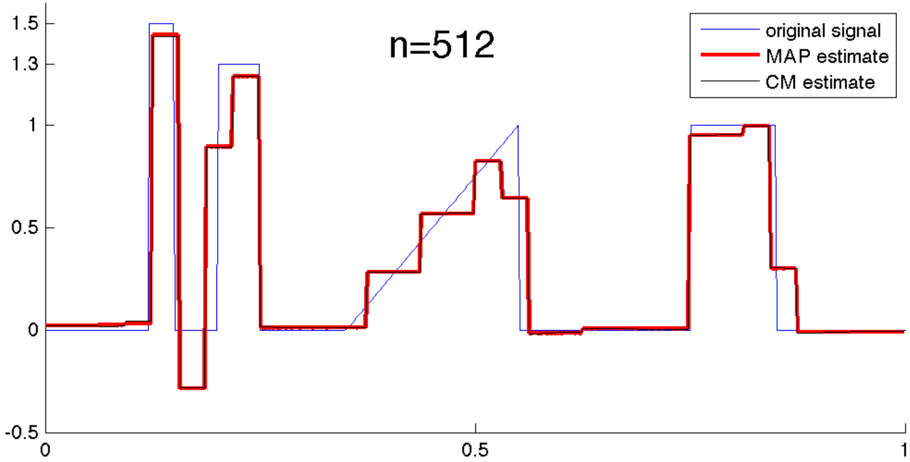
n=64
original signal
MAP estimate
CM estimate
1.5
1.3
1
0.5
0
-0.5
0
0.5
1



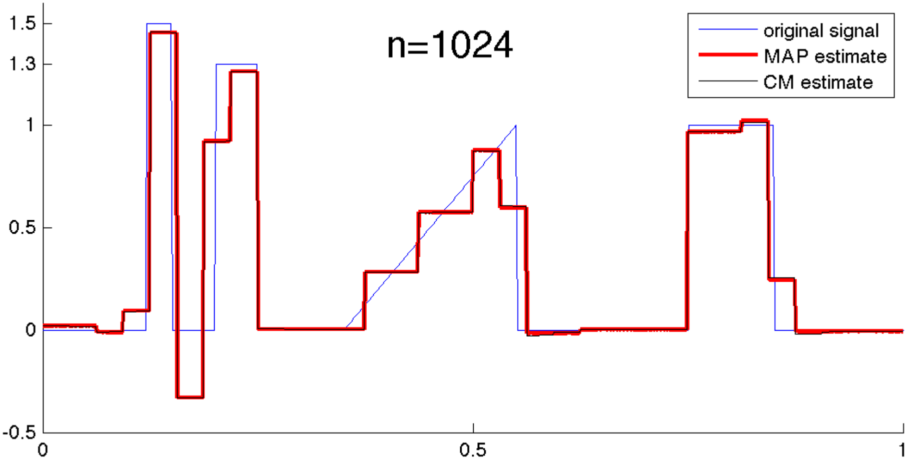
n=128
original signal
MAP estimate
CM estimate
1.5
1.3
1
0.5
0
-0.5
0
0.5
1



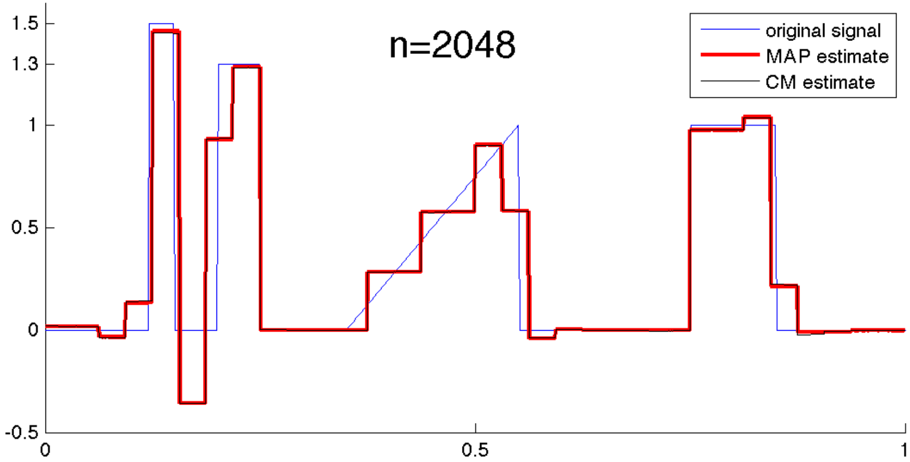
n=256
original signal
MAP estimate
CM estimate
1.5
1.3
1
0.5
0
-0.5
0
0.5
1



