Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

MATEMATICA

Ciclo XXXII

Settore Concorsuale: 01/A5

Settore Scientifico Disciplinare: MAT/08

SPACE ADAPTIVE AND HIERARCHICAL BAYESIAN VARIATIONAL MODELS FOR IMAGE RESTORATION

Presentata da: Monica Pragliola

Coordinatore Dottorato	**Supervisori**
Fabrizio Caselli	Fiorella Sgallari
	Co - supervisori
	Daniela Calvetti Erkki Somersalo

Esame finale anno 2020

Abstract

The main contribution of this thesis is the proposal of novel space-variant regularization or penalty terms motivated by a strong statistical rational. In light of the connection between the classical variational framework and the Bayesian formulation, we will focus on the design of highly flexible priors characterized by a large number of unknown parameters. The latter will be automatically estimated by setting up a hierarchical modeling framework, i.e. introducing informative or non-informative hyperpriors depending on the information at hand on the parameters.

More specifically, in the first part of the thesis we will focus on the restoration of natural images, by introducing highly parametrized distribution to model the local behavior of the gradients in the image. The resulting regularizers hold the potential to adapt to the local smoothness, directionality and sparsity in the data. The estimation of the unknown parameters will be addressed by means of non-informative hyperpriors, namely uniform distributions over the parameter domain, thus leading to the classical Maximum Likelihood approach.

In the second part of the thesis, we will address the problem of designing suitable penalty terms for the recovery of sparse signals. The space-variance in the proposed penalties, corresponding to a family of informative hyperpriors, namely generalized gamma hyperpriors, will follow directly from the assumption of the independence of the components in the signal. The study of the properties of the resulting energy functionals will thus lead to the introduction of two hybrid algorithms, aimed at combining the strong sparsity promotion characterizing non-convex penalty terms with the desirable guarantees of convex optimization.

Contents

List of symbols

$\mathbb{R}$	The set of real numbers
$\mathbb{R}_+$	The set of non-negative real numbers
$\partial\Omega$	Boundary of the domain Ω
∇u	gradient of u in continuous setting
$C(\Omega)$	Space of continuous functions in Ω
$\mathrm{C}_c^1(\Omega, \mathbb{R}^2)$	Space of continuously differentiable functions with compact support in $\Omega \subset \mathbb{R}^2$
$L_p\,(\Omega)$	Space of functions $f : \Omega \to \mathbb{R}$ such that $\int_\Omega \|f\|^p dx < \infty$, with $1 \leq p < \infty$
$L_\infty(\Omega)$	Space of essentially bounded measurable functions on Ω
$W^{k,p}(\Omega)$	Sobolev space of functions $u \in L_p(\Omega)$ with weak derivatives up to order k also in $L_p(\Omega)$
$\mathrm{BV}(\Omega)$	Space of functions of bounded variation on Ω
$\\|\cdot\\|_p$	ℓ_p-norm
I_n	Identity matrix $n \times n$
$\mathrm{null}(\mathrm{L})$	Null space of matrix L
$\mathbf{0}_n$	n-dimensional null-vector
$\mathrm{span}\{v_1, \ldots, v_n\}$	Linear span of the set of vectors $v_1, \ldots, v_n$
$\mathbb{P}(F)$	Probability measure of the event F
$\mathrm{P}_X(x)$	Probability density function of the random variable X

Part I

Introduction

Chapter 1

Preamble

The main goal of this thesis is the design of novel space-variant regularization terms to be used in the framework of variational methods for signal restoration. As one would expect, the flexibility of the proposed regularizers is mainly due to the presence, in their definition, of a number of free parameters at least equal to the number of pixels. The estimation of the values of these parameters is clearly crucial to make the aforementioned regularizers applicable in practice. Hence, as a further contribution, we propose robust, efficient and, more importantly, automatic parameter estimation strategies.

The probabilistic regularizers, as well as the parameter estimation procedures proposed, are introduced in a *Bayesian* perspective, where the unknown quantities involved in the image restoration problem, i.e. the unknown signal and all the unknown parameters, will be modeled as random variables. The proposed regularization terms will be thus related to the probability density functions (pdfs), encoding information or assumptions on the unknown signal that are believed reasonable a priori, whence commonly called *prior* pdfs. Analogously, by adopting a *hierarchical* modeling, a further layer of assumptions about characteristics of the unknown solution will be introduced with the definition of *hyperpriors*, i.e. the pdfs of the unknown parameters of the prior. The prior pdf and the hyperprior pdf will be coupled with a suitable likelihood pdf, encoding information on the degradation model, in particular on the corrupting noise, in order to obtain the analytical expression of the *posterior* pdf. The posterior pdf jointly models the behavior of the signal and of the parameters once that the information available has been exploited in light of the a priori beliefs. A classical Maximum A Posteriori (MAP) approach is then employed in order to approximate the posterior pdf with a single-point estimate.

This thesis is the result of the collaboration with two research groups, one affiliated to the University of Bologna, where I joined the PhD program in Mathematics, and the other based in Cleveland, Ohio, at Case Western Reserve University, where I have been twice as a visiting PhD student. The dual approach that has characterized my research in the last three years is reflected in the structure of the thesis. In fact, the proposed results are here presented divided into two parts, namely Part II and Part III of this work. In fact, although the researches carried out in both collaborations share a space-variant perspective and a statistical basis, they also present significant differences that should be stated in advance. In terms of content, the focus in Part II is on how the total variation (TV) prior can be modified, based on statistical assumptions, so as to integrate in the regularization local information about the gradient structure of the processed image. The main effort will thus consists in representing the images in terms of suitable, space-variant, image regularizers. In Part III, on the other hand, we will consider signals admitting a sparse representation in a given basis. Here, the space-variance comes directly from assuming the components of the unknown signal to be independently distributed as zero-mean Gaussian random variables with different variances.

