Inverse Problems **28** (2012) 025005 (28pp) doi:10.1088/0266-5611/28/2/025005

# Sparsity-promoting Bayesian inversion

**V Kolehmainen[1], M Lassas[2], K Niinimäki[1] and S Siltanen[2]**

[1] Department of Applied Physics, University of Eastern Finland, Kuopio, Finland
[2] Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

E-mail: samuli.siltanen@helsinki.fi

Received 17 January 2011, in final form 1 November 2011
Published 10 January 2012
Online at stacks.iop.org/IP/28/025005

**Abstract**
A computational Bayesian inversion model is demonstrated. It is discretization invariant, describes prior information using function spaces with a wavelet basis and promotes reconstructions that are sparse in the wavelet transform domain. The method makes use of the Besov space prior with $p = 1$, $q = 1$ and $s = 1$, which is related to the total variation prior. Numerical evidence is presented in the context of a one-dimensional deconvolution task, suggesting that edge-preserving and noise-robust reconstructions can be achieved consistently at various resolutions.

## 1. Introduction

We demonstrate a new computational Bayesian inversion model that combines the following three properties. First, discretization invariance ensures that refining the computational grid leads to more accurate approximation of a well-defined limit estimate. Second, prior information about the quantity of interest can be described using an infinite-dimensional function space that allows a wavelet basis. Third, it is possible to promote estimates with an *a priori* known level of sparsity in the wavelet domain.

Consider an ill-posed inverse problem where a physical quantity $f$ needs to be recovered from an indirect measurement $m$ modelled by

$$m = Af + \epsilon, \tag{1.1}$$

where $m \in \mathbb{R}^k$ is the measurement data vector, $A$ is a linear operator modelling the measurement and $\epsilon$ is white noise taking values in $\mathbb{R}^k$. The unknown quantity $f$ is defined on the one-dimensional interval $[0, 1]$ with periodic boundary conditions: we write $f : \mathbb{T} \to \mathbb{R}$ with $\mathbb{T}$ denoting the one-dimensional torus[3].

There are several reasons for considering the unknown $f$ as a function as opposed to a discrete quantity. Many measurements of practical interest can be modelled very accurately by (1.1), where $A$ is a partial differential operator or integral operator arising from mathematical physics in a continuous setting. Also, analytic and geometric theory of inverse problems

[3] Theoretical treatment of the multidimensional case is analogous.

  

typically assumes *a priori* knowledge in the form $f \in X$, where $X$ is an infinite-dimensional function space defined on a subset of $\mathbb{R}^d$ or on a manifold.

A computational solution of the inverse problem is based on a discrete model for approximating the function $f \in X$. Let $T_n : X \to X$ be a linear projection operator with the $n$-dimensional range: $\dim(T_n(X)) = n < \infty$. We denote $f_{(n)} = T_n f \in X$ and study the finite computational model

$$m_n = A f_{(n)} + \epsilon. \tag{1.2}$$

Note that $n$ can be chosen independently of the number $k$ of measurements.

In Bayesian inversion, the quantities $m_n$ and $f_{(n)}$ and $\epsilon$ are considered as random variables. Here, $m_n$ and $\epsilon$ take values in $\mathbb{R}^k$, and $f_{(n)}$ takes values in $T_n(X)$. The complete solution to the discrete inverse problem is the *posterior distribution* defined using conditional probabilities and the Bayes formula:

$$\pi_{\text{post}}(f_{(n)} | \, m_n) = \frac{\pi\,(f_{(n)})\,\pi\,(m_n \mid f_{(n)})}{\pi\,(m_n)}, \tag{1.3}$$

where $\pi\,(f_{(n)})$ is the *prior model* representing *a priori* information about the unknown and $\pi\,(m_n)$ is a normalizing constant. The conditional probability $\pi\,(m_n|f_{(n)})$ is the *likelihood model* describing the measurement. In the case of independent white noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, we have

$$\pi\,(m_n \mid f_{(n)}) = \pi_\epsilon(m_n - A f_{(n)}) = C \exp\left(-\frac{1}{2\sigma^2}\|A f_{(n)} - m_n\|_2^2\right), \tag{1.4}$$

where $C$ is a normalization constant and $\| \cdot \|_2$ stands for the Euclidean norm on $\mathbb{R}^k$.

In practice, we are given a realization $\widehat{m} \in \mathbb{R}^k$ of the measurement modelled by (1.1), which is substituted into (1.3):

$$\pi_{\text{post}}(f_{(n)} \mid \widehat{m}) = \frac{\pi\,(f_{(n)})\,\pi\,(\widehat{m} \mid f_{(n)})}{\pi\,(\widehat{m})}, \tag{1.5}$$

and a representative estimate is computed from the distribution (1.5). In this work, we study two such estimates: the conditional mean (cm) estimate

$$f_{(n)}^{\text{CM}} := \int_{T_n(X)} f_{(n)}\, \pi_{\text{post}}(f_{(n)} \mid \widehat{m})\, \mathrm{d}\mu(f_{(n)}), \tag{1.6}$$

and the maximum *a posteriori* (MAP) estimate $f_{(n)}^{\text{MAP}}$ defined by

$$\pi_{\text{post}}\big(f_{(n)}^{\text{MAP}} | \, \widehat{m}\big) = \max\{\pi_{\text{post}}(f_{(n)} | \, \widehat{m}) \ : \ f_{(n)} \in T_n(X)\}. \tag{1.7}$$

For general references on Bayesian inversion, see [30, 34, 42, 50, 51].

The concept of *sparsity* is an active research topic in the fields of signal processing and deterministic inversion. If $f$ can be finitely represented in some transform domain, then effective computational algorithms often become available for recovering $f$ from $\widehat{m}$. The use of $\| \cdot \|_1$ norms instead of $\| \cdot \|_2$ norms in regularization is known to promote sparsity [9, 18, 19, 45, 58]; in particular, the total variation norm $\|f\|_{\text{BV}}$ is widely used in computational inversion. Sparse wavelet representations and Besov spaces are useful for denoising [12, 20], for linear ill-posed inverse problems [13, 16, 21], and for nonlinear inverse problems [3, 4, 68, 10, 17, 22, 24, 25, 27, 32, 36, 55, 56].

Bayesian inversion generalizes deterministic regularization and yields information about the structure of ill posedness of the inverse problem. Therefore, it is of practical and conceptual interest to develop sparsity-promoting Bayesian inversion methods.

The total variation prior of the form '$\exp(-\alpha\|f\|_{\text{BV}})$' has been studied in the Bayesian framework. With fixed discretizations, edge-preserving MAP and CM estimates are achieved

in various applications [33, 37, 38, 61]. However, it was shown in [39] that the Bayesian total variation prior is not discretization invariant. This so-called discretization dilemma will be discussed in more detail in sections 3 and 6.2.

Two classes of priors are known to be discretization invariant: Gaussian priors [31, 41, 43, 54] and Besov space priors [40]. In the latter case, *a priori* knowledge can be expressed in the infinite-dimensional continuum setting as a probability distribution on the coefficients of a wavelet basis for a Besov space, and the same prior information is approximately represented at any finite discretization level by the projection operator $T_n$ that truncates the wavelet expansion to $n$ coarsest terms. (Actually, the wavelet-based approach works also for the Gaussian case since the Besov space $B^s_{22}$ equals the Sobolev space $H^s$, but more efficient computational methods are typically used in practice.)

Bayesian inversion based on Besov space priors has been applied successfully to practical tomographic problems in [2, 35, 52, 57, 65]. However, the spaces used there were $B^s_{pp}$ with $p > 1$, leading to continuously differentiable objective functionals in the computation of MAP estimates. From the point of view of sparse reconstruction, the most promising Besov prior space is $B^1_{11}$ whose norm involves $L^1$ integrals of band-limited first derivatives. This is closely related to the total variation norm and leads to a nonsmooth posterior.

We report new theoretical and numerical results concerning the $B^1_{11}$ prior. We prove the existence of $\lim_{n\to\infty} f^{\mathrm{MAP}}_{(n)}$, ensuring that MAP estimates behave consistently at different resolutions (existence of the limit $\lim_{n\to\infty} f^{\mathrm{CM}}_{(n)}$ is shown in [40]). Furthermore, we present the first numerical computations of both MAP and CM estimates from the same data using the $B^1_{11}$ prior, including convergence studies of estimates as the computational grid is refined. The total variation prior is used as a reference.

The computation of MAP estimates is equivalent to minimizing

$$\|Af - m\|_2^2 + \alpha \sum_{j=1}^{\infty} W_j |\langle f, \psi_j \rangle|, \tag{1.8}$$

where $\psi_j$ is a wavelet basis, $\alpha > 0$ is a constant and $W_j$ are the positive weights. The seminal article [16] and most of its extensions rely on iterative thresholding schemes that are simple to implement and computationally efficient. In contrast, we minimize (1.8) using quadratic programming (QP); see [5, 53, 66]. Implementation of physically relevant constraints (such as non-negativity of $f$) is easier to implement in the QP framework than in the thresholding methods. Therefore, our results provide a preferred approach for inverse problems with *a priori* known constraints.

Another novelty here is a sparsity-based choice rule for the prior parameter $\alpha > 0$ in (1.8). Many methods have been suggested for the choice of optimal value for $\alpha$ which also can be used in sparse recovery applications; see the review [1]. See also [46] for a parameter-choice method designed for sparse reconstruction. According to the Bayesian paradigm, the parameter $\alpha$ is considered as an auxiliary unknown of the inverse problem. See [11, 29, 47] as general references to the Bayesian parameter-choice principles. We propose a new method for choosing the parameter $\alpha$ in (1.8), based on *a priori* knowledge of the number of nonzero wavelet coefficients in $f$. The numerical experiments suggest that the method is robust and simple, and yields useful choices of $\alpha$. We remark that the method does not use the magnitude of the noise $\varepsilon$.

This paper is organized as follows. In section 2, we give the details of the wavelet-based construction of the $B^1_{11}(\mathbb{T})$ prior. In Section 3, we discuss the discretization dilemma related to the total variation prior. In sections 4 and 5, we describe computational methods for the direct and inverse problems, respectively. Numerical results concerning the one-dimensional

test problem are presented in section 6. We discuss and conclude our findings in sections 7 and 8, respectively.

In what follows, $C$ and $C_n$ denote generic constants whose values may change from line to line.

## 2. Wavelets and Besov space priors in dimension 1

### *2.1. The continuum case*

Following [15, section 9.3], we construct a wavelet representation for 1-periodic functions on $\mathbb{R}$, in other words, for functions on the one-dimensional torus $\mathbb{T}$ constructed by identifying the two boundary points of the interval $[0, 1]$.

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

$$\phi_{j,h}(x) = \sum_{\ell\in\mathbb{Z}} 2^{j/2}\phi^C(2^j(x+\ell) - h), \tag{2.1}$$

$$\psi_{j,h}(x) = \sum_{\ell\in\mathbb{Z}} 2^{j/2}\psi^C(2^j(x+\ell) - h), \tag{2.2}$$

and define spaces $V_j := \overline{\text{span}\{\phi_{j,h} \,|\, h \in \mathbb{Z}\}}$ and $W_j := \overline{\text{span}\{\psi_{j,h} \,|\, h \in \mathbb{Z}\}}$. It turns out that $V_j$ are spaces of constant functions for $j \leqslant 0$. Thus, we have a ladder $V_0 \subset V_1 \subset V_2 \subset \cdots$ of multiresolution spaces satisfying $\overline{\cup_{j\geqslant 0} V_j} = L^2(\mathbb{T})$.

The spaces $W_j$ are the successive orthogonal complements of $V_j$ in $V_{j+1}$ for $j \geqslant 0$. Each space $W_j$ has the orthonormal basis $\{\psi_{j,h} \,|\, h = 0, \ldots, 2^j - 1\}$, and we can represent functions as follows:

$$f(x) = c_0 + \sum_{j=0}^{\infty}\sum_{h=0}^{2^j-1} w_{j,h}\,\psi_{j,h}(x), \tag{2.3}$$

where the coefficients are defined by

$$c_0 := \langle f, 1\rangle = \textstyle\int_{\mathbb{T}} f(x)\,\mathrm{d}x, \tag{2.4}$$

$$w_{j,h} := \langle f, \psi_{j,h}\rangle = \textstyle\int_{\mathbb{T}} f(x)\psi_{j,h}(x)\,\mathrm{d}x. \tag{2.5}$$

Besov spaces are the function spaces denoted by $B^s_{pq}(\mathbb{T})$, where $s \in \mathbb{R}$ is a smoothness index, and $1 \leqslant p < \infty$ and $1 \leqslant q < \infty$ are integrability exponents. Roughly speaking, the greater is the $s$, the more derivatives a function $f \in B^s_{pq}(\mathbb{T})$ has with the finite $L^p(\mathbb{T})$ norm. According to Meyer [49, section 6.10] and Triebel [62], we can characterize periodic Besov space functions using wavelets. Assuming enough smoothness in the mother wavelet $\psi_{0,0}$ and scaling function $\phi_{0,0}$, a function $f$ belongs to $B^s_{pq}(\mathbb{T})$ if and only if $f$ can be represented as in (2.3) and

$$\|f\|_{B^s_{pq}(\mathbb{T})} := \left(|c_0|^q + \sum_{j=0}^{\infty} 2^{jq(s+\frac{1}{2}-\frac{1}{p})}\left(\sum_{h=0}^{2^j-1}|w_{j,h}|^p\right)^{\frac{q}{p}}\right)^{\frac{1}{q}} < \infty. \tag{2.6}$$