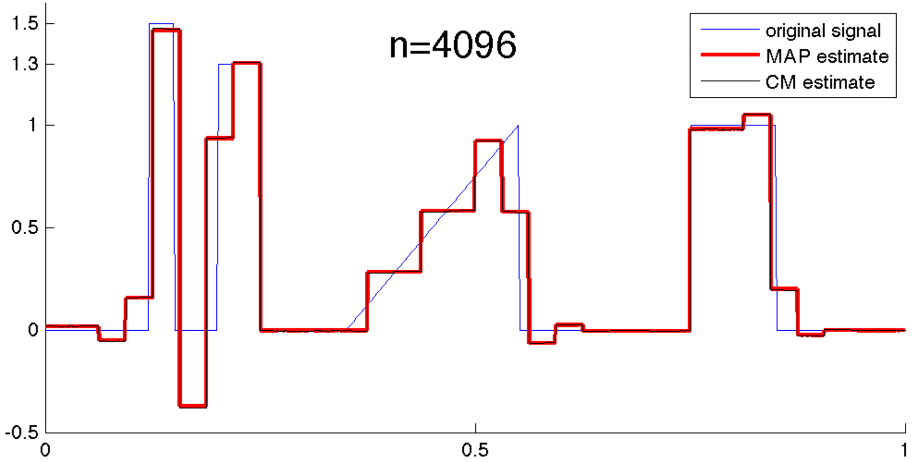
n=512
original signal
MAP estimate
CM estimate
1.5
1.3
1
0.5
0
-0.5
0
0.5
1



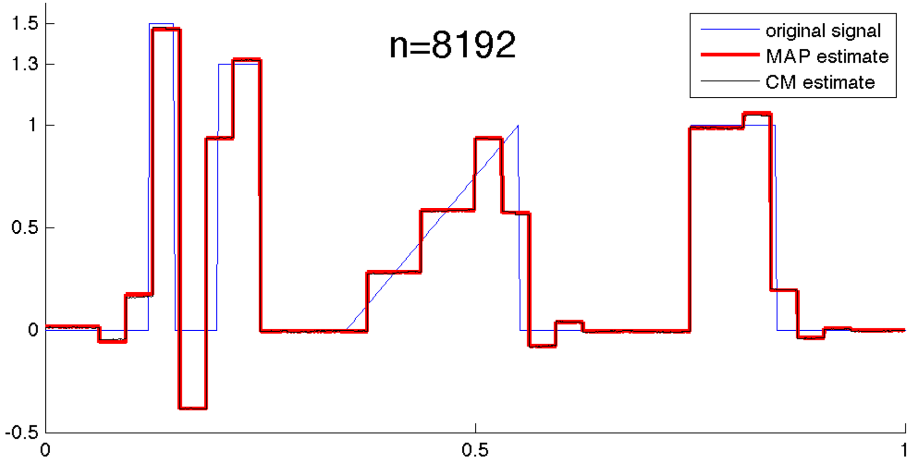
n=1024
original signal
MAP estimate
CM estimate
1.5
1.3
1
0.5
0
-0.5
0
0.5
1



n=2048
original signal
MAP estimate
CM estimate
1.5
1.3
1
0.5
0
-0.5
0
0.5
1



n=4096
original signal
MAP estimate
CM estimate
1.5
1.3
1
0.5
0
-0.5
0
0.5
1


n=8192
original signal
MAP estimate
CM estimate
1.5
1.3
1
0.5
0
-0.5
0
0.5
1

Computational implementation required special algorithms as the dimension went up to $n = 8192$

The MAP estimates were computed using a tailored Quadratic Programming algorithm.