There are several dissimilarities arising in the discussion of these two topics, both from a modeling and a conceptual point of view. The first one concerns the choice of the hyperpriors modeling the behavior of the unknown parameters. In fact, whilst it can be easy to get an intuition on the properties of the signal that we want to restore and design a prior accordingly, making an assumption on the parameters can be challenging or even unfeasible. For this reason, we will make a distinction between non-informative and informative hyperpriors; the former, treated in Part II, are uniform pdfs over the parameters domain, while the latter encode more substantial information and will be matter of discussion of Part III. In other words, in Part II the hierarchical modeling coupled with non-informative hyperpriors is along the lines of the classical Maximum Likelihood approach for the estimation of the parameters, that can be also interpreted as an *a posteriori* estimation procedure. *A priori* information available on the unknown space-variant parameters will be exploited only in Part III.

A further difference lies in the way one generally thinks at variational methods and at their relation with the Bayesian framework. In Part II, variational models are the main actors and the Bayesian framework will only provide a stronger rational basis behind their introduction. In other words, the variational models proposed could be defined independently from any probabilistic interpretation. In Part III, on the other hand, the variational models we will end up with are inextricably related to the Bayesian framework, in particular to the choice of hyperpriors. Clearly, such dissimilarity is not formal, since our focus will be the minimization of a functional in both cases, but rather conceptual.

In addition, using the terminology from [67], in Part II we embrace the *analysis* approach, that is we derive the pdf of the unknown signal by applying linear filters to the signal itself. More in detail, we will make assumption on the behavior of the discrete gradient of the image. In Part III, we will instead recover a vector representing the unknown signal in a given basis. This framework goes under the name of *synthesis* approach. In the synthesis perspective, one usually looks for a sparse representation of the signal. In fact, the focus of Part III will be on sparse recovery.

1.1 Organization of the thesis

In Chapter 2, we will set the notations used throughout the thesis. We will introduce the degradation model related to the image restoration problem. In particular, we will pay attention to the ill-posedeness arising for such problem. Then, in Chapter 3, variational methods for image restoration will be introduced, together with their interpretation within a Bayesian framework. More in detail, we will highlight the connections between the regularization term of the variational functional and the prior pdf set on the unknown image. Moreover, we will derive some of the classical variational models, such as Tikhonov-L$_2$ and TV-L$_2$ relying on the Bayesian interpretation. A detailed analysis of the drawbacks arising when adopting a TV regularizer will be carried out in Chapter 4, together with a revision of the space-variant and directional regularization terms already introduced in literature. As far as the problem of sparse recovery is concerned, we will highlight some issues related to the convex and non-convex penalty terms proposed so far. After the *chapeau* aimed at introducing the problems and motivating the results discussed in the following chapters, in Chapter 5 a weighted TV regularizer [19] is introduced. A generalization [110], obtained by adding a space-variant parameter tuning the type of regularization, is then considered in Chapter 6. The last contribution of the first part is proposed in Chapter 7, where the information on local orientations in the image are added to the previously introduced regularizers [20]. Sparse recovery problem is discussed in the last three chapters of the thesis. In particular, in Chapter 8, we consider the problem of recovering a sparse signal on which a conditionally Gaussian prior with a generalized gamma hyperprior is set [27]. In particular, we extend to more general settings a previous work [32], where the gamma hyperprior has been considered. In Chapter 9, starting from the discussion carried out in Chapter 8, we will thus propose two hybrid algorithmic schemes aimed at enforcing sparsity while trying to preserve a convex regime. Finally, in Chapter 10, we consider an application of the outlined framework to the recovery of a signal admitting a sparse representation in an over-complete basis. In the computed examples, the space-variance perspective will arise again and interpreted as using different bases to more efficiently represent different local features.

1.2 Related publications

A large part of the discussion reported in the thesis refers to upcoming works or works that have already been published. In particular, the three chapters presented in the first part of the thesis, namely Chapter 5 to 7 are based, respectively, on the following works:

- Calatroni, L., Lanza, A., Pragliola, M., Sgallari, F. Adaptive parameter selection for weighted-TV image reconstruction problems. To appear on Journal of Physiscs: Conference series.

- Lanza, A., Morigi, S., Pragliola, M., Sgallari, F. Space-variant generalised Gaussian regularisation for image restoration. Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization, 7:490-503, 2019.

- Calatroni, L., Lanza, A., Pragliola, M., Sgallari, F. A Flexible Space-Variant Anisotropic Regularization for Image Restoration with Automated Parameter Selection. SIAM Journal on Imaging Sciences, 12:1001-1037, 2019.

As far as the second part of the thesis is concerned, Chapter 8 refers to the following work:

- Calvetti, D., Pragliola, M., Somersalo, E., Strang, A. Sparse reconstructions from few noisy data: analysis of hierarchical Bayesian models with generalized gamma hyperpriors. Inverse Problems, 36, 2020.

Moreover, Chapter 9 and Chapter 10 will be further developed in future works.

Chapter 2

A world of images (to be processed)

In the last few decades, images have become one of the most widespread way of communication. The main reason behind this increasing diffusion is the variety of different applications involving data that can be rearranged so as to be visualized as images, making visible what is not due to its inaccessibility or because of the limits of human eyes. Typical examples are represented by seismic imaging and astronomical imaging, focused on collecting information about subsoil and celestial structures, respectively. Images can also be recorded from sources of radiations, as in the case of PET (positron emission tomography) and MRI (magnetic resonance imaging) in a medical scenario. From industrial applications to life sciences, we simply can not avoid dealing with images.