There is another (but equivalent) formulation of the Besov norm, showing the connection to the total variation norm more explicitly. Let $\mathcal{F}$ denote the operator which maps a function $f$ to its Fourier coefficients: $\mathcal{F}f = (\widehat{f}(n))_{n\in\mathbb{Z}}$. Also, let $\kappa \in C_0^\infty(\mathbb{R})$ be a non-negative function supported in $[\frac{1}{2}, 2]$, which satisfies $\sum_{j=1}^{\infty}\kappa(js) = 1$ for $s > 2$, and define $\kappa_j(t) = \kappa(2^{-j}|t|)$ for $j \geqslant 1$ and $\kappa_0(t) = 1 - \sum_{j=1}^{\infty}\kappa_j(t)$ for $t \in \mathbb{R}$. Then, supp $(\kappa_0) \subset [-2, 2]$,

supp $(\kappa_j) \subset I_j = [-2^{j+1}, -2^{j-1}] \cup [2^{j-1}, 2^{j+1}]$, and $\sum_{j=0}^{\infty} \kappa_j(t) = 1$. Let $K_j$ be the band-pass filters $K_j f = \mathcal{F}^{-1}(\kappa_j \, \mathcal{F} f)$ defined by the point-wise multiplication in the frequency domain by functions supported near specific annuli. The Besov norm $\|f\|_{B^s_{pq}(\mathbb{T})}$ is equivalent to the following expression [60, 64]:

$$\|f\|_{\tilde{B}^s_{pq}(\mathbb{T})} = \left( \sum_{j=0}^{\infty} 2^{jsq} \|K_j f\|^q_{L^p(\mathbb{T})} \right)^{1/q}. \tag{2.7}$$

In this paper, $p = 1$, $q = 1$ and $s = 1$; then, the Besov norm (2.6) takes the form

$$\|f\|_{B^1_{11}(\mathbb{T})} = |c_0| + \sum_{j=0}^{\infty} \sum_{h=0}^{2^j-1} 2^{j/2} |w_{j,h}|, \tag{2.8}$$

provided that $\psi_{0,0}$ and $\phi_{0,0}$ are once continuously differentiable [62, theorem 1.20]. Furthermore, expression (2.7) takes the form

$$\|f\|_{\tilde{B}^1_{11}(\mathbb{T})} = \sum_{j=0}^{\infty} 2^j \|K_j f\|_{L^1(\mathbb{T})} = \sum_{j=0}^{\infty} \|\mathcal{F}^{-1}(2^j \kappa_j \, \widehat{f})\|_{L^1(\mathbb{T})}. \tag{2.9}$$

Let us give an intuitive reason why (2.9) is related to the total variation norm. Note that $\mathcal{F}^{-1}(|n|\widehat{f}(n))$ is a pseudodifferential operator of order 1, and $\sum_{j=0}^{\infty} \mathcal{F}^{-1}(2^j \kappa_j \, \widehat{f})$ is a similar operation due to the definition of $\kappa_j$. The total variation norm deals with the $L^1(\mathbb{T})$ norm of the first derivatives of $f$, and so does (2.9) in a generalized sense.

We say that the distribution of a random generalized function $f$ is the continuum Besov–$B^1_{11}$ prior if $f$ is of the form (2.3), where $\alpha c_0$ and $\alpha 2^{j/2} w_{j,h}$ are independent real-valued random variables having the Laplace distribution with the scale 1. Recall when a real-valued random variable $W$ has the Laplace distribution with the scale $b > 0$, its probability density function is $\pi_W(w) = (2b)^{-1} e^{-|w|/b}$.

The continuum Besov–$B^1_{11}$ prior can be written formally as

$$\pi(f) \overset{\text{formally}}{=} \exp(-\alpha \|f\|_{B^1_{11}(\mathbb{T})}). \tag{2.10}$$

### 2.2. Discretization

We start by truncating the wavelet expansion (2.3) to a finite number of the coarsest scales. Take $n = 2^N$ with some $N \geqslant 0$ and define $T_n$ by

$$(T_n f)(x) = c_0 + \sum_{j=0}^{N-1} \sum_{h=0}^{2^j-1} w_{j,h} \, \psi_{j,h}(x). \tag{2.11}$$

The right-hand side of (2.11) spans the $n$-dimensional subspace $T_n(B^1_{11}(\mathbb{T}))$, which is parameterized by the wavelet coefficient vector

$$w = [c_0 \quad w_{0,0} \quad w_{1,0} \quad w_{1,1} \quad \cdots \quad w_{(N-1),(2^{(N-1)}-1)}]^T \in \mathbb{R}^n. \tag{2.12}$$

Now we can use (2.11) to construct a prior model for $f_{(n)} = T_n f$ as the following truncated version of formula (2.10):

$$\pi(f_{(n)}) = C_n \exp(-\alpha \|f_{(n)}\|_{B^1_{11}(\mathbb{T})}), \tag{2.13}$$

where $C_n = C_n(\alpha) > 0$ is a normalization constant. The Besov norm $\|f_{(n)}\|_{B^1_{11}(\mathbb{T})}$ in (2.13) takes the simple form

$$\|f_{(n)}\|_{B^1_{11}(\mathbb{T})} = |c_0| + \sum_{j=0}^{N-1} \sum_{h=0}^{2^j-1} 2^{j/2} |w_{j,h}|. \tag{2.14}$$

Note that the definition (2.10) is formal but (2.13) is well defined.

Next, we connect point values and wavelet coefficients of the function $f_{(n)}$ using the fast wavelet transform. Define a grid of $n$ points in the interval $[0, 1)$:

$$x_n(j) = \frac{j}{n} \qquad \text{for} \quad j = 0, 1, \ldots, n-1. \tag{2.15}$$

The 1-periodic real-valued function $f_{(n)} : \mathbb{T} \to \mathbb{R}$ is represented by a vector $f_n$ containing its values at the grid points:

$$f_n = [f_{(n)}(x_n(0)) \quad f_{(n)}(x_n(1)) \quad \cdots \quad f_{(n)}(x_n(n-1))]^T. \tag{2.16}$$

Note that $f_{(n)} \in B^1_{11}(\mathbb{T})$ is a function and $f_n \in \mathbb{R}^n$ is a vector.

Given the vector $f_n$, it is quite easy and computationally effective to evaluate approximately the wavelet coefficients $c_0$ and $w_{j,k}$. Namely, two finite filters $g = [g_0 \cdots g_L]$ and $h = [h_0 \cdots h_L]$, are related to any fixed pair of mother wavelet $\psi_{0,0}$ and scaling function $\phi_{0,0}$. Denote the discrete periodic convolutions between the two filters and the data vector by

$$Gf_n := g * f_n \in \mathbb{R}^n \qquad \text{and} \qquad Hf_n := h * f_n \in \mathbb{R}^n.$$

Furthermore, define a downsampling operator $S : \mathbb{R}^n \to \mathbb{R}^{n/2}$ by

$$Sv = S[v_1\ v_2\ v_3 \cdots v_{n-1}\ v_n]^T := [v_1\ v_3 \cdots v_{n-1}]^T.$$

The wavelet coefficients corresponding to the finest scale basis functions are then given by

$$2^{-n/2} SHf_n = [w_{(N-1),0}\ w_{(N-1),1} \cdots w_{(N-1),(2^{N-1}-1)}]^T.$$

The next coarser level is given by

$$2^{-n/2} SHSGf_n = [w_{(N-2),0}\ w_{(N-2),1} \cdots w_{(N-2),(2^{N-2}-1)}]^T,$$

and finally, the process stops by the computation of the scalars

$$c_0 = 2^{-n/2}(SG)^N f_n,$$
$$w_{0,0} = 2^{-n/2} SH(SG)^{N-1} f_n.$$

We denote the discrete direct and inverse wavelet transform by

$$w = B^{-1} f_n \qquad \text{and} \qquad f_n = Bw, \tag{2.17}$$

respectively. Both direct and inverse transforms can be implemented efficiently using convolutions.

The Besov prior defined by (2.13) and (2.14) can now be written as a probability density on $\mathbb{R}^n$ in terms of the point values (2.16):

$$\pi(f_n) = C_n \exp\left( -\alpha \sum_{\nu=1}^{n} |(WB^{-1} f_n)_\nu| \right), \tag{2.18}$$

where the parameter $\alpha > 0$ controls the width of the distribution and $W$ is a diagonal matrix containing the power-of-2 weights that appear in formula (2.14).

Let us mention explicitly the Haar wavelet case as it will be used below. The periodic Haar mother wavelet $\psi_{0,0} : \mathbb{T} \to \mathbb{R}$ is the discontinuous function defined by

$$\psi_{0,0}(x) = \begin{cases} 1 & \text{for} \quad 0 \leqslant x < 1/2, \\ -1 & \text{for} \quad 1/2 \leqslant x < 1, \end{cases}$$

and the corresponding periodic scaling function is $\phi_{0,0}(x) \equiv 1$. The filters for the Haar wavelet basis are

$$h = \left[\tfrac{1}{\sqrt{2}} \quad \tfrac{-1}{\sqrt{2}}\right], \qquad g = \left[\tfrac{1}{\sqrt{2}} \quad \tfrac{1}{\sqrt{2}}\right].$$

## 3. Total variation prior and the discretization dilemma

### *3.1. The continuum case*

A real-valued function $f \in L^1(\mathbb{T})$ is said to be of bounded variation (BV), $f \in \mathrm{BV}(\mathbb{T})$, if its distributional derivative $Df$ is a signed Radon measure on $\mathbb{T}$ and the total variation of the measure $Df$, denoted by $|Df|$, is a finite measure on $\mathbb{T}$. The variation of the function $\|f\|_{\mathrm{BV}}$ is equal to $|Df|(\mathbb{T})$, i.e. the measure of $\mathbb{T}$ in $|Df|$. If the function $f$ is absolutely continuous, it is differentiable almost everywhere and its derivative $f'$ is in $L^1(\mathbb{T})$, implying that $f$ is in the Sobolev space $f \in W^{1,1}(\mathbb{T})$. For an absolutely continuous function $f$, $Df = f'(x)\mathrm{d}x$, where $\mathrm{d}x$ is the Lebesgue measure on $\mathbb{T}$ and

$$\|f\|_{\mathrm{BV}} = \int_{\mathbb{T}} \left| \frac{\mathrm{d}f}{\mathrm{d}x} \right| \mathrm{d}x.$$

By (2.7), the Besov norm $\|f\|_{\tilde{B}^s_{pq}(\mathbb{T})}$ is equivalent to the norm

$$\|f\|_{\tilde{B}^1_{11}(\mathbb{T})} \equiv \|K_0 f\|_{L^1(\mathbb{T})} + \sum_{j=0}^{\infty} \left\| \frac{d}{dt}(K_j f) \right\|_{L^1(\mathbb{T})}. \tag{3.1}$$

This means that if we decompose $f \in B^1_{11}(\mathbb{T})$ into the components corresponding to dyadic frequency bands, $f = \sum_{j=0}^{\infty} K_j f$, and compute the sum of the BV-norms of all components $K_j f$, $j \geqslant 1$ and the $L^1$ norm of $K_0 f$, we obtain a norm equivalent to $\|f\|_{B^1_{11}(\mathbb{T})}$.

Analogously to formula (2.10), one could try to define a (formal) continuum total variation prior for $f \in W^{1,1}(\mathbb{T})$ or $f \in \mathrm{BV}(\mathbb{T})$ by using the probability density functions

$$\pi(f) \overset{\text{formally}}{=} \exp(-\alpha \|f'\|_{L^1(\mathbb{T})}) \quad \text{or} \quad \pi(f) \overset{\text{formally}}{=} \exp(-\alpha \|f\|_{\mathrm{BV}}). \tag{3.2}$$

However, it is a delicate question whether such formal expressions correspond to well-defined probability distributions. Below, we will define rigorously discrete versions of these priors but depending on the choice of the parameter $\alpha$, these discrete priors do not converge or they converge to a Gaussian distribution when the discretization becomes finer.

### *3.2. Discretization*

Fix any $n \geqslant 2$. Consider piecewise linear roof-top functions $\theta^n_j : \mathbb{T} \to \mathbb{R}$ uniquely determined by the requirement $\theta^n_j(x_n(\nu)) = \delta_{j\nu}$ for $j = 0, \ldots, n-1$ and $\nu = 0, \ldots, n-1$, where the grid points $x_n(\nu)$ are defined by (2.15). Set

$$(T_n f)(x) = \sum_{j=0}^{n-1} f(x_n(j))\, \theta^n_j(x). \tag{3.3}$$

Define the finite-dimensional (improper) total variation prior for $f_{(n)} = T_n f$:

$$\pi(f_{(n)}) = \exp\left( -\alpha_n \int_{\mathbb{T}} \left| \frac{\mathrm{d}f_{(n)}(x)}{\mathrm{d}x} \right| \mathrm{d}x \right), \tag{3.4}$$

where the parameter $\alpha_n > 0$ is determined later. Note that (3.4) gives the same value for any constant function $f_{(n)}(x) \equiv c \in \mathbb{R}$, so the prior does not have a finite integral; thus, the word 'improper' above. However, when combined later with a likelihood model, it will lead to an integrable posterior distribution.

We set $\Delta_n := x_n(2) - x_n(1)$ and proceed to discretize the integrals in (3.4) as Riemann sums. Use the form (3.3) to obtain

$$\int_{\mathbb{T}} \left| \frac{\mathrm{d}f_{(n)}(x)}{\mathrm{d}x} \right| \mathrm{d}x \approx \sum_{\nu=0}^{n-1} \Delta_n \left| \frac{f_{(n)}(x_n(\nu+1)) - f_{(n)}(x_n(\nu))}{\Delta_n} \right|$$
$$= \sum_{\nu=0}^{n-1} |f(x_n(\nu+1)) - f(x_n(\nu))|, \tag{3.5}$$

where $x_n(n) = x_n(0)$ in accordance with the periodic boundary condition.

We wish to work with the vector $f_n \in \mathbb{R}^n$ of point values defined by (2.16). Use (3.5) to write (3.4) as

$$\pi(f_{(n)}) = \exp\left( -\alpha_n \sum_{\nu=1}^{n} |(Df_n)_\nu| \right), \tag{3.6}$$

where the periodic difference matrix $D$ of size $n \times n$ is defined by

$$D = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & -1 & 1 & 0 & \cdots & 0 \\ \vdots & & & \ddots & & & \\ \vdots & & & & \ddots & & \\ 0 & \cdots & & 0 & -1 & 1 & 0 \\ 0 & \cdots & & 0 & 0 & -1 & 1 \\ 1 & \cdots & & 0 & 0 & 0 & -1 \end{bmatrix}. \tag{3.7}$$

Compare formula (3.6) to (2.18).