The CM estimates were computed using the Metropolis-Hastings algorithm and a random one-element scanning scheme.

Details of the implementation can be found in **Kolehmainen, Lassas, Niinimäki and S (submitted)**.

How sparse are the MAP estimates?

n:	64	128	256	512	1024	2048	4096	8192
nonzeroes:	30	48	103	68	208	334	328	385

It seems that the number of nonzero wavelet coefficients in the MAP estimate practically freezes for $n \geq 2048$.

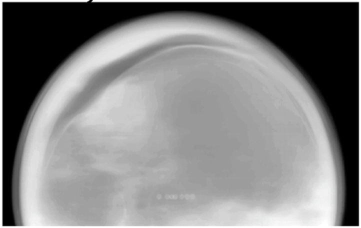
This numerical evidence is in line with the results in
Grasmair, Haltmeier and Scherzer,
Sparse regularization with ℓ^q penalty term,
Inverse Problems 2008.

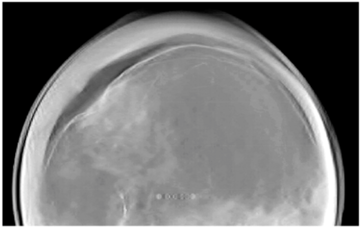
Outline

Limited angle tomography results for X-ray mammography

[Rantala *et al.* 2006]
Thanks to GE Healthcare

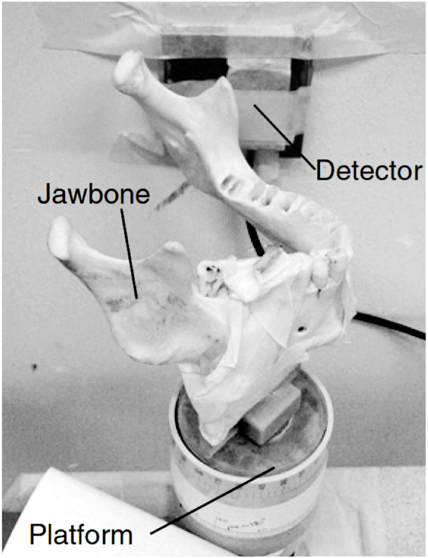
Tomosynthesis

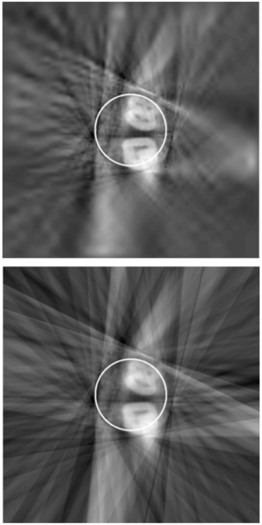




MAP estimate, Besov prior, $p=1.5=q$ and $s=0.5$

Local tomography results for dental X-ray imaging; data measured from specimen

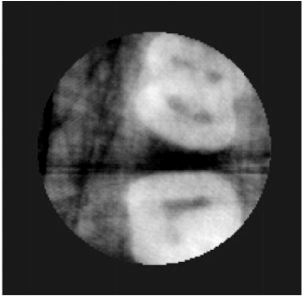




Besov space method compared to Λ-tomography in the region of interest

Lambda-tomography

MAP using Besov prior with p=q=1.5 and s=0.5



Niinimäki, S and Kolehmainen (2007)
Thanks to Palodex Group

Summary

- Refining discretization in Bayesian inversion should increase the accuracy of the results. Hence we introduce the notion of discretization-invariance.
- Total variation prior is not discretization-invariant, and should not be used in Bayesian inversion.
  Deterministic total variation regularization is OK.
- Wavelet-based Besov space priors promote sparsity and provide discretization-invariant Bayesian estimates.

Thank you for your attention!

Preprints available at **www.siltanen-research.net.**