The aim of *image processing* is to develop strategies for manipulating, efficiently transmitting and even improving 2-dimensional (2D) signals, whose formation, in general, is carried out by recasting indirect information, possibly coming from non-imaging modalities, meaning that the measurement space differs from the image space. We can thus classify the generic imaging problem as an *inverse problem*.

2.1 Inverse problems in imaging

An inverse problem arises every time that an unknown cause producing an observed data - or effect - is investigated, provided that the cause-to-effect forward model is available. Moving against causality implies taking some risks, that here mainly consist in the loss of well-posedness. According to the definition given by J. Hadamard in the early 20th century, a problem is said to be *well-posed* if:

(i) the problem admits a solution;

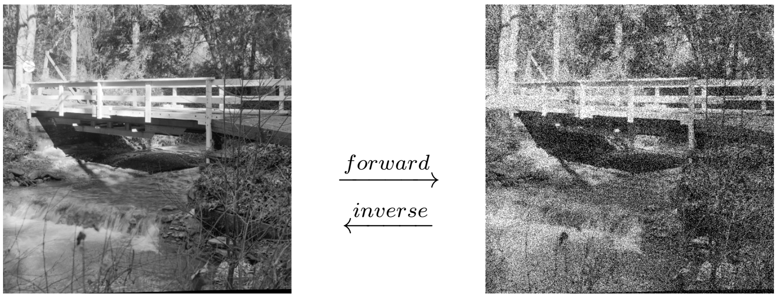
(ii) the solution is unique;

(iii) the solution depends continuously on the data.

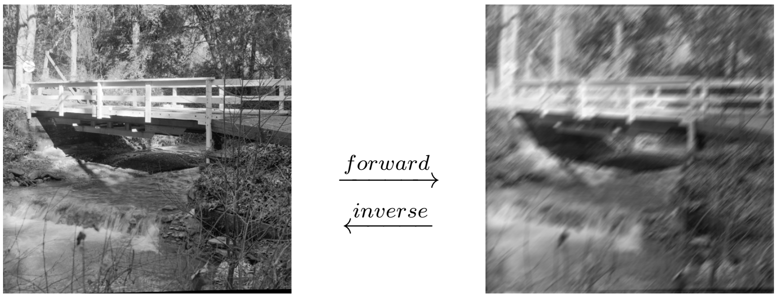
Inverse problems may turn out to be *ill-posed*, that means that at least one of the properties mentioned above is not satisfied. In general, the existence and the uniqueness of the solution are easier to tackle. If the existence of the solution is not guaranteed, a possible strategy is relaxing the regularity properties that we ask the candidate solution to satisfy. The non-uniqueness can be overcome by imposing a further condition on the solution, such as being the solution with minimum norm. Therefore, the most challenging issue is the lack of continuous dependence on the data, meaning that small errors in the data can lead to very large perturbations in the final solution.

Before exploring the reason behind the occurrence of ill-posedness for imaging problems and possible strategies to overcome this issue, we are giving a brief review of widely studied problems in the field of image processing:

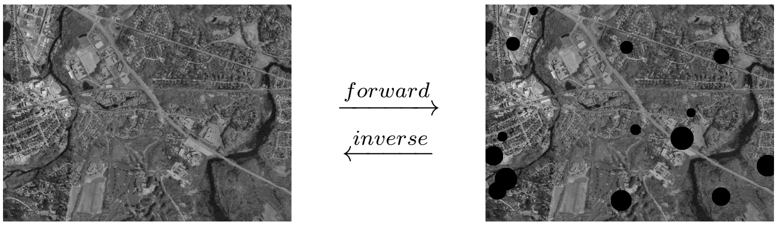
- *Denoising*, is the problem of removing noise from the image keeping the sensible information unchanged. The presence of noise can be related to the quantum nature of electromagnetic radiation or to atmospheric distortions.



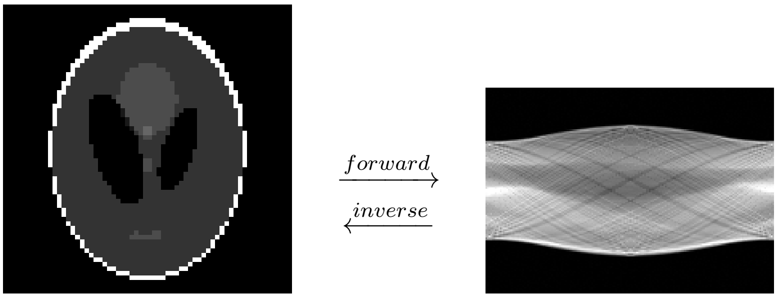
- *Deblurring*, is the enhancement of images corrupted by blur. The blur, as well as the noise, can be an unavoidable consequence of the tool used for capturing the data - that is the case of microscopic images - or can be due to the relative motion between the object of interest and the camera, or still can be introduced by the device being out of focus.



- *Inpainting*, is the task of filling parts of the image missing because of occlusions or other damages.



- *Reconstruction*, aimed at retrieving an image starting from physical data that do not belong to the image space. This is a typical problem arising in medical imaging.



Mathematically, a continuous signal u can be modeled as a function in $L^1(\Omega)$

$$u : \Omega \subset \mathbb{R}^d \to \mathbb{R}^s \,, \tag{2.1}$$

where Ω is a compact set in $\mathbb{R}^d$, with $d \geq 1$ and $s \geq 1$. When $d = 1$ and $s = 1$, u is a one-dimensional signal, while $d = 2, 3$ for 2D and 3D images, respectively. In particular, when $s = 1$, u is a gray scale image, or single-channel. For $s \geq 2$, u is a color image, or multi-channel. As an example, when $s = 3$, u assumes values on three color channels, as in the case of RGB (red-green-blue) images. The continuous formulation provides a straightforward modeling of the human acquisition process, based on the coupled action of eyes and brain. Moreover, some image structures find their natural interpretation in continuous settings, as in the case of edges that can be described as jump discontinuities in the image function.