### *3.3. The discretization dilemma*

It was shown in [39] that the MAP and CM estimates based on the prior density (3.6) are incompatible in the high-resolution limit (as $n \to \infty$).

- If $\alpha_n$ is chosen in (3.6) as a function of $n$ so that the CM estimates converge, then the MAP estimates converge to zero. Also, the CM estimates converge to a Gaussian limit case that favours smoothness and does not preserve edges.
- If $\alpha_n$ is chosen in (3.6) as a function of $n$ so that the MAP estimates converge, then the CM estimates diverge. In this case, the limit of MAP estimates is edge preserving and sparse.

We will illustrate the discretization dilemma numerically in section 6.2.

## 4. Computational measurement models

### *4.1. Simulation of data*

In the measurement model (1.2), we take $A$ to be the periodic convolution operator:

$$(Af)(x) = \int_{\mathbb{T}} K(x-y) f(y)\, \mathrm{d}\sigma(y) \tag{4.1}$$

with the kernel $K$ defined in the interval $x \in [-1/2, 1/2]$ by

$$K(x) = \begin{cases} C_a(x+a)^2(x-a)^2, & \text{for} \quad |x| < a, \\ 0, & \text{for} \quad a \leqslant |x| \leqslant 1/2, \end{cases} \tag{4.2}$$

and the shape (4.2) copied around each integer to make $K$ a 1-periodic function. In (4.2), the parameter $0 < a < 1/2$ is the kernel width and $C_a > 0$ is a normalizing constant ensuring that the non-negative function $K$ has integral 1 over $\mathbb{T}$.

Measurement data $\widehat{m} \in \mathbb{R}^k$ can then be simulated by choosing a suitable noise level $\sigma > 0$ and computing

$$\widehat{m}_\nu = (Af)(x_k(\nu)) + \sigma \cdot \texttt{randn(k,1)}, \qquad 0 \leqslant \nu \leqslant k-1,$$

where $(Af)(x_k(\nu))$ is evaluated numerically by using a high-resolution discretization of (4.1). Thus, we can avoid inverse crime: the data vector $\widehat{m}$ is simulated using a higher density computation grid than the ones that will be used in the likelihood model.

### *4.2. The likelihood model*

For any given $n$, we sample the convolution kernel appropriately and construct a $k \times n$ convolution matrix $A_n$ in the standard way such that

$$A_n f_n \approx [Af_{(n)}(x_k(0)) \quad Af_{(n)}(x_k(1)) \quad \cdots \quad Af_{(n)}(x_k(k-1))]^T.$$

Then, we can implement the likelihood model (1.4) approximately by

$$\pi(m_n|f_n) = C\mathrm{e}^{-\frac{1}{2\sigma^2}\|A_n f_n - m_n\|_2^2} \approx C\mathrm{e}^{-\frac{1}{2\sigma^2}\|Af_{(n)} - m_n\|_2^2} = \pi(m_n|f_{(n)}).$$

## 5. Computing Bayesian estimates

### *5.1. Computing the MAP estimate by optimization*

In view of (3.6) and (2.18), the problem of finding the MAP estimate for either the $B^1_{11}(\mathbb{T})$ prior or the total variation prior becomes an optimization problem with mixed $\ell^2$ and $\ell^1$ norms:

$$f_n^{\mathrm{MAP}} = f_n^{\mathrm{MAP}}(\alpha_n) = \operatorname*{argmin}_{f_n \in \mathbb{R}^n} \left\{ \frac{1}{2\sigma^2}\|A_n f_n - \widehat{m}\|_2^2 + \alpha_n \sum_{\nu=1}^{n} |(Uf_n)_\nu| \right\}, \tag{5.1}$$

where $U = D$ for the total variation prior and $U = WB^{-1}$ for the Besov space prior. The vector $f_n^{\mathrm{MAP}} \in \mathbb{R}^n$ then contains the values of the function $f_{(n)}^{\mathrm{MAP}}(x)$, defined in (1.7), at the grid points.

We use QP methods to solve problems of the form (5.1); for general reference on such methods see [5, 53, 66]. A gradient projection method was used in [24] to solve a problem like (5.1); we use a different method here.

The minimization of (5.1) can be reformulated into a QP form as follows. Denote $Uf_n = u^+ - u^-$, where $u^+, u^- \geqslant 0$. Now, problem (5.1) can be written as

$$\operatorname*{argmin}_{f_n} \left\{ \frac{1}{2\sigma^2} f_n^T A_n^T A_n f_n - \frac{1}{\sigma^2} f_n^T A_n^T \widehat{m} + \alpha_n \mathsf{1}^T u^+ + \alpha_n \mathsf{1}^T u^- + \frac{1}{2\sigma^2}\widehat{m}^T\widehat{m} \right\},$$

where $\mathsf{1} \in \mathbb{R}^n$ is vector of all 1's. Furthermore, we denote

$$z = \begin{bmatrix} f_n \\ u^+ \\ u^- \end{bmatrix}, \qquad Q = \begin{bmatrix} \frac{1}{\sigma^2}A_n^T A_n & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad c = \begin{bmatrix} -\frac{1}{\sigma^2}A_n^T\widehat{m} \\ \alpha_n \mathsf{1} \\ \alpha_n \mathsf{1} \end{bmatrix},$$

and $\delta = \frac{1}{2\sigma^2}\widehat{m}^T\widehat{m}$. Now, the minimization of (5.1) becomes

$$\begin{aligned} &\min_z \quad \frac{1}{2} z^T Q z + c^T z + \delta \\ &\text{such that } \mathcal{A}z = b \quad \text{and} \quad z \geqslant \Lambda = \begin{bmatrix} -M \\ 0 \\ 0 \end{bmatrix}, \end{aligned} \tag{5.2}$$

where $\Lambda$ is a lower bound for the primal variable $z$ related to the following three inequalities: $u^+ \geqslant 0$, $u^- \geqslant 0$ and $f_n > -M$ (with some $M \gg 0$ so that the constraint on $f_n$ is practically ineffective). The matrix $\mathcal{A}$ is an equality constraint matrix related to $Uf_n = u^+ - u^-$. By (3.6), the total variation prior corresponds to

$$U = D, \qquad \mathcal{A} = [D \quad -I \quad I].$$

By (2.18), the Besov space $B^1_{11}(\mathbb{T})$ prior corresponds to

$$U = WB^{-1}, \qquad \mathcal{A} = [WB^{-1} \quad -I \quad I].$$

Now, the estimate $f_n^{\mathrm{MAP}}$ can be computed using QP optimization methods. In this work, we use a primal-dual path-following interior-point method [23, 48, 53, 63, 66].

### 5.2. *Computing the CM estimate by Monte Carlo integration*

We generate a set of (dependent) random samples $\{f_n^{(t)}, t = 1, 2, \ldots, n_s\}$ from the posterior density $\pi_{\mathrm{post}}(f_n|\widehat{m})$ using the Metropolis–Hastings Markov chain Monte Carlo (MCMC) algorithm. As a general reference to MCMC methods, see [44]. The basic idea of using MCMC here is the approximation property

$$f_n^{\mathrm{CM}} = \int_{\mathbb{R}^n} f_n \pi_{\mathrm{post}}(f_n|\widehat{m})\, \mathrm{d}f_n \approx \frac{1}{n_s} \sum_{t=1}^{n_s} f_n^{(t)}.$$

The vector $f_n^{\mathrm{CM}} \in \mathbb{R}^n$ then contains the values of the function $f_{(n)}^{\mathrm{CM}}(x)$, defined in (1.6), at the grid points.

For the generation of the candidate state in the MCMC algorithm, we employ a random one-element update scheme, where we first select the index $\ell$ with the (uniform) probability $1/n$, and then sample a new value for the element $\ell$ from the normal distribution $(f_n')_\ell \sim \mathcal{N}\big((f_n^{(t)})_\ell, \eta^2\big)$, where $f_n^{(t)} \in \mathbb{R}^n$ is the current state of the Markov chain and $\eta$ is the variance of the 1D Gaussian. Thus, we have the proposal density

$$q\big(f_n^{(t)}, f_n'\big) = \frac{1}{n} \frac{1}{\sqrt{2\pi}\,\eta} \exp\left\{ -\frac{1}{2\eta^2} \big((f_n')_\ell - \big(f_n^{(t)}\big)_\ell\big)^2 \right\} \tag{5.3}$$

for random drawing of the candidate state. Using the proposal density (5.3), the implementation of the Metropolis–Hastings algorithm can be described by the following simple procedure. Given the current state $f_n^{(t)}$ of the Markov chain, set $t = 1$ and then do the following.

(1) Draw a candidate state $f_n'$ from the proposal distribution $q\big(f_n^{(t)}, f_n'\big)$.
(2) Compute the Metropolis–Hastings acceptance factor

$$\alpha\big(f_n^{(t)}, f_n'\big) = \min\left\{1, \frac{\pi_{\mathrm{post}}(f_n'|\widehat{m}) q\big(f_n', f_n^{(t)}\big)}{\pi_{\mathrm{post}}\big(f_n^{(t)}|\widehat{m}\big) q\big(f_n^{(t)}, f_n'\big)}\right\} = \min\left\{1, \frac{\pi_{\mathrm{post}}(f_n'|\widehat{m})}{\pi_{\mathrm{post}}\big(f_n^{(t)}|\widehat{m}\big)}\right\}$$

since $q\big(f_n^{(t)}, f_n'\big) = q\big(f_n', f_n^{(t)}\big)$.
(3) Draw $\xi \in [0\ 1]$ from the uniform probability density.
(4) If $\alpha\big(f_n^{(t)}, f_n'\big) \geqslant \xi$, then set $f_n^{(t+1)} = f_n'$, else $f_n^{(t+1)} = f_n^{(t)}$. When $t = n_s$, the desired sample size, stop, else increase $t \to t + 1$ and repeat from step 1.

For the selection of the variance $\eta^2$ of the proposal distribution, we perform trial runs for each model and discretization level, and monitor the acceptance rates, length of the burn-in sequence (the set of initial samples that is discarded) and the integrated autocorrelation time $\tau$ of the chain [44]. Once a suitable value for $\eta$ has been found, we set the final simulation length such that we sample $d\tau$ (with $d = 15\,000$) states after the burn-in sequence. Roughly speaking, this corresponds to having 15 000 independent samples from the posterior distribution for the Monte Carlo integration.

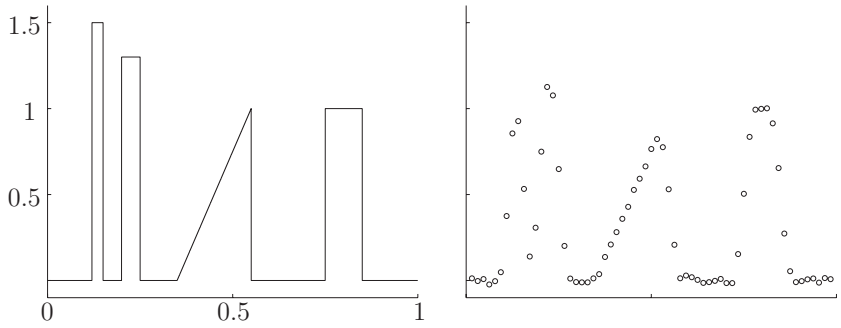


**Figure 1.** (Left) The piecewise linear target function $f$. (Right) The simulated measurement vector $\widehat{m} \in \mathbb{R}^{63}$ corrupted by white noise with a relative amplitude 1%. The coordinate axes limits are the same in both plots.

## 6. Numerical results

### *6.1. Simulation of measurement data*

The data used in the numerical evaluations were simulated using $a = 0.04$ and a piecewise linear function $f$ in (4.1). See figure 1 (left) for a plot of $f$.

Gaussian white noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with a standard deviation of 1% of the maximum of the measurements was added to the simulated data. The resulting measurement vector $\widehat{m}$ consists of $k = 2^6 - 1 = 63$ elements. See figure 1 (right) for a plot of the realization of the measurement.

In the computations below, we use a more dense grid for the generation of data than is used in any of the MAP and CM estimates, thus avoiding the inverse crime (unrealistically good reconstructions).

### *6.2. Bayesian deconvolution using the total variation prior*

Consider the following posterior distribution for the vector $f_n \in \mathbb{R}^n$:

$$\pi_{\text{post}}(f_n|\widehat{m}) = C \exp\left(-\frac{1}{2\sigma^2}\|A_n f_n - \widehat{m}\|_2^2 - \alpha_n \sum_{\nu=1}^{n} |(Df_n)_\nu|\right), \tag{6.1}$$

where $C > 0$ is a normalization constant, $\alpha_n > 0$ is a parameter, $\widehat{m}$ is a realization of the measurement (1.1), the matrix $A_n$ as explained in section 4.2, and $\sigma > 0$ is the standard deviation of the additive Gaussian white noise. The matrix $D$ implements consecutive differences and is defined in (3.7).

A suitable parameter value for $\alpha_n$ for the MAP estimates with the total variation prior was determined using numerical experiments and visual inspection at the fixed discretization level $n = 512$. We took $\alpha_{512} = 89$.

We want to illustrate the discretization dilemma discussed in section 3. To that end, we consider two ways of choosing the parameter $\alpha_n$ as a function of $n$:

$$\text{(i) constant } \alpha_n = 89, \quad \text{(ii) } \alpha_n = 3.93\sqrt{n}.$$

Both (i) and (ii) satisfy $\alpha_{512} = 89$. See figure 2 for the plots of the resulting estimates.