A substantial literature has been devoted to the solution of imaging inverse problems of the form

$$\text{find } u \text{ such that} \quad b = \mathcal{T}(u) = \mathcal{N}\left(\mathcal{A}\left(u\right)\right) \,, \tag{2.2}$$

where b is the observed image defined on Ω and $\mathcal{T} : \mathbb{R}^s \to \mathbb{R}^s$ is a model of the measurement process, which is typically the combination of a deterministic mapping $\mathcal{A}$ acting on u and a random noise operator, modeled by operator $\mathcal{N}$. We can also refer to the linear model in (2.2) as the *degradation model*.

All the problems listed above can be described by a linear degradation model. For image denoising problems $\mathcal{A}$ is the identity operator, i.e. $\mathcal{A}(u) = u$, while when dealing with image deblurring, the transformation is of the form

$$\mathcal{A}(u) = \int_\Omega k(x,y)u(y)dy\,,$$

with k denoting the blur kernel. In the case of image inpainting, $\mathcal{A}(u) = \chi_C\, u$, with χ_C the indicator function of the subset C of the image domain Ω in which the information is available, and, finally, in the image reconstruction framework, $\mathcal{A}$ is, e.g, the Radon or Fourier transformation function - just to name a few - depending on the applications.

2.2 The inverse problem of interest: image restoration

In this thesis, we focus on the so-called *image restoration* problem, namely the problem of recovering images corrupted by both blur and noise. In particular, we will consider gray-level images. Thus, the continuous degradation model of reference is

$$b(x) = \mathcal{N}\Big(\int_\Omega k(x,y)u(y)dy\Big)\,, \quad \forall x \in \Omega\,, \quad \text{with} \quad b : \Omega \to \mathbb{R}\,, \quad b \in L^1(\Omega)\,. \tag{2.3}$$

In the case of additive noise, which is the one considered in this thesis, the model in (2.3) takes the specific form

$$b(x) = \int_\Omega k(x,y)u(y)dy + e(x)\,, \quad \forall x \in \Omega\,. \tag{2.4}$$

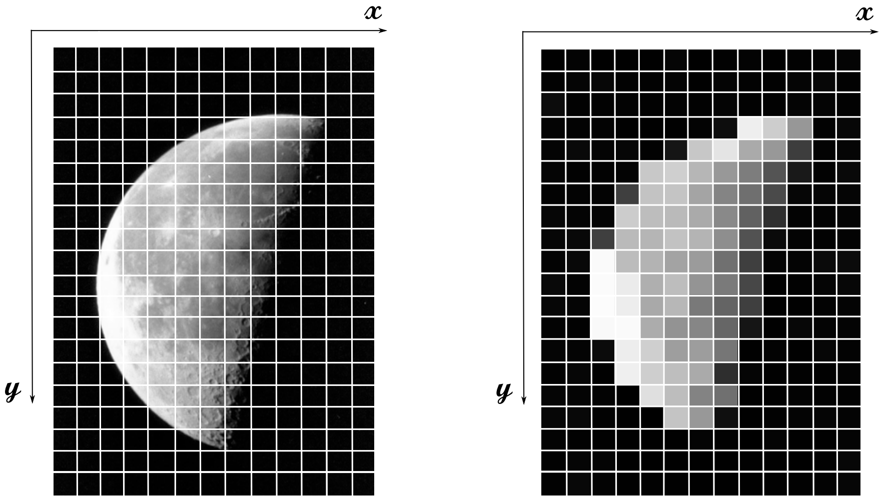
We remark that, at least in principle, the work presented here could be extended to the multi-channel case and, in a non trivial way, to other kinds of inverse problems related to images, as the ones mentioned above.

Discretization It is worthwhile to spend now a few words about how to make a continuous image *computer-readable*, that is how the *digitalization* process is carried out [13]. The continuous gray-scale image u is converted into a discrete image u_d which is a matrix whose elements are referred to as *picture elements*, or *pixels*. The action of transforming u in u_d is also known as *sampling*. Sampling may or may not lead to a loss of sensible information; this is, of course, due to the resolution of the device used for the acquisition. We can interpret this operation as approximating the continuous domain Ω, introduced in (2.1), with a discrete grid Ω_d. In Figure 2.1, a continuous image u and its sampled version u_d, corresponding to a very rough discretization grid, are shown.

Despite the pleasant properties of a continuous formulation highlighted above, here we are rather adopting a discrete framework. The discrete version of the continuous degradation model in (2.4) related to the image restoration problem thus reads as

$$b_d = \mathrm{K}u_d + e_d, \tag{2.5}$$

where u_d, b_d and e_d are the discretizations, in vectorized form, of the observed continuous image b, of the continuous unknown u and of the additive corrupting noise e, respectively. In particular, u_d, b_d, e_d $\in \mathbb{R}^n$ and $n = m \times \ell$, in case u_d being originally discretized over an



(a) (b)

Figure 2.1: Original image $u : \Omega \to \mathbb{R}$ with a superimposed uniform sampling grid (a), sampled discrete image (b).

$m \times \ell$ grid. The matrix K models the action of the blur on the unknown image. More details about e_d and K will be given in the next sections. Nonetheless, in the following, in order to avoid heavy notations, we are denoting both the continuous and the discrete versions of the quantities involved in (2.5) without subscriptions, with the caveat of specifying whether we are in continuous or discrete settings.

In conclusion, the discrete ill-posed inverse problem of interest in this thesis reads:

$$\text{find } u \in \mathbb{R}^n \text{ such that } \quad b = \mathrm{K}u + e\,. \tag{2.6}$$

2.2.1 Noise

In this section, we review some of the noise models that are commonly encountered in the applications.

Additive noise. Consider the linear problem modeling the corruption of u by means of an additive noise:

$$b = u + e\,,$$

where u, b and e are vectorized images, that is $u, b, e \in \mathbb{R}^n$ and $n = m\ell$, in case u being originally discretized over an $m \times \ell$ grid. In other words, the elements of the original u are re-arranged so as to be ordered columnwise. The entries of the noise vector e can be interpreted as realizations drawn from a fixed random variable.

A noise arising in many applications is the *Additive White Gaussian Noise* (AWGN)

$$e \sim \mathcal{N}(0, \Sigma)\,, \quad \text{with} \quad \Sigma = \sigma^2 \mathrm{I}_n\,.$$

The property of *whiteness* implies that the entries of e are identically independently distributed (i.i.d.) realizations of the given distribution - in the case of AWGN, for each $i = 1, \ldots, n$, e_i is drawn from a zero-mean Gaussian distribution with standard deviation σ. This reflects in

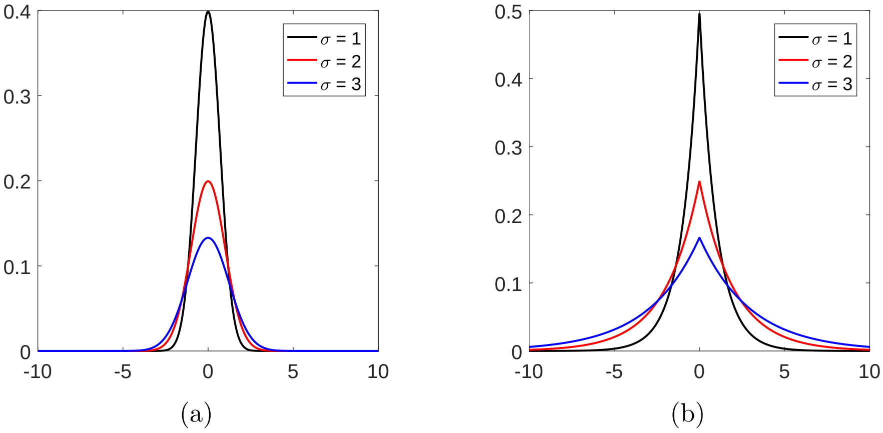


Figure 2.2: Gaussian (a) and Laplacian (b) pdf for different values of σ.

the covariance matrix Σ being a scaled identity. The independence of the entries also allows to factorize the probability density function of e as follows,

$$\mathrm{P}(e) = \mathrm{P}(e_1)\,\mathrm{P}(e_2)\cdots\mathrm{P}(e_n)\,,$$

thus giving

$$\mathrm{P}(e) = \prod_{i=1}^{n}\mathrm{P}(e_i) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{e_i^2}{2\sigma^2}\right) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{\|e\|_2^2}{2\sigma^2}\right).$$

One of the reason behind the wide use of the AWGN is the fact that, in general, the final noise corrupting the acquired data is the result of the contribution of different sources, and, according to the central limit theorem, it can be approximated with a Gaussian distribution. However, we remark that usually the number of sources is not large enough to justify the application of the central limit theorem. Therefore, often the adoption of AWGN is improper, but in most cases it produces a reasonable and satisfying approximation of the observed noise.
As a further example of additive noise, we mention the *Additive White Laplacian noise* (AWLN)

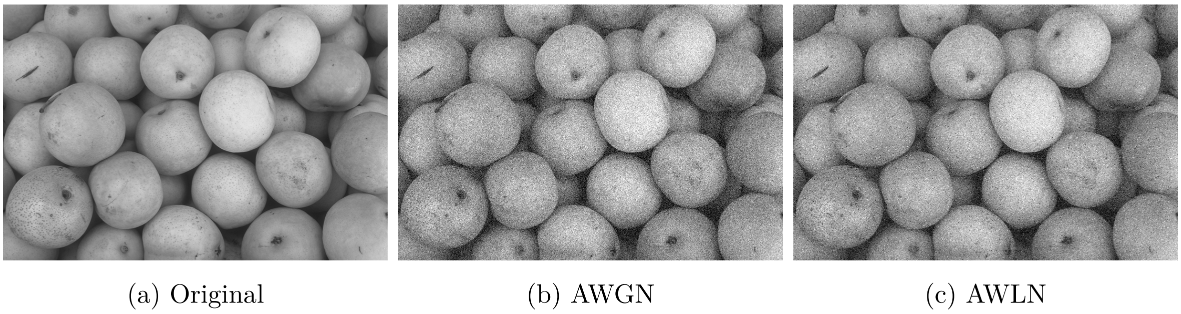
$$e \sim \mathrm{Laplace}(0,\Sigma)\,, \quad \text{with} \quad \Sigma = \sigma^2\mathrm{I}_n\,,$$

with each entry e_i having variance equal to $2\sigma^2$. The pdf of e is defined as

$$\mathrm{P}(e) = \prod_{i=1}^{n}\mathrm{P}(e_i) = \prod_{i=1}^{n}\frac{1}{2\sigma}\exp\left(-\frac{|e_i|}{\sigma}\right) = \frac{1}{(2\sigma)^n}\exp\left(-\frac{\|e\|_1}{\sigma}\right).$$

The AWLN results to be more impulsive than AWGN. This is a consequence of the heavy tails of the Laplace pdf, which allow the occurrence of realizations that can be very far away from the mean value with non-negligible probability. The shape of the Gaussian and Laplace pdf for different values of σ are shown in Figure 2.2a-2.2b, respectively.
In Figure 2.3, an image corrupted by both AWGN and AWLN is shown.