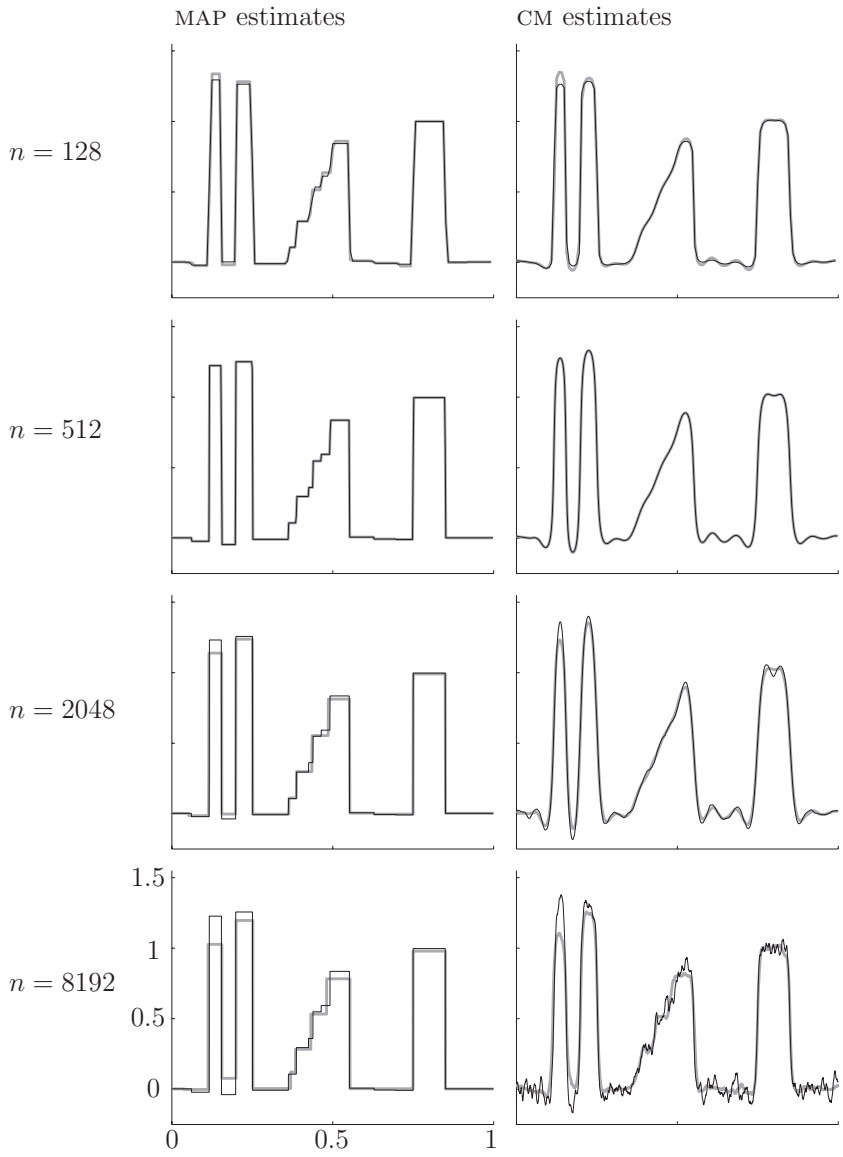


**Figure 2.** Bayesian deconvolutions related to the total variation prior. We computed MAP estimates (left column) and CM estimates (right column) using two choices of $\alpha_n$ parameters as a function of $n$. Both MAP and CM estimates computed using the constant value $\alpha_n = 89$ are drawn with a thin black line. Both MAP and CM estimates computed using the choice $\alpha_n = 3.93\sqrt{n}$ are drawn with a grey thick line. The discretization levels are the same for any two plots in the same row, and the corresponding value of $n$ is indicated on the left. The coordinate axes limits are the same in all the eight plots.

### *6.3. Bayesian deconvolution using the Besov space prior*

The posterior distribution for $f_n$ is now

$$\pi_{\text{post}}(f_n|\widehat{m}) = C \exp\left(-\frac{1}{2\sigma^2}\|A_n f_n - \widehat{m}\|_2^2 - \alpha_n \sum_{\nu=1}^{n} |(WB^{-1}f_n)_\nu|\right), \tag{6.2}$$

where $C > 0$ is a normalization constant, $\alpha_n > 0$ is a parameter, $\widehat{m}$ is a realization of the measurement (1.1), the matrix $A_n$ as explained in section 4.2, and $\sigma > 0$ is the standard deviation of the additive Gaussian white noise. The matrix $B^{-1}$ denotes the wavelet transform and $W$ is a diagonal weighting matrix defined by (2.14) and (2.18).

We consider two different wavelet bases: the Haar basis with discontinuous mother wavelet and filter length 2, and the Daubechies 7 basis with continuously differentiable mother wavelet and scaling function and filter length 14. The Daubechies 7 wavelets provide a theoretically rigorous basis for the space $B^1_{11}(\mathbb{T})$ according to [62, theorem 1.20].

We start with a numerical experiment at a fixed resolution $n = 128$. We compare the Bayesian estimates using the Haar and Daubechies 7 bases and parameters $\alpha_{128}$ with different orders of magnitude. The result is shown in figure 3. It seems that estimates using the Daubechies 7 basis do not preserve edges, so we concentrate below only on estimates using the Haar basis. Note from figure 3 that the CM and MAP estimates are very similar when $\alpha$ is large, but when $\alpha$ is small, the CM estimate becomes less sparse and less edge preserving than the MAP estimate. Furthermore, in the following, we propose a novel sparsity-based method for the selection of the parameter $\alpha$ in (6.2).

### *6.4. Sparsity-based selection of the prior parameter*

If the parameter $\alpha_n$ in (6.2) is very large, then the prior part dominates when computing MAP estimates, leading asymptotically to a zero estimate as $\alpha_n \to \infty$. For a proof of this fact, see the remark at the end of appendix A. Hence, there will be no wavelet coefficients that differ from zero. This can be seen in figure 5: the sparsity levels of the estimates approach zero as $\alpha_{256}$ grows.

On the other hand, taking $\alpha_n$ close to zero removes all regularizing properties of the inversion, and the resulting MAP estimates will be typically quite oscillatory, and also very unstable with respect to changes in the noise or computational parameters. Presumably, almost all wavelet coefficients need to be nonzero to represent such erratic functions. This is also evident from figure 5: the number of nonzero coefficients becomes maximal when $\alpha_{256}$ tends to zero.

For finite $0 < \alpha_n < \infty$ ranging between the extremes, we expect some kind of transition from very sparse to not sparse. Figure 5 suggests that as $\alpha_{256}$ decreases from very large values towards zero, there is a region of almost monotonic growth followed by a regime featuring erratic behaviour as the sparsity levels vary wildly as a function of $\alpha_{256}$.

Motivated by the above observations, we propose to choose $\alpha_n$ in (6.2) as follows.

Assume that we have *a priori* knowledge about the sparsity of our unknown, i.e. we know the number $\widehat{S}$ of nonzero wavelet coefficients $\langle f, \psi_{j,h}\rangle$. We denote

$$S(\alpha_n) = \#\Big(\big\{ j \,:\, 1 \leqslant j \leqslant N,\ \big|\langle f_n^{\mathrm{MAP}}(\alpha_n), \psi_{j,h}\rangle\big| > \kappa \big\}\Big), \tag{6.3}$$

where $\#(E)$ denotes the number of elements in the set $E$ and (for theoretical purposes) $\kappa = 0$. In practical computations, it is advisable to take a small but positive $\kappa > 0$.

Now, compute the estimates

$$f_n^{\mathrm{MAP}}\big(\alpha_n^{(1)}\big), f_n^{\mathrm{MAP}}\big(\alpha_n^{(2)}\big), \ldots, f_n^{\mathrm{MAP}}\big(\alpha_n^{(M)}\big) \tag{6.4}$$

corresponding to a finely spaced collection of parameters

$$0 < \alpha_n^{(1)} < \alpha_n^{(2)} < \ldots < \alpha_n^{(M)} < \infty.$$

Here, it is important to take $\alpha_n^{(1)}$ so small that the wavelet coefficients of $f_n^{\mathrm{MAP}}\big(\alpha_n^{(1)}\big)$ are (almost) all nonzero, and $\alpha_n^{(M)}$ so large that the estimate $f_n^{\mathrm{MAP}}\big(\alpha_n^{(M)}\big)$ has less than $\widehat{S}$ nonzero wavelet coefficients.

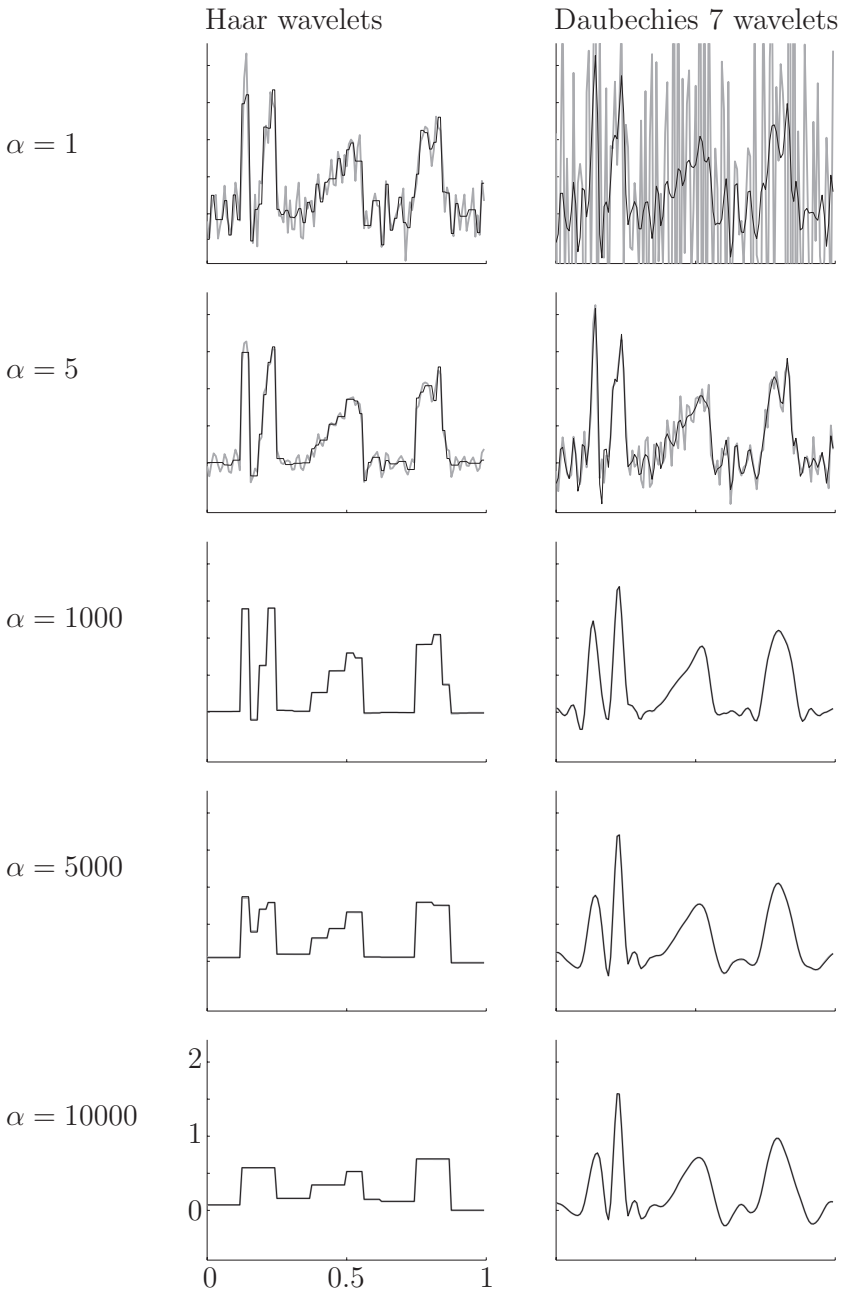


**Figure 3.** Computational MAP estimates (thin black line) and CM estimates (thick solid grey line) with the Besov space $B^1_{11}(\mathbb{T})$ prior at a fixed discretization level $n = 128$ and varying value of the parameter $\alpha_{128}$. (Left) Estimates computed using Haar wavelets. (Right) Estimates computed using Daubechies 7 wavelets. The axis limits and scales in all the ten plots are the same.

The idea now is to choose $\alpha_n := \alpha_n^{(i)}$ for such an index $i$ that

$$S(\alpha_n) = S\big(\alpha_n^{(i)}\big) \approx \widehat{S}.$$

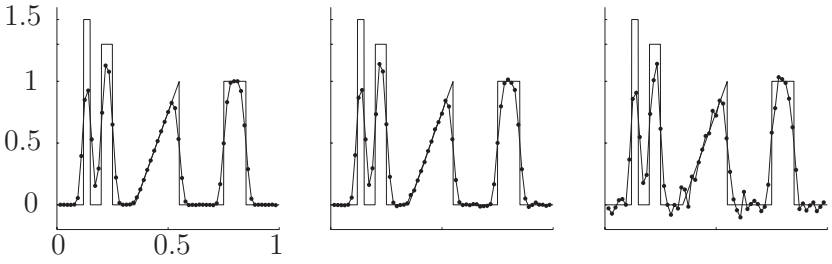


**Figure 4.** Discrete data with the standard deviation of the noise 0.1%, 1% and 5% (left to right), plotted using dots and connecting lines.

**Table 1.** Relative errors of MAP and CM estimates with the Besov space $B^1_{11}(\mathbb{T})$ prior using the Haar wavelet basis. Relative errors are computed with respect to the estimates at the resolution $n = 8192$.

| $n$ | 64 | 128 | 256 | 512 | 1024 | 2048 | 4096 | 8192 |
|---|---|---|---|---|---|---|---|---|
| $\delta_{\mathrm{MAP}}$ | 0.28 | 0.31 | 0.23 | 0.16 | 0.10 | 0.07 | 0.03 | 0 |
| $\delta_{\mathrm{CM}}$ | 0.25 | 0.25 | 0.20 | 0.15 | 0.11 | 0.09 | 0.07 | 0 |

However, such a choice is not unique in general as we see from figure 5: the sparsity levels of the estimates (6.4) take values above and below $\widehat{S}$ in an oscillatory fashion. We decide to find the largest index $j$ for which $S(\alpha_n^{(i)}) > \widehat{S}$ and set $\alpha_n := \alpha_n^{(j+1)}$. There are surely other strategies available that may do a better job in some sense; finding them is left for future studies.