Signal-dependent noise. Beside the additive case, noise observed in medical imaging problems is often signal-dependent, in the sense that the pdf, or, more specifically, the standard



(a) Original (b) AWGN (c) AWLN

Figure 2.3: *Additive noise.* Original image (a) corrupted by AWGN (b) and AWLN (c) with $\sigma = 0.1$.

deviation σ of the noise, depends on the intensity of the underlying noiseless signal. Here, we only cite the *Poisson noise*, which is related to the inherent quantum nature of light, as in the case of computer tomography (CT). At each pixel i, the observed intensity b_i is the realization of a Poisson random variable with mean value μ_i:

$$\mathrm{P}(b_i) = \frac{\mu_i^{b_i}\,\mathrm{e}^{-\mu_i}}{b_i!}\,, \quad i = 1, \dots, n.$$

The mean value μ_i, and, as a consequence, the noise standard deviation $\sigma_i = \sqrt{\mu_i}$, increases with the number of photons hitting the sensors. When the number of photons is sufficiently large, the Poisson distribution can be approximated by a Gaussian distribution, but the noise standard deviation depends on the number of photons. An image corrupted by Poisson noise is shown in Figure 2.4b.

Impulse noise. Another common class of noise is the one of impulse noises, including, among the others, the *Salt-and-Pepper noise* (SPN). Typically, SPN arises when the device used to measure the data presents some malfunctioning pixels, or when an error occurs in the transmission of the data. According to [44], the degradation via salt and pepper noise can be modeled as follows,

$$b = (1-s)u + sc\,,$$

with $s, c \in \{0,1\}^n$ the realization of two independent binary random fields such that, for any $i = 1, \dots, n$,

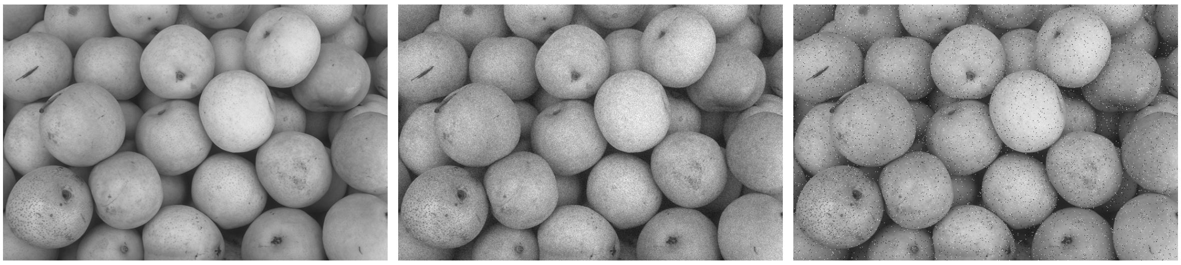
$$c_i = \begin{cases} 0 & \text{with probability } \frac{1}{2} \\ 1 & \text{with probability } \frac{1}{2} \end{cases}, \quad s_i = \begin{cases} 0 & \text{with probability } p \\ 1 & \text{with probability } 1-p \end{cases},$$

where $p \in [0,1]$ controls the noise level, i.e. the number of corrupted pixels.
In Figure 2.4c, an image corrupted by SPN is shown.

2.2.2 Blur

Consider the following formulation of the continuous degradation model

$$\forall x \in \Omega\,, \quad b(x) = b_0(x) + e(x)\,, \quad \text{where} \quad b_0(x) = \int_\Omega k(x,y)u(y)dy \tag{2.7}$$



(a) Original (b) Poisson (c) SPN

Figure 2.4: *Signal-dependent and impulse noise.* Original image (a) corrupted by Poisson (b) and Salt & Pepper noises (c).

and where b is a noisy data corrupted by additive noise e. We refer to k as the *blur kernel* or, as it is often called in microscopy imaging, *point spread function* (PSF). The latter definition is motivated from the fact that the action of k on an image consisting of a point source would lead to a spread of the intensity around the point.

The PSF k takes only real and non-negative values

$$k(x,y) \geq 0\,, \quad \forall x, y \in \Omega\,,$$

and satisfies,

$$\int_\Omega k(x,y)\,dy \; = 1\,, \quad \forall x \in \Omega\,.$$

In particular, k usually has a compact support, or is at least absolutely integrable - see e.g. [68, 13]. For simplicity, we focus on *space-invariant* blur, i.e. blur whose action is the same in every point of the image and depends only on the difference between the two pixel locations. In formula,

$$k(x,y) = k(x-y)\,.$$

In this case, the integral in (2.7) takes the form of a convolution product, that is,

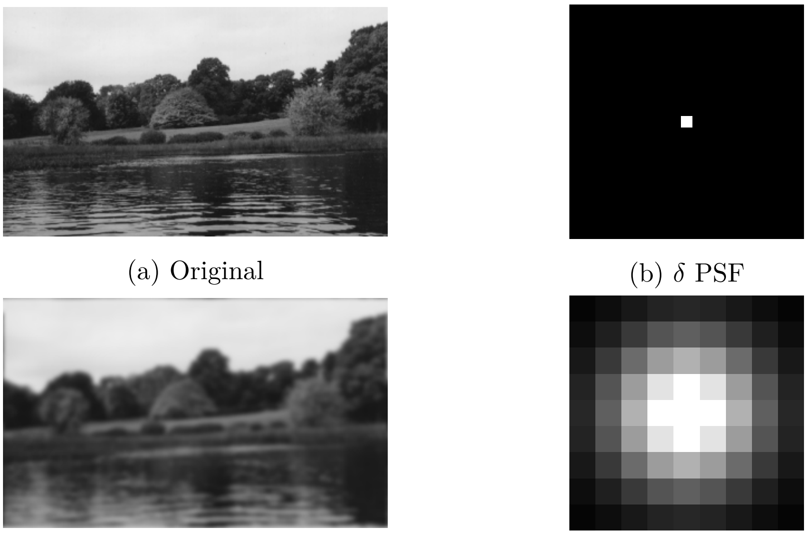
$$b_0(x) = (k * u)(x).$$

The hypothesis of space-invariance can be easily relaxed, although it yields to computational advantages, as it will be detailed later. When the PSF k is unknown, we refer to the problem of solving the integral equation (2.7) as *blind deconvolution* problem. In our analysis, k is assumed to be known.

As an example of space-invariant blur, we consider Gaussian blur

$$k(x-y) = C \exp\left(- \frac{\|x-y\|_2^2}{2w^2} \right),$$

where C is a normalizing constant and w is clearly related to the *width* or spread of the blur. In Figure 2.5c, we show an image corrupted by a Gaussian blur of `band` = 9 and `width` = 2, the latter parameters being the input arguments of the Matlab function `fspecial`. The `band` parameter represents the side length (in pixels) of the square support of the kernel, whereas `width` is the standard deviation (in pixels) of the isotropic bivariate Gaussian distribution defining the kernel in continuous setting. The Gaussian PSF is shown in Figure 2.5d. Notice that one can think at the original image u as the convolution product of u itself by a very narrow PSF whose limit can be modeled as a δ PSF, with δ denoting the Dirac function - see Figure 2.5a-2.5b.



(a) Original

(b) δ PSF

(c) Corrupted by Gaussian blur

(d) Gaussian PSF

Figure 2.5: Original image (a) corresponding to the convolution of u with a δ PSF (b), test image corrupted by Gaussian blur (c), generated by Gaussian PSF of `band` $= 9$ and `width` $= 2$ (d).

2.3 Ill-posedness

A possible strategy to solve problem (2.7) in case k being space-invariant and under periodic boundary conditions consists in applying the *convolution theorem* - see e.g. [68], by which

$$\mathcal{F}\,(s * t) = \mathcal{F}(s) \cdot \mathcal{F}(t)\,, \quad \forall s, t \in L^1(\mathbb{R}^2),$$

where

$$\mathcal{F}f(\xi,\psi) = \int_{\mathbb{R}^2} e^{-\xi x - \psi y} f(x,y)dxdy\,,$$

is the Fourier transform of the generic $f \in L^1(\mathbb{R}^2)$, and (ξ,ψ) is the pair of variables in the frequency domain. We thus have:

$$\mathcal{F}\,b_0 = \mathcal{F}\,(k * u) = \mathcal{F}\,k \cdot \mathcal{F}\,u\,,$$

for all the frequencies $(\xi,\psi) \in \mathbb{R}^2$. Hence,

$$\mathcal{F}\,u = \frac{\mathcal{F}\,b_0}{\mathcal{F}\,k}\,,$$

where the division is intended point-wise, and u can be obtained by applying the inverse Fourier transform to $\mathcal{F}\,u$. Nevertheless, we recall that k has a compact support and, for the Riemann-Lebesgue Lemma, its Fourier transform goes to zero as the norm of the variables in the frequence space (ξ, ψ) increases. Nonetheless, usually b_0 is not available and one has at disposal only a noisy data b. As a consequence, a little perturbation in the data is amplified in the final solution. This phenomenon determines the ill-posedness of the image restoration problem - because of the lack of continuous dependence on the data - and makes necessary the design of different methods to recover u.

By means of standard quadrature formulas, the integral in (2.7) can be discretized thus leading to the discrete degradation model in (2.6). In particular, will refer to $\mathrm{K} \in \mathbb{R}^{r \times n}$ as the *blur matrix*. Typically, K is a square matrix - when $r = n$ - or a wide matrix - when $r < n$, that is the case of an *under-determined* system.

The blur matrix K is typically a large and sparse matrix. Based on linear algebra considerations, the previous discussion can be interpreted in terms of the spectral properties of K, that can be decomposed via singular value decomposition (SVD) as

$$\mathrm{K} = \mathrm{W}\,\Lambda\,\mathrm{V}^T\,, \mathrm{W} \in \mathbb{R}^{r\times r}\,,\ \Lambda \in \mathbb{R}^{r\times n}\,,\ \mathrm{V} \in \mathbb{R}^{n\times n}\,,$$

where W and V are orthonormal matrices, and Λ is a diagonal matrix in case K being square, and as close as possible to a diagonal matrix in case K being wide. In formulas,

$$\Lambda = \begin{pmatrix} \lambda_1 & \dots & 0 \\ & \ddots & \\ 0 & \dots & \lambda_n \end{pmatrix} \text{ if } r = n \quad \text{and} \quad \Lambda = \begin{pmatrix} \lambda_1 & \dots & 0 & 0 & \dots & 0 \\ & \ddots & 0 & 0 & \dots & 0 \\ 0 & \dots & \lambda_r & 0 & \dots & 0 \end{pmatrix} \text{ if } r < n\,,$$

with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{\min\{n,r\}} \geq 0$ denoting the *singular values* of K. We can thus define the condition number κ as the ratio between the largest and the smallest singular value, i.e. $\kappa := \lambda_1/\lambda_n$. The magnitude of κ is usually taken as an indicator of the ill-conditioning of the matrix: the larger κ, the more likely K will be singular. When K is square and non-singular, we can define K^{-1}, whose SVD is inherited from the SVD of K. Namely,

$$\mathrm{K}^{-1} = \mathrm{V}\,\Lambda^{-1}\,\mathrm{W}^T. \tag{2.8}$$

More in general, when K is not invertible, we can still consider the Moore-Penrose or pseudo-inverse matrix $\mathrm{K}^\dagger$ and introduce its SVD. Starting from (2.8), one may thus think to solve the linear system

$$\mathrm{K}u = b_0\,,$$

by simply inverting K, i.e.