Figure 5 shows the results of applying the choice rule to three data sets with different standard deviations of the noise, i.e. $\sigma = 5\%$, $\sigma = 1\%$ and $\sigma = 0.1\%$. Figure 4 shows the realizations of these measurements. Here, $n = 256$, $\hat{S} = 82$, $\kappa = 10^{-9}$ and $\alpha_n^{(j)}$ ranges in the interval $[10^{-5}, 10^8]$ with $M = 500$. Here, the selection $\hat{S} = 82$ comes from the number of nonzero wavelet coefficients in the representation of the true unknown in the Haar wavelet basis at the level $n = 256$, representing an optimal case where the number of nonzero coefficients is exactly known.

According to the proposed method for selecting $\alpha$ only based on the *a priori* knowledge of the sparsity of the target function, we have $\alpha = 5.39$ for the test case with the 1% noise level as figure 5 illustrates. With this choice, we obtain $S(5.39) = 44$ nonzero coefficients in the MAP estimate. In the following, the selected value $\alpha = 5.39$ is used to compute the MAP and CM estimates with Besov space prior in all discretization levels.

Figure 6 shows Bayesian estimates with the Besov space prior using Haar wavelets and various discretization levels. The superposition of the same estimates in the same plot is shown in figure 7.

In order to study the convergence of the computational estimates quantitatively, we computed relative errors

$$\delta = \frac{\|f_{\mathrm{E}} - f\|_{L^2}}{\|f_{\mathrm{E}}\|_{L^2}} \tag{6.5}$$

with respect to the highest resolution computed estimate with $n = 8192$. See table 1 for the results.

We wish to measure the sparsity of the estimates quantitatively. To that end we count the percentage of scaled wavelet coefficients $\alpha 2^{j/2} w_{j,h}$ that have $|\alpha 2^{j/2} w_{j,h}| < t$ with the prior

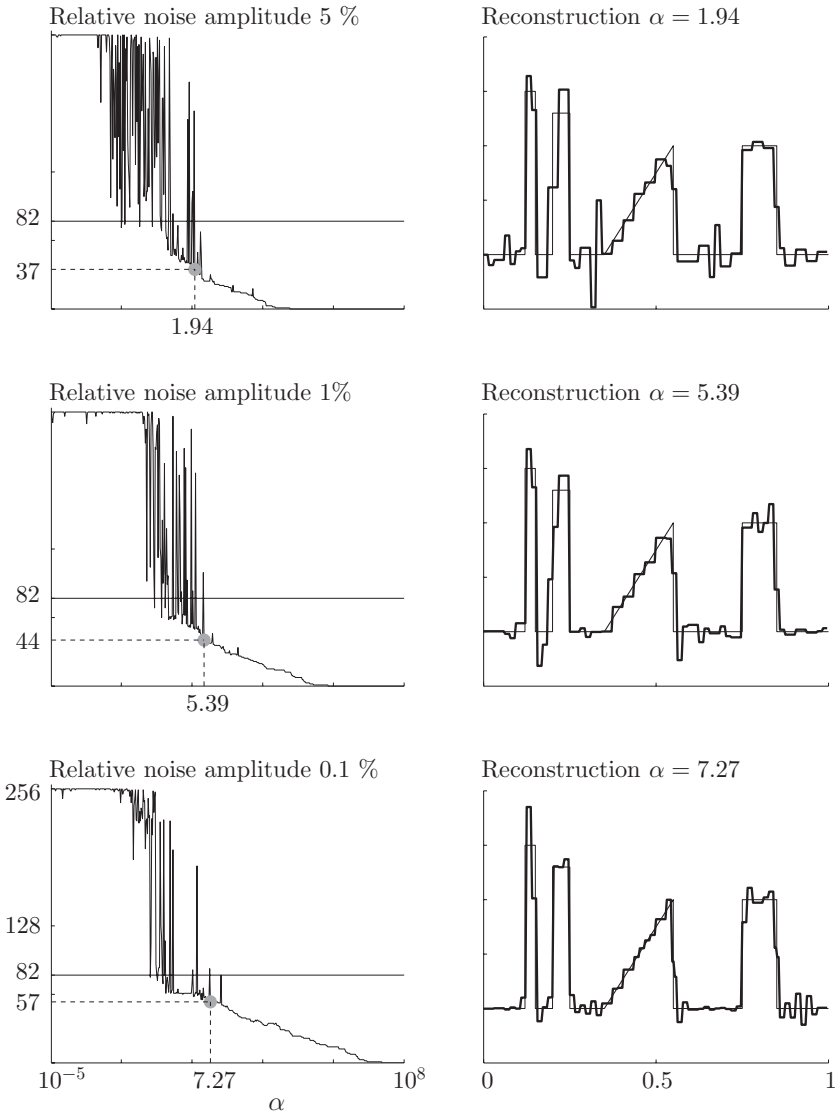


**Figure 5.** Sparsity-based choice of the regularization parameter. (Left) Plots of numbers of nonzero wavelet coefficients in MAP estimates computed with various values of parameter $\alpha$. Results with three different levels of noise $\varepsilon$ are presented. (Right) Original function (thin line) and MAP estimate (thick line) using parameter $\alpha$ determined by *a priori* knowledge of the numbers of nonzero wavelet coefficients in the unknown.

probability of $10^{-6}$. By this criterion, we obtain $t = 10^{-6}$ as tolerance for the numerical zero in the scaled wavelet coefficients when measuring sparsity. The results are shown in table 2.

## 7. Discussion

We investigate computational Bayesian inversion and look for noise-robust estimates that are edge preserving and behave consistently at different resolutions.

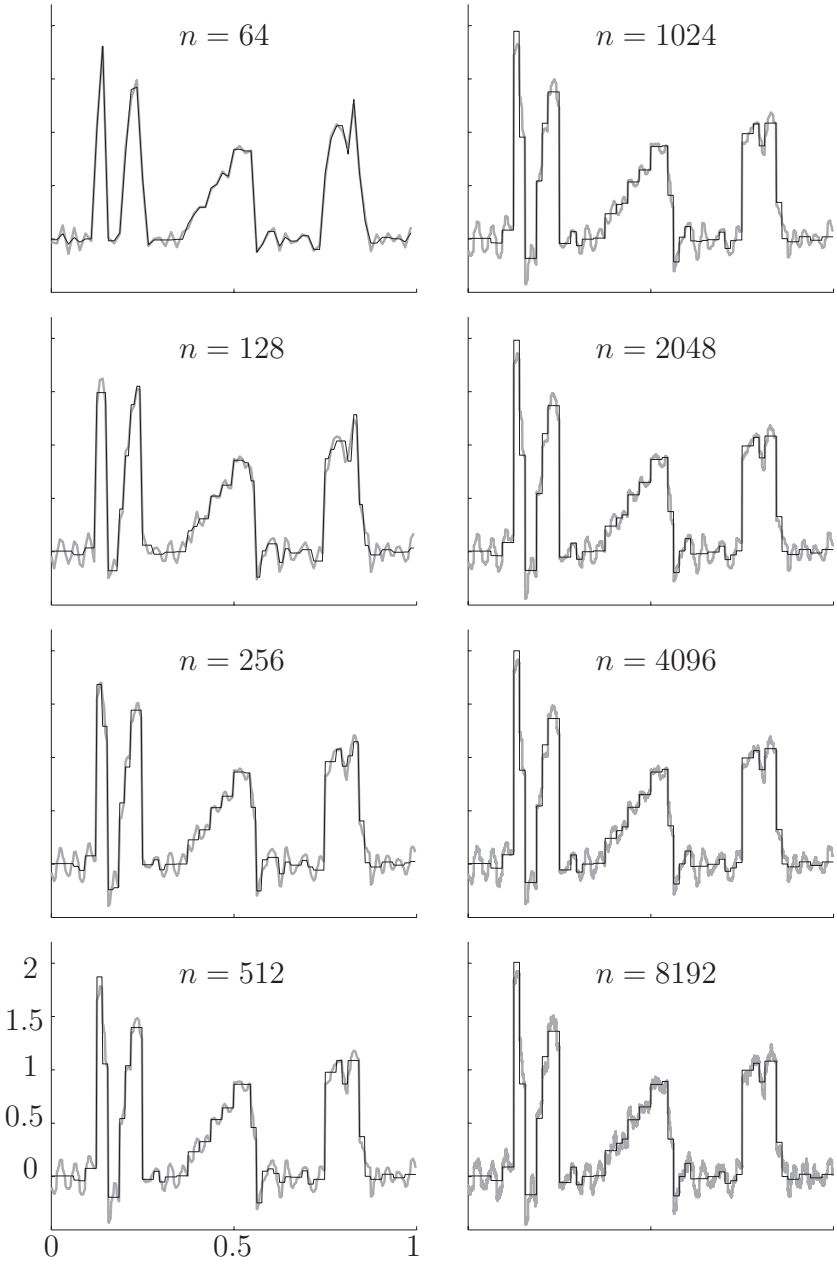


**Figure 6.** Computational MAP estimates (thin black line) and CM estimates (thick solid grey line) with the Besov space $B^1_{11}(\mathbb{T})$ prior using Haar wavelets and various levels of discretization. The coordinate axes limits are the same in all the eight plots. See figure 7 for a superposition of all the estimates in this figure.

As a reference method to the presented Besov-space-based method, we computed MAP and CM estimates using the total variation prior. In addition, we provide new numerical examples illustrating the discretization dilemma explained in [39] and section 3. In figure 2, we see how using a constant parameter $\alpha_n \equiv \alpha \in \mathbb{R}$ leads to convergent MAP estimates

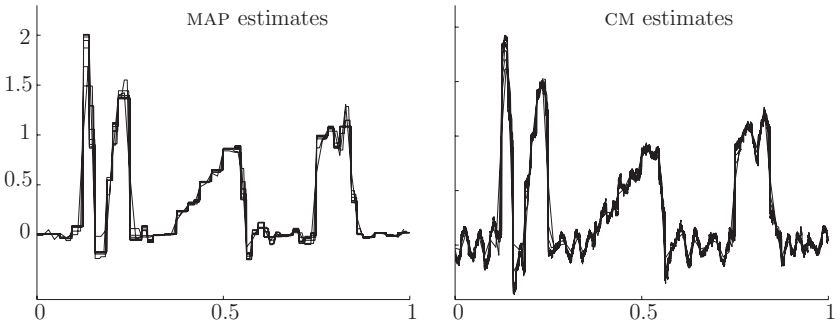


**Figure 7.** Superposition of the estimates shown in figure 6. Note the quite close agreement of the estimates over a significant range of discretization levels ($64 \leqslant n \leqslant 8192$). This is due to the convergence properties of the discretization-invariant Besov space prior.

**Table 2.** Sparsity of MAP and CM estimates with the Besov space $B^1_{11}(\mathbb{T})$ prior using the Haar wavelet basis. The numbers present the numbers of nonzero scaled wavelet coefficients $\alpha 2^{j/2} w_{j,h}$ in the estimates computed using $(10^{-6})$ as a tolerance value. Note that the number of nonzero coefficients in the MAP estimates is roughly equal at all discretizations.

| | $n = 64$ | 128 | 256 | 512 | 1024 | 2048 | 4096 | 8192 |
|---|---|---|---|---|---|---|---|---|
| Number of nonzero coefficients in CM | 64 | 128 | 256 | 512 | 1024 | 2048 | 4095 | 8191 |
| Number of nonzero coefficients in MAP | 47 | 41 | 44 | 40 | 42 | 41 | 41 | 42 |

and divergent CM estimates. On the other hand, a choice of the form $\alpha_n = \alpha_0\sqrt{n}$ is the only possibility to achieve convergent CM estimates. However, such a choice leads to MAP estimates converging to zero and to CM estimates that are smooth in the high resolution limit, thus failing to preserve edges. Our conclusion from the results shown in figure 2 is the same as in [39]: *the total variation prior cannot be used for discretization-invariant Bayesian inversion*.

The Besov space $B^1_{11}(\mathbb{T}^d)$ prior was introduced and studied theoretically in [40]. The space is interesting because it allows a norm based on $\ell^1$ sums of wavelet coefficients and therefore promotes sparsity analogously to a total variation penalty. However, unlike the total variation prior, the $B^1_{11}(\mathbb{T}^d)$ prior is discretization invariant: convergence of $B^1_{11}(\mathbb{T}^d)$-based CM estimates in the high-resolution limit is proved in [40], and the corresponding convergence result for MAP estimates is given in appendix A.

This paper reports the first numerical study of $B^1_{11}$ priors involving the computation of both CM and MAP estimates. The test problem is one-dimensional periodic convolution with the $B^1_{11}(\mathbb{T})$ prior. Our initial approach was to choose a smooth enough mother wavelet and scaling function (Daubechies 7) to provide a theoretically valid wavelet basis for the $B^1_{11}(\mathbb{T})$ space. However, while producing convergent estimates and sparsity in the wavelet transform domain, the Daubechies 7 approach failed to preserve edges. This is a consequence of the smoothness of the basis functions combined with the truncation of the wavelet expansion. See figure 3 (right) for plots of the estimates using the Daubechies 7 basis. We remark that the lack of staircasing in the reconstruction of the linear ramp part of the signal may be an advantage for applications where edge preservation is not the top priority.

Next, we replaced the smooth Daubechies 7 basis by the discontinuous Haar basis, but still used the weighted sum formula corresponding to the $B^1_{11}(\mathbb{T})$ norm. In other words, we used formula (6.2) with $B^{-1}$ denoting the wavelet transform with the Haar basis.

We propose a new, sparsity-based method for choosing $\alpha_n$ for (6.2) in section 6.4. Assume that *a priori* information is available about the number of nonzero wavelet coefficients in the unknown function. (Such information may be available from technical properties of the measurement, as in the recovery of corrupted dual-tone multi-frequency signals used for telecommunications over analog telephone lines; there, one knows that the clean signal consists of exactly two pure frequencies. Another source is the use of image atlases: for example, in medical imaging one can use a set of CT slice images for measuring typical sparsity values and then use this knowledge as *a priori* in limited-angle tomography.) We compute MAP estimates with a large set of parameters as shown in figure 5 and select the value that leads to an estimate with approximately the *a priori* known level of sparsity. This choice rule seems to work robustly over a range of noise levels, as illustrated in figure 5. We compared our new choice rule to the L-curve method with the following result: the L-curve method did not always work as the curve did not have a recognizable L-shape; when it did work, it gave almost the same parameter values as the new rule.

However, there are a couple of obvious concerns with the suggested choice rule. First, the behaviour of the sparsity levels of the estimates as a function of $\alpha_n$ is very erratic when $\alpha_n$ is small. Second, the computational cost is quite high as one needs to solve a large number of optimization problems. Both of these problems may be overcome, e.g., by computing less estimates and fitting a parametric curve such as a polynomial to the data points. According to our preliminary tests, this method works nicely. Such improvements are outside the scope of this paper, however, as we want to illustrate the new choice rule in the 'raw' form in this first report about it.

We illustrate Bayesian estimates with the Besov prior at different discretization levels by computing MAP and CM estimates using $\alpha_n \equiv \alpha_{256}$, where $\alpha_{256}$ is chosen with the new choice rule at the resolution $n = 256$. The estimates are shown in figures 6 and 7; the MAP estimates are clearly sparse and edge preserving, while the CM estimates with the selected $\alpha$ preserve edges only approximately and are not sparse. However, figure 3 indicates that the CM estimates become sparse and very similar to the MAP estimate when the width of the prior decreases ($\alpha$ increases). Table 1 shows that both MAP and CM estimates converge as the discretization is refined. This convergence is in accordance with the theory of discretization invariance and can be intuitively understood as follows: when $n$ is taken larger in the expression

$$\frac{1}{2\sigma^2}\|A_n f_n - \widehat{m}\|_2^2 + \alpha \sum_{\nu=1}^{n} |(WB^{-1} f_n)_\nu|,$$

then both terms approximate their well-defined continuum counterparts, namely $\frac{1}{2\sigma^2}\|Af - \widehat{m}\|_2^2$ and $\alpha\|f\|_{B^1_{11}(\mathbb{T})}$, respectively. If these continuum quantities are in balance, then that balance is approximated by finite computations better and better as $n$ grows.

Let us stress that the new choice rule is used only at the resolution $n = 256$, and the chosen parameter is then fixed and used at all resolutions. The resulting MAP estimates are nevertheless quite sparse for all $n$; according to table 2, they even have roughly the same number of nonzero wavelet coefficients. Why is this? The answer is given in appendix A, where we show that the number of nonzero coefficients 'freezes' at some $N_0$ and remains constant for $n \geqslant N_0$. Table 2 seems to indicate that such freezing happens in our example problem for quite a small value of $N_0$. This is good from the computational point of view as there is no need to compute with very high values of $n$ because the same coefficients can be recovered with less effort using smaller $n$.

It is surprising how close the MAP and CM estimates are to each other in figure 6 when $\alpha$ is large and how they differ for smaller $\alpha$. In the case of purely Gaussian posterior density, the MAP and CM estimates are known to coincide, but there is no obvious reason why this should be the case in this non-Gaussian Besov prior framework. Actually, it is easy to construct low-dimensional examples where the $B^1_{11}$-based MAP and CM estimates are very different; see appendix C. Thus, there seems to be some (as of now unknown) reason for the estimates to be almost the same when $\alpha$ is large and different when $\alpha$ is small (but not zero), at least in the deconvolution example considered here. Let us stress that the MAP and CM estimates were computed with different algorithms and different initial guesses, so the observed closeness is not due to numerical inaccuracy or computational convergence failure.

## 8. Conclusion

We have shown by numerical evidence and mathematical proofs that the combination of Haar basis functions and $B^1_{11}(\mathbb{T})$ norm formula leads to discretization-invariant Bayesian estimates. Furthermore, the MAP estimates are sparsity promoting and edge preserving, and the CM estimates appear to be very close to the MAP estimates for large values of the prior parameter $\alpha$. In addition, we proposed and demonstrated a completely new Bayesian method for choosing the prior parameter. The new method seems to perform robustly under noisy conditions.

## Acknowledgments

This work was supported by the Academy of Finland under Projects 119270, 218183, 140731, 141094, the CSI Speech project 134868 belonging to the LASTU Programme on Computational Science, and the Finnish Centre of Excellence in Inverse Problems Research 2006–2011 (decision number 213476). Also, the work was supported in part by the Qvision project funded by Forestcluster Ltd. This project was partially conducted at the Mathematical Sciences Research Institute, Berkeley, CA, whose hospitality is gratefully acknowledged. The authors thank Janne Tamminen, MSc, for compiling the MCMC codes to a grid computer for parallelization of the MCMC simulations.

## Appendix A. Convergence of sparse MAP estimates

We start with the convergence result for MAP estimates when $A$ is a linear finite-rank operator and the Besov norm on the $d$-dimensional torus $\mathbb{T}^d$ is given in terms of the wavelet coefficients. In appendix B, we will consider the convergence results for more general operators and for other (but equivalent) Besov norms.

To formulate the Besov norm in terms of wavelets, we use the standard tensor product wavelet basis construction in dimensions $d > 1$ and number the wavelet coefficients by just one integer index. Regardless of the dimension $d$, this is done starting from the coarsest ones and moving on to the finer scale wavelets as explained in detail in [40]. Then, the coefficients are collected into the following infinite-dimensional vector:

$$\begin{aligned} w &= [w_1, w_2, w_3, \ldots]^T \\ &:= [c_{0,0}, \ldots, c_{0,K(0)}, w_{0,0}, \ldots, w_{0,K(0)}, w_{1,0}, \ldots, w_{1,K(1)}, \ldots]^T, \end{aligned}$$

where $K(j)$ denotes the finite number of wavelet basis functions at scale $j$.

According to [40, 49], the Besov space can be endowed with the norm $\|f\|_{B^1_{11}(\mathbb{T}^d)}$ written in the form

$$\|f\|_{B^1_{11}(\mathbb{T}^d)} = \sum_{\nu=1}^{\infty} 2^{(1-d/2)j(\nu)}|w_\nu| =: \sum_{\nu=1}^{\infty} \beta_\nu |w_\nu| := \|w\|_{\ell^1_\beta}, \tag{A.1}$$

where $j(\nu)$ denotes the scale corresponding to the $\nu$th wavelet coefficient listed in the vector $w$. The essential point of formula (A.1) is the sequence $\beta = (\beta_\nu)_{\nu=1}^{\infty}$ of strictly positive weights.

Then, the computation of the MAP estimate

$$\operatorname*{argmin}_{f_{(n)}\in T_n(X)} \left\{ \frac{1}{2\sigma^2}\|Af_{(n)} - \widehat{m}\|^2_{\mathbb{R}^k} + \alpha \|f_{(n)}\|_{B^1_{11}(\mathbb{T}^d)} \right\}$$

with $\alpha > 0$ is equivalent to the following optimization problem:

$$\operatorname*{argmin}_{w\in \mathcal{T}_n\ell^1_\beta} \left\{ \frac{1}{2\sigma^2}\|ABw - \widehat{m}\|^2_{\mathbb{R}^k} + \alpha \|w\|_{\ell^1_\beta} \right\}, \tag{A.2}$$

where

$$\mathcal{T}_n : \ell^1 \to \ell^1, \qquad (\mathcal{T}_n w)_j = \begin{cases} w_j & \text{if } j \leqslant n, \\ 0 & \text{otherwise.} \end{cases} \tag{A.3}$$

Denoting $v_\nu = \beta_\nu w_\nu$ and $\mathcal{A} = AB\,\mathrm{diag}(\beta_\nu^{-1})$, we come to the minimization problem

$$\operatorname*{argmin}_{v\in \mathcal{T}_n\ell^1_\beta} \left\{ \frac{1}{2\sigma^2}\|\mathcal{A}v - \widehat{m}\|^2_{\mathbb{R}^k} + \alpha \|v\|_{\ell^1} \right\}. \tag{A.4}$$

Let us consider the function $F : \ell^1 \to \mathbb{R}$,

$$F(v) = \frac{1}{2\sigma^2}\|\mathcal{A}v - \widehat{m}\|^2_{\mathbb{R}^k} + \alpha \|v\|_{\ell^1}.$$

Below, we will assume that $\mathcal{A} : \ell^p \to \mathbb{R}^k$ is a bounded linear operator for $p = 2$. Note that, then, $\mathcal{A} : \ell^1 \to \mathbb{R}^k$ is bounded, too.

By [27], there exists a (possibly non-unique) point $v^0 = (v^0_j)_{j=1}^{\infty} \in \ell^1$, where $F$ obtains its minimal value, i.e. $F(v^0) = \min_{v\in\ell^1} F(v)$. It is also well known that $v^0$ is sparse, i.e. only a finite number of elements $v^0_j$ of the sequence $v^0$ are nonzero; see, e.g., [16, 26–28]. Below, we give explicit estimates for the parameter $N_0 = N_0(\alpha)$ associated below with the sparsity. For the convenience of the reader, we next recall the proof of the sparsity property of the solution using the convex analysis techniques used, e.g., in [26, 27].

Observe that the minimum of $F$ is less or equal than $C_1 = F(0) = \frac{1}{2}\sigma^{-2}\|\widehat{m}\|^2_{\mathbb{R}^k}$. Let us consider the subgradient $\partial F(v) \subset (\ell^1)^* = \ell^\infty$ of the function $F$ at $v \in \ell^1$; see, e.g., [59, definition 10.14]. We see that

$$b = (b_1, b_2, b_3, \ldots) \in \partial F(v)$$

if and only if

$$\begin{aligned} b_j &= \frac{1}{\sigma^2}(\mathcal{A}^*(\mathcal{A}v - \widehat{m}))_j + \alpha \quad \text{if } v_j > 0, \\ b_j &= \frac{1}{\sigma^2}(\mathcal{A}^*(\mathcal{A}v - \widehat{m}))_j - \alpha \quad \text{if } v_j < 0, \\ b_j &- \frac{1}{\sigma^2}(\mathcal{A}^*(\mathcal{A}v - \widehat{m}))_j \in [-\alpha, \alpha], \quad \text{if } v_j = 0. \end{aligned}$$

As $F$ obtains its minimal value at $v^0$, we have that $0 \in \partial F(v^0)$; see, e.g., [59, lemma 10.15].

Let us write $\mathcal{A}$ as $(\mathcal{A}v)_j = \sum_{\nu=1}^{\infty} a_{j\nu} v_\nu$ for $j = 1, 2, \ldots, k$. As $\mathcal{A}^* : \mathbb{R}^k \to \ell^2$ is a bounded linear operator, we see that $(a_{j\nu})_{\nu=1}^{\infty} \in \ell^2$. Thus, there is $N_0(\alpha) \in \mathbb{Z}_+$ such that if $n > N_0(\alpha)$, then

$$\sum_{j=1}^{k} |a_{jn}| \leqslant \varepsilon_0, \quad \varepsilon_0 := \frac{\sigma^2 \alpha}{2} (1 + C_1 \alpha^{-1} \|A\|_{L(\ell^2, \mathbb{R}^k)} + \|\widehat{m}\|_{\mathbb{R}^k})^{-1}.$$

Note that $N_0(\alpha)$ is determined by the elements $a_{j\nu}$ of the infinite matrix corresponding to $\mathcal{A}$ and the norm $\|\widehat{m}\|_{\mathbb{R}^k}$. Moreover, as $\|v^0\|_{\ell^2} \leqslant \|v^0\|_{\ell^1} \leqslant \alpha^{-1} C_1$, we see that if $j > N_0(\alpha)$, then

$$\frac{1}{\sigma^2} |(\mathcal{A}^*(\mathcal{A}v^0 - \widehat{m}))_j| \leqslant \frac{\alpha}{2}.$$

As $0 \in \partial F(v^0)$, we see thus that if $j > N_0(\alpha)$, then $v_j^0 = 0$. This means that the minimizing point $v^0 \in \ell^1$ of $F$ is sparse, i.e. only finitely many elements of the sequence $v^0$ are nonzero. Moreover, for $n > N_0(\alpha)$, the solution of the minimization problem

$$\operatorname*{argmin}_{v \in \ell^1} \left\{ \frac{1}{2\sigma^2} \|\mathcal{A}v - \widehat{m}\|_{\mathbb{R}^k}^2 + \alpha \|v\|_{\ell^1} \right\} \tag{A.5}$$

is $v^0 \in T_n \ell^1$. Hence, when $n > N_0(\alpha)$, the solution of the minimization problem

$$\operatorname*{argmin}_{v \in T_n \ell^1} \left\{ \frac{1}{2\sigma^2} \|\mathcal{A}v - \widehat{m}\|_{\mathbb{R}^k}^2 + \alpha \|v\|_{\ell^1} \right\} \tag{A.6}$$

coincides with the solution of (A.5). Roughly speaking, this means that if we increase the parameter $n$, the solutions of the minimization problems (A.6) 'freeze' at some point, i.e. they do not change after $n$ exceeds the value $N_0(\alpha)$.

We note that there exists $\alpha_0$ such that if $\alpha > \alpha_0$, then

$$\frac{\sigma^2 \alpha}{2} (1 + C_1 \alpha^{-1} \|A\|_{L(\ell^2, \mathbb{R}^k)} + \|\widehat{m}\|_{\mathbb{R}^k})^{-1} > \sup_{n \in \mathbb{Z}_+} \sum_{j=1}^{\infty} |a_{jn}|.$$

Then, $N(\alpha) = 0$ for $\alpha > \alpha_0$. This implies that for $\alpha > \alpha_0$ the minimizer $v$ has to be $v = 0$.

## Appendix B. Equivalent Besov norms and convergence

Let us consider the issue that in $B_{11}^1(\mathbb{T}^d)$ there are many equivalent norms. For instance, one of the standard norms used in the Besov spaces is the one based on the Paley–Littlewood decomposition of a function; see [62, definition 1.1]. We note that if the previous norm based on the wavelet coefficients is replaced with an equivalent norm, then the sparsity analysis presented in appendix A does not hold. However, using $\Gamma$-convergence techniques, we will show that the solutions of the discretized minimization problems converge to the solution of the infinite-dimensional problem, at least up to a subsequence. Below, as $B_{11}^1(\mathbb{T}^d)$ is isomorphic to $\ell^1$, instead of considering the minimization problem in $B_{11}^1(\mathbb{T}^d)$, we consider the corresponding problem in the space $\ell^1$, which is endowed with a non-standard norm.

The minimization problem (A.2) is a particular case of the more general problem

$$\operatorname*{argmin}_{w \in \mathcal{T}_n \ell^1} \left\{ \|G(w) - m\|_{\ell^2}^2 + \|w\|_E \right\}, \tag{B.1}$$

where $G : \ell^1 \to \ell^2$ is a possibly nonlinear operator and $\|w\|_E$ is a norm in $\ell^1$ being equivalent to its usual norm, i.e. $c_1 \|w\|_{\ell^1} \leqslant \|w\|_E \leqslant c_2 \|w\|_{\ell^1}$ with some $c_1, c_2 > 0$.

In the following, we will consider $\Gamma$-convergence in the weak-$*$ topology of $\ell^1$. The analysis below is closely related to techniques used in [26].

Next, let $\tau^*$ denote the weak-$*$ topology in the dual $X^*$ of a separable Banach space $X$, i.e. $x_n^* \to x^*$ in $\tau^*$ topology of $X^*$ if for all $x \in X$

$$\lim_{n\to\infty} \langle x_n^*, x\rangle = \langle x^*, x\rangle.$$

We note that in a bounded set $B \subset X^*$, the relative topology of $B$ inherited from $(X^*, \tau^*)$ is metrizable, i.e. equivalent to a topology induced by some metric.

Also, let $\tau^s$ be the strong topology (i.e. the norm topology) of the Banach space $X$.

Our goal is to prove the following proposition.

**Proposition 1.** *Let $w \mapsto G(w)$ be a continuous, possibly nonlinear operator from $(\ell^1, \tau^*)$ to $(\ell^2, \tau^s)$, $\|\ \|_E$ be a norm in $\ell^1$ which is equivalent to the usual norm of $\ell^1$, and $\mathcal{T}_n : \ell^1 \to \ell^1$, $n \in \mathbb{Z}_+$ be the projectors given in (A.3). Moreover, let $w_{\mathrm{MIN}}^n$ be some minimizers of*

$$F_n(w) = \begin{cases} \|G(w) - m\|_{\ell^2}^2 + \|w\|_E, & \text{if } w \in \mathcal{T}_n\ell^1, \\ \infty, & \text{otherwise,} \end{cases} \tag{B.2}$$

*i.e. the MAP estimates for the discrete problems. Then, $w_{\mathrm{MIN}}^n$ converge up to a subsequence, i.e. for any strictly increasing sequence $(n_j)_{j=1}^\infty \subset \mathbb{Z}$, there is a subsequence $n_{j_k}$ such that there exists a limit*

$$\lim_{k\to\infty} w_{\mathrm{MIN}}^{n_{j_k}} = w_{\mathrm{MIN}} \quad \text{in } (\ell^1, \tau^*). \tag{B.3}$$

*Moreover, any limit $w_{\mathrm{MIN}}$ in (B.3) is a minimizer of the function $F : \ell^1 \to \mathbb{R}$, $F(w) = \|G(w) - m\|_{\ell^2}^2 + \|w\|_E$.*

*If, in addition to the above assumptions, the map $w \mapsto \|G(w) - m\|_{\ell^2}^2$ is strictly convex, the minimizer of $F$ is unique, and thus, we have the convergence of MAP estimates:*

$$\lim_{n\to\infty} w_{\mathrm{MIN}}^n = w_{\mathrm{MIN}} \quad \text{in } (\ell^1, \tau^*).$$

We use the $\Gamma$-convergence in a separable Banach space $X$ with a topology $\tau$. Let us recall some basic definitions.

In the following, let $\mathcal{B} \subset X$ be a set which, when being endowed with the topology inherited from $(X, \tau)$, is metrizable.

**Definition 1.** *We say that $F_j : \mathcal{B} \to (-\infty, \infty]$ $\Gamma$-converges to $F : \mathcal{B} \to (-\infty, \infty]$ for the topology $\tau$ and denote $F = \Gamma\text{-}\lim_{j\to\infty} F_j$ if the following conditions hold.*

- *For every $x \in \mathcal{B}$ and for every sequence $x_j$ $\tau$-converging to $x$ in $\mathcal{B}$, we have $F(x) \leqslant \liminf_{j\to\infty} F_j(x_j)$.*
- *For every $x \in \mathcal{B}$, there exists a sequence $x_j$ $\tau$-converging to $x$ in $\mathcal{B}$ such that $F(x) \geqslant \limsup_{j\to\infty} F_j(x_j)$.*

We need also the concept of equicoercivity.

**Definition 2.** *A functional $F : \mathcal{B} \to (-\infty, \infty]$ is said to be coercive if for any $t \in \mathbb{R}$ the set $\{x \in \mathcal{B} \mid F(x) \leqslant t\}$ is precompact. Also, we call a sequence of functionals $F_j : \mathcal{B} \to [-\infty, \infty]$, $j \in \mathbb{Z}_+$, equicoercive in topology $\tau$ if for every $t \in \mathbb{R}$, there exists a compact set $K_t \subset \mathcal{B}$ such that $\{x \in \mathcal{B} \mid F_j(x) \leqslant t\} \subset K_t$ for all $j \in \mathbb{Z}_+$.*

Let $F, F_j : \mathcal{B} \to (-\infty, \infty]$, $j \in \mathbb{Z}_+$ be nonlinear maps. The following results are proven in [14].

**Theorem 1.** *Let $F, F_j : \mathcal{B} \to (-\infty, \infty]$, $j \in \mathbb{Z}_+$, be a sequence of equicoercive functionals in topology $\tau$ and $F = \Gamma\text{-}\lim_{j\to\infty} F_j$. Then, the following three properties hold.*

- *If the* $\Gamma$*-limit of* $F_j$ *exists, it is unique and lower semi-continuous.*
- *For any continuous* $H : \mathcal{B} \to \mathbb{R}$*, we have* $F + H = \Gamma\text{-}\lim_{j\to\infty}(F_j + H)$.
- *Let* $x_j \in \mathcal{B}$ *be such that* $\left|F_j(x_j) - \inf_{x\in\mathcal{B}} F_j(x)\right| \leqslant \delta_j$*, where* $\delta_j \to 0$*. Then, any accumulation point* $y \in \mathcal{B}$ *of* $(x_j)_{j=1}^{\infty} \subset \mathcal{B}$ *is a minimizer of* $F$ *and moreover* $\lim_{j\to\infty} F_j(x_j) = F(y)$.

It follows from (iii) that if the assumptions in theorem 1 hold and $x_j$ is a minimizer of $F_j$ for $j \in \mathbb{Z}_+$, then any converging subsequence of $(x_j)_{j=1}^{\infty} \subset \mathcal{B}$ converges to a minimizer of $F$.

Let us endow $\mathbb{Z}_+$ with its co-finite topology $\tau_{\text{cf}}$ so the $C_0(\mathbb{Z}_+)$ coincide with the set of sequences $c_0 = \{(w_j)_{j=1}^{\infty} \in \mathbb{R}^{\mathbb{Z}_+} : \lim_{n\to\infty} w_j = 0\}$. We endow $c_0$ with the sup-norm $\|w\|_\infty = \sum_j |w_j|$. Then, by the Riesz representation theorem in $\mathbb{Z}_+$, the space $\ell^1$ has a pre-dual $c_0$, i.e. $\ell^1 = c_0^*$. By the Banach–Alaoglu theorem, the closed unit ball of $\ell^1$ is weak-$*$ compact.

Assume now that $w^n = (w_j^n)_{j=1}^{\infty} \in \ell^1$, $n = 1, 2, \ldots$, is a sequence which converges to $w = (w_j)_{j=1}^{\infty} \in \ell^1$ in the weak-$*$ topology of $\ell^1$. By the Banach–Steinhaus theorem, $w^n$ is uniformly bounded sequence in $\ell^1$, i.e. there is $C_0 > 0$ such that $\|w_n\|_E \leqslant C_0$. Moreover, the norm is lower semi-continuous in the weak-$*$ topology:

$$\|w\|_E \leqslant \liminf_{n\to\infty} \|w^n\|_E,$$

and using the definition of weak-$*$ convergence with the elements $e^k = (\delta_{kj})_{j=1}^{\infty} \in c_0(\mathbb{Z}_+)$, we see that for any fixed index $j$,

$$\lim_{n\to\infty} w_j^n = w_j.$$

In the following, we will assume that $w \mapsto G(w)$ is a continuous (possibly nonlinear) operator from $(\ell^1, \tau^*)$ to $(\ell^2, \tau^s)$. An example of such an operator is $\mathcal{A}J$, where $J : \ell^1 \to \ell^2$ is the identical embedding and $\mathcal{A} : \ell^2 \to \ell^2$ is a compact linear operator.

Let us denote $B(R) = \{w \in \ell^1 : \|w\|_E < R\}$ and $\overline{B}(R) = \{w \in \ell^1 : \|w\|_E \leqslant R\}$, $R > 0$. The sets $\overline{B}(R) \subset \ell^1$ are closed and metrizable both with respect to the topologies $\tau^*$ and $\tau^s$. Observe that as $F(0) = F_n(0) = G(0) + \|m\|_{\ell^2}^2$, the minima of both $F$ and $F_n$ are obtained, if it exists, in the set $\overline{B}(R_0)$, where $R_0 = G(0) + \|m\|_{\ell^2}^2 < \infty$. In the following, we denote $R_1 = R_0 + 1$ and analyse the restrictions of the maps $F$ and $F_n$ to $B(R_1)$.

As the norm $w \mapsto \|w\|_E$ is lower semi-continuous in $(\ell^1, \tau^*)$, we see that

$$F(w) = \|G(w) - m\|_{\ell^2}^2 + \|w\|_E$$

is lower semi-continuous in $(\ell^1, \tau^*)$. As the set $\overline{B}(R_0)$ is non-empty and compact in $(\ell^1, \tau^*)$, we see that $F$ has a minimum $w_{\text{MIN}} \in \overline{B}(R_0)$:

$$F(w_{\text{MIN}}) = \min_{w\in\ell^1} F(w).$$

Let us consider now the operator $F_n$ defined in (B.2). As $\mathcal{T}_n$ is a finite-rank operator, we see that there exist (possibly non-unique) $w_{\text{MIN}}^n \in \overline{B}(R_0)$ at which $F_n$ obtains its minimum. Moreover,

$$F_n\big(w_{\text{MIN}}^n\big) = \min_{w\in\ell^1} F_n(w) = \min_{w\in\text{Ran}(\mathcal{T}_n)} F(w).$$

Now if $w = (w_j)_{j=1}^{\infty} \in \ell^1$, we see that for all $f \in c_0$

$$\lim_{n\to\infty} \langle \mathcal{T}_n w, f\rangle = \langle w, f\rangle$$

and thus $\mathcal{T}_n w \to w$ in $\tau^*$ as $n \to \infty$.

Let $w \in B(R_1)$. Then, $\lim_{n\to\infty} \mathcal{T}_n w = w$ both in $(\ell^1, \tau^*)$ and in $(\ell^1, \tau^s)$. Thus, $\mathcal{T}_n w \in B(R_1)$ for $n$ large enough. Moreover, as $G : (\ell^1, \tau^*) \to (\ell^2, \tau^s)$ is continuous and $\|\ \|_E$ is continuous in $(\ell^1, \tau^s)$, we see using the definition (B.2) that

$$\lim_{n\to\infty} F_n(\mathcal{T}_n w) = F(w). \tag{B.4}$$

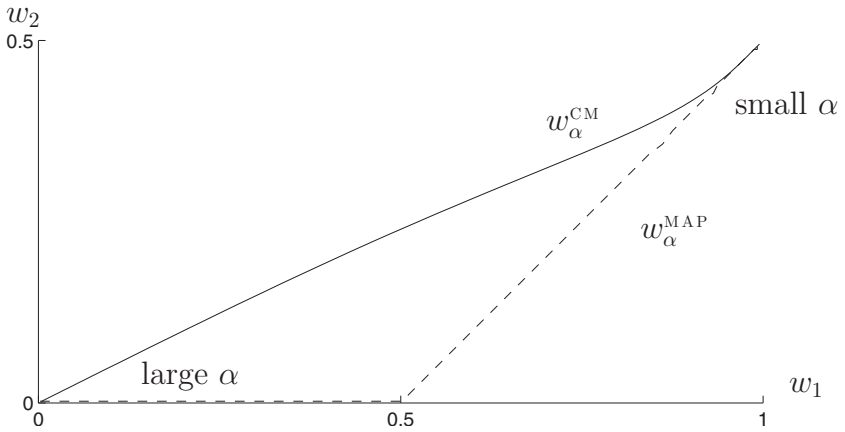


**Figure C1.** The CM estimate $w_\alpha^{\mathrm{CM}}$ is shown with a solid line and the MAP estimate $w_\alpha^{\mathrm{MAP}}$ is shown with a dashed line, both as the functions of $\alpha$. For large enough $\alpha$, the second component of $w_\alpha^{\mathrm{MAP}}$ becomes identically zero. This is a simple example of sparsity.

On the other hand, assume that $w^n \to w$ in $(B(R_1), \tau^*)$ as $n \to \infty$. By (B.2), $F_n(w^n) \geqslant F(w^n)$. Since $G : (\ell^1, \tau^*) \to (\ell^2, \tau^s)$ is continuous and the norm $\| \ \|_E$ is lower semi-continuous in $(\ell^1, \tau^*)$, we have

$$\liminf_{n\to\infty} F_n(w^n) \geqslant \liminf_{n\to\infty} F(w^n) \geqslant F(w). \tag{B.5}$$

This and (B.4) yield that $F = \Gamma\text{-}\lim_{j\to\infty} F_n$ in $(B(R_1), \tau^*)$.

As the sets $\{w \in \ell^1 : \ F_n(w) \leqslant t\}$ and $\{w \in \ell^1 : \ F(w) \leqslant t\}$ are subsets of $t$-balls $\overline{B}(t)$, which are compact in $(\ell^1, \tau^*)$, we see that $F_n$, $n \in \mathbb{Z}_+$, and $F$ form an equicoercive family. Due to this, theorem 1 proves proposition 1.

## Appendix C. Sparsity and the $B_{11}^1$ prior: a simple experiment

Here, we let $w = [w_1 \ w_2]^T \in \mathbb{R}^2$ and $m = [m_1 \ m_2]^T \in \mathbb{R}^2$ and consider the posterior distribution

$$\pi(w|m) = C \exp\left(-\|Aw - m\|_2^2 - \alpha\|w\|_1\right), \tag{C.1}$$

where $C$ is the normalization constant, $0 < \alpha < \infty$ is a parameter, and $A$ is a $2 \times 2$ matrix. The posterior distribution (C.1) is related to the Bayesian use of the $B_{11}^1$ prior in the case of two orthonormal wavelet coefficients $w_1$ and $w_2$.

The CM estimate is defined as the integral

$$w_\alpha^{\mathrm{CM}} := \int_{\mathbb{R}^2} w\, \pi(w|m)\, \mathrm{d}w, \tag{C.2}$$

and the MAP estimate can be defined as a minimizer:

$$w_\alpha^{\mathrm{MAP}} := \operatorname*{argmin}_{w\in\mathbb{R}^2} \{\|Aw - m\|_2^2 + \alpha|w_1| + \alpha|w_2|\}. \tag{C.3}$$

Consider the limit $\alpha \to 0$, when only the quadratic term is left in (C.3). Then, the posterior becomes Gaussian, implying that the MAP and the CM estimates coincide. Consequently, we have

$$\lim_{\alpha\to 0} w_\alpha^{\mathrm{MAP}} = A^{-1}m = \lim_{\alpha\to 0} w_\alpha^{\mathrm{CM}}. \tag{C.4}$$

On the other hand, the quadratic term becomes irrelevant in (C.3) in the limit $\alpha \to \infty$, and we obtain the identities

$$\lim_{\alpha\to\infty} w_\alpha^{\mathrm{MAP}} = 0 = \lim_{\alpha\to\infty} w_\alpha^{\mathrm{CM}}, \tag{C.5}$$

where the latter follows from a symmetry argument in the integral (C.2).

For our numerical example, we take

$$A = \begin{bmatrix} \cos\frac{\pi}{6} & \sin\frac{\pi}{6} \\ -\sin\frac{\pi}{6} & \cos\frac{\pi}{6} \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{3}}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{\sqrt{3}}{2} \end{bmatrix},$$

representing rotation counterclockwise by 30°, and let $m = A[1\ \frac{1}{2}]^T$. We let the $\alpha$ range in an interval and compute both vectors $w_\alpha^{\mathrm{MAP}}$ and $w_\alpha^{\mathrm{CM}}$ numerically. We should observe for large $\alpha$ the zero limit (C.5), and for small $\alpha$, the limit

$$\lim_{\alpha\to\infty} w_\alpha^{\mathrm{MAP}} = A[1\ \frac{1}{2}]^T = \lim_{\alpha\to\infty} w_\alpha^{\mathrm{CM}}.$$

See figure C1 for a plot. Clearly, the MAP estimate exhibits sparsity, while the CM estimate does not.

## References


[1] Bauer F and Lukas M A 2011 Comparing parameter choice methods for regularization of ill-posed problems *Math. Comput. Simul.* **81** 1795–1841

[2] Bioucas-Dias J M 2006 Bayesian wavelet-based image deconvolution: a GEM algorithm exploiting a class of heavy-tailed priors *IEEE Trans. Image Process.* **15** 937–51

[3] Bonesky T, Bredies K, Lorenz D A and Maass P 2007 A generalized conditional gradient method for nonlinear operator equations with sparsity constraints *Inverse Problems* **23** 2041–58

[4] Bonesky T, Dahlke S, Maass P and Raasch T 2010 Adaptive wavelet methods and sparsity reconstruction for inverse heat conduction problems *Adv. Comput. Math.* **33** 385–411

[5] Boyd S and Vandenberghe L 2007 *Convex Optimization* (Cambridge: Cambridge University Press)

[6] Bredies K and Lorenz D 2008 Iterated hard shrinkage for minimization problems with sparsity constraints *SIAM J. Sci. Comput.* **30** 657–83

[7] Bredies K, Lorenz D A and Maass P 2009 A generalized conditional gradient method and its connection to an iterative shrinkage method *Comput. Optim. Appl.* **42** 173

[8] Burger M and Osher S 2004 Convergence rates of convex variational regularization *Inverse Problems* **20** 1411–21

[9] Candès E J, Romberg J and Tao T 2006 Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information *IEEE Trans. Inform. Theory* **52** 489–509

[10] Candès E, Romberg K J and Tao T 2006 Stable signal recovery from incomplete and inaccurate measurements *Commun. Pure Appl. Math.* **LIX** 1207–23

[11] Carlin B P and Louis T A 2000 *Bayes and Empirical Bayesian Methods for Data Analysis* (Boca Raton, FL: Chapman and Hall/CRC Press)

[12] Chambolle A, DeVore R A, Lee N-Y and Lucier B J 1998 Nonlinear wavelet image processing: variational problems, compression, and noise removal through wavelet shrinkage *IEEE Trans. Image Process.* **7** 319–35

[13] Chaux C, Combettes P L, Pesquet J-C and Wajs V R 2007 A variational formulation for frame-based inverse problems *Inverse Problems* **23** 1495–518

[14] Dal Maso G 1993 *An Introduction to* $\Gamma$*-Convergence* (Basel: Birkhäuser)

[15] Daubechies I 2006 *Ten Lectures on Wavelets* 9th edn (Philadelphia, PA: SIAM)

[16] Daubechies I, Defrise M and De Mol C 2004 An iterative thresholding algorithm for linear inverse problems with a sparsity constraint *Commun. Pure Appl. Math.* **57** 1413–57

[17] Daubechies I, Fornasier M and Loris I 2008 Accelerated projected gradient method for linear inverse problems with sparsity constraints *J. Fourier Anal. Appl.* **14** 764–92

[18] Dobson D C and Santosa F 1994 An image enhancement technique for electrical impedance tomography *Inverse Problems* **10** 317–34

[19] Dobson D C and Santosa F 1996 Recovery of blocky images from noisy and blurred data *SIAM J. Appl. Math.* **56** 1181–98

[20] Donoho D L and Johnstone I M 1994 Ideal spatial adaptation via wavelet shrinkage *Biometrika* **81** 425–55
[21] Donoho D L 1995 Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition *Appl. Comput. Harmon. Anal.* **2** 101–26
[22] Donoho D L 2006 Compressed sensing *IEEE Trans. Inform. Theory* **52** 1289–306
[23] Fiacco A V and McCormick G P 1968 *Nonlinear Programming: Sequential Unconstrained Minimization Techniques* (New York: Wiley)
[24] Figueiredo M A T, Nowak R D and Wright S J 2007 Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems *IEEE J. Sel. Top. Signal Process.* **1** 586–97
[25] Fornasier M 2007 Domain decomposition methods for linear inverse problems with sparsity constraints *Inverse Problems* **23** 2505–26
[26] Grasmair M 2009 Well-posedness and convergence rates for sparse regularization with sublinear $l^q$ penalty term *Inverse Problems Imaging* **3** 383–7
[27] Grasmair M, Haltmeier M and Scherzer O 2008 Sparse regularization with $l^q$ penalty term *Inverse Problems* **24** 055020
[28] Grasmair M, Haltmeier M and Scherzer O 2011 Necessary and sufficient conditions for linear convergence of $l^1$-regularization *Commun. Pure Appl. Math.* **64** 161–82
[29] Gelman A B, Carlin J S, Sterm H S and Rubin D B 2004 *Bayesian Data Analysis* (Boca Raton, FL: Chapman and Hall/CRC Press)
[30] Hanson K M 1987 Bayesian and related methods in image reconstruction from incomplete data *Image Recovery: Theory and Applications* (Orlando, FL: Academic) pp 79–125
[31] Bissantz N, Hohage T and Munk A 2004 Consistency and rates of convergence of nonlinear Tikhonov regularization with random noise *Inverse Problems* **20** 1773–89
[32] Jin B, Khan T and Maass P 2011 A reconstruction algorithm for electrical impedance tomography based on sparsity regularization *Int. J. Num. Methods Eng.* DOI: 10.1002/nme.3247
[33] Kaipio J P, Kolehmainen V, Somersalo E and Vauhkonen M 2000 Statistical inversion and Monte Carlo sampling methods in electrical impedance tomography *Inverse Problems* **16** 1487–522
[34] Kaipio J and Somersalo E 2005 *Statistical and Computational Inverse Problems* (*Applied Mathematical Sciences* vol 160) (New York: Springer)
[35] Kalke M, Siltanen S, Vänskä S, Lassas M and Rantala M 2007 Method and arrangement for multiresolutive reconstruction for medical x-ray imaging *US Patent* 7215730
[36] Klann E, Kuhn M, Lorenz D A, Maass P and Thiele H 2007 Shrinkage versus deconvolution *Inverse Problems* **23** 2231–48
[37] Kolehmainen V, Siltanen S, Järvenpää S, Kaipio J P, Koistinen P, Lassas M, Pirttilä J and Somersalo E 2003 Statistical inversion for x-ray tomography with few radiographs: Part II. Application to dental radiology *Phys. Med. Biol.* **48** 1465–90
[38] Kolehmainen V, Vanne S, Siltanen S, Järvenpää S, Kaipio J P, Lassas M and Kalke M 2006 Parallelized Bayesian inversion for three-dimensional dental x-ray imaging *IEEE Trans. Med. Imaging* **25** 218–28
[39] Lassas M and Siltanen S 2004 Can one use total variation prior for edge preserving Bayesian inversion? *Inverse Problems* **20** 1537–64
[40] Lassas M, Saksman E and Siltanen S 2009 Discretization-invariant Bayesian inversion and Besov space priors *Inverse Problems Imaging* **3** 87–122
[41] Lasanen S 2002 Discretizations of generalized random variables with applications to inverse problems *PhD Thesis* University of Oulu, Oulu, Finland
[42] Lehtinen M S 1988 On statistical inversion theory *Theory and Applications of Inverse Problems* ed H Haario (London: Longmans Green)
[43] Lehtinen M, Päivärinta L and Somersalo E 1989 Linear inverse problems for generalized random variables *Inverse Problems* **5** 599–612
[44] Liu J 2004 *Monte Carlo Strategies in Scientific Computing (Springer Series in Statistics)* (New York: Springer)
[45] Lorenz D A 2008 Convergence rates and source conditions for Tikhonov regularization with sparsity constraints *J. Inverse Ill-Posed Problems* **16** 463–78
[46] Lu S and Pereverzev S V 2009 Sparse recovery by the standard Tikhonov method *Numer. Math.* **112** 403–24
[47] Matinverno A and Briggs V A 2004 Expanded uncertainty qualification in inverse problems: hierarchical Bayes and empirical Bayes *Geophysics* **69** 1005–16
[48] Mehrotra S 1992 On the implementation of a primal-dual interior point method *SIAM J. Optim.* **2** 575–601
[49] Meyer Y 2001 *Oscillating Patterns in Image Processing and Nonlinear Evolution Equations* (*University Lecture Series* vol 22) (Providence, RI: American Mathematical Society)
[50] Mosegaard K and Tarantola A 1995 Monte Carlo sampling of solutions to inverse problems *J. Geophys. Res.* B **14** 12431–47

[51] Nicholls G K and Fox C 1998 Prior modelling and posterior sampling in impedance imaging *Bayesian Inference for Inverse Problems* ed A Mohammad-Djafari (*Proc. SPIE* vol 3459) (Bellingham, WA: SPIE Optical Engineering Press) pp 116–27

[52] Niinimäki K, Kolehmainen V and Siltanen S 2007 Bayesian multiresolution method for local tomography in dental x-ray imaging *Phys. Med. Biol.* **52** 6663–78

[53] Nocedal J and Wright S J 2006 *Numerical Optimization* (*Springer Series in Operations Research*) (New York: Springer)

[54] Piiroinen P 2005 Statistical measurements, experiments and applications *Ann. Acad. Sci. Fenn. Mathematia Dissertationes* **143**

[55] Ramlau R 2008 Regularization properties of Tikhonov regularization with sparsity constraints *Electron. Trans. Numer. Anal.* **30** 54–74

[56] Ramlau R and Teschke G 2006 A Tikhonov-based projection iteration for nonlinear ill-posed problems with sparsity constraints *Numer. Math.* **104** 177–203

[57] Rantala M, Vänskä S, Järvenpää S, Kalke M, Lassas M, Moberg J and Siltanen S 2006 Wavelet-based reconstruction for limited-angle x-ray tomography *IEEE Trans. Med. Imaging* **25** 210–7

[58] Rudin L I, Osher S and Fatemi E 1992 Nonlinear total variation based noise removal algorithms *Physica* D **60** 259–68

[59] Scherzer O, Grasmair M, Grossauer H and Haltmeier M 2008 *Variational Methods in Imaging* (*Applied Mathematical Sciences*) (Berlin: Springer)

[60] Schmeisser H-J and Triebel H 1987 *Topics in Fourier Analysis and Function Spaces* (New York: Wiley)

[61] Siltanen S, Kolehmainen V, Järvenpää S, Kaipio J P, Koistinen P, Lassas M, Pirttilä J and Somersalo E 2003 Statistical inversion for x-ray tomography with few radiographs: Part I. General theory *Phys. Med. Biol.* **48** 1437–63

[62] Triebel H 2008 *Function Spaces and Wavelets on Domains* (*European Mathematical Society Tracts in Mathematics* vol 7) (Zurich: European Mathematical Society Publishing House)

[63] Vanderbei R J 2008 *Linear Programming: Foundations and Extensions* (New York: Springer Science and Business Media)

[64] Veraar M C 2012 Regularity of Gaussian white noise on the *d*-dimensional torus *Banach Center Publications 'Józef Marcinkiewicz Centenary Volume'* at press

[65] Vänskä S, Lassas M and Siltanen S 2009 Statistical x-ray tomography using empirical Besov priors *Int. J. Tomography Stat.* **11** 3–32

[66] Wright S J 1997 *Primal-Dual Interior-Point Methods* (Philadelphia, PA: SIAM)

[67] Zarzer C A 2009 On Tikhonov regularization with non-convex sparsity constraints *Inverse Problems* **25** 025006