$$u = \mathrm{K}^{-1}\,b_0 = \mathrm{V}\,\Lambda^{-1}\,\mathrm{W}^T\,b_0 = \sum_{i=1}^{n} \frac{w_i^T b_0}{\lambda_i} v_i\,,$$

with w_i, v_i denoting the i-th column of W and V, respectively. Typically, the blur matrix K presents very small singular values. This property can be interpreted as the discrete counterpart of the absolute integrability of the blur kernel k arising in continuous settings. Clearly, small singular values can lead to the amplification of small perturbations in the data. We thus expect the condition number κ to be very large and K to be very ill-conditioned. As a consequence, its inversion is not feasible and different strategies must be proposed to address the solution of the far from harmless linear system in (2.6).

Beyond the spectral properties of K, the structure of the blur matrix is determined by the property of the continuous kernel k and by the fixed boundary conditions, the latter expressing our assumption on the scene outside the acquired image. In fact, formally the blur acts on the image via a convolution product. This means that on the boundary of the acquisition domain the PSF overlaps the image border, "falling out" of it. As a consequence, what is *outside* the image has an influence on the action of the blur *inside* the image. In Figure 2.6a, the blue line represents the boundary of the acquisition domain. The compact support and space-invariant PSF acts on the top of the boundary, partially outside the image domain. It is clear that, in order to compute the convolution product, the image outside the blue box, which a priori is unknown, must be fixed somehow.

Before going on with the description of popular choices of boundary conditions, we recall some basic linear algebra definitions.

- The matrix A is said to be a *Toeplitz* matrix if the entries of A are constant on each diagonal. In formula,
$$\mathrm{A} = \begin{pmatrix} a_0 & a_1 & a_2 & a_3 & a_4 \\ a_{-1} & a_0 & a_1 & a_2 & a_3 \\ a_{-2} & a_{-1} & a_0 & a_1 & a_2 \\ a_{-3} & a_{-2} & a_{-1} & a_0 & a_1 \\ a_{-4} & a_{-3} & a_{-2} & a_{-1} & a_0 \end{pmatrix}$$

- The matrix A is said to be a *circulant* matrix if it is a Toeplitz matrix in which each row (and column) is a periodic shift of its previous row (and column). Namely,
$$\mathrm{A} = \begin{pmatrix} a_0 & a_1 & a_2 & a_3 & a_4 \\ a_4 & a_0 & a_1 & a_2 & a_3 \\ a_3 & a_4 & a_0 & a_1 & a_2 \\ a_2 & a_3 & a_4 & a_0 & a_1 \\ a_1 & a_2 & a_3 & a_4 & a_0 \end{pmatrix}$$

- The matrix A is said to be a *Hankel* matrix if its entries are constant on each anti-diagonal, that is:
$$\mathrm{A} = \begin{pmatrix} a_{11} & a_1 & a_5 & a_8 & a_{10} \\ a_1 & a_{12} & a_2 & a_6 & a_9 \\ a_5 & a_2 & a_{13} & a_3 & a_7 \\ a_8 & a_6 & a_3 & a_{14} & a_4 \\ a_{10} & a_9 & a_7 & a_4 & a_{15} \end{pmatrix}$$

One of the simplest way of imposing boundary conditions is to assume the original image to be 0 outside the image domain Ω. We refer to this choice as adopting *zero* or *homogeneous Dirichlet* boundary conditions. Under the adoption of zero boundary conditions, the matrix K is a block Toeplitz matrix with Toeplitz blocks. In general, Dirichlet boundary conditions may result very artificial and lead to black artifacts at the boundary, unless the original image is

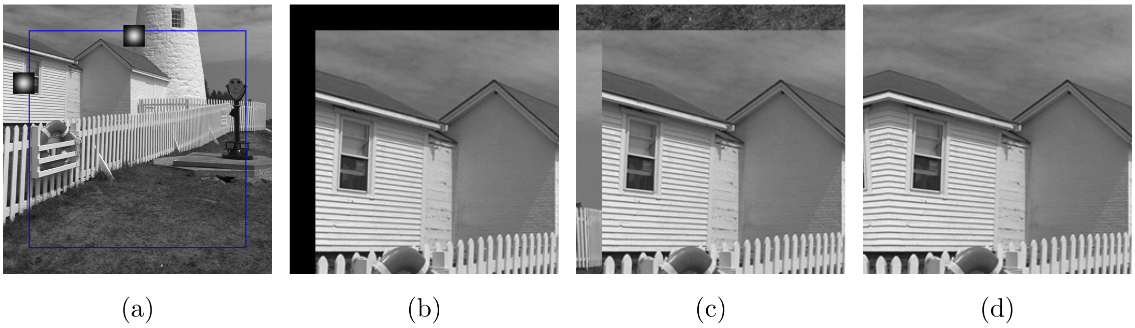

(a) (b) (c) (d)

Figure 2.6: Original image with acquisition domain delimited by a blue line and PSF acting on the boundary (a); zero-Dirichlet (b), periodic (c) and reflective boundary conditions (d).

naturally embedded in a black background.

Alternatively, one can think at u as infinitely repeating itself and *periodic* boundary conditions can be set. The blur matrix K corresponding to periodic boundary conditions is a block circulant matrix with circulant blocks. In this case, K can be diagonalized as follows

$$\mathrm{K} = \mathrm{F}\mathrm{H}\mathrm{F}^* \,, \tag{2.9}$$

where H is a diagonal matrix, F is the 2D Fourier transform matrix and F^* is its adjoint. In the following, it will be made clear how, on the algorithmic side, computations can take advantage of property (2.9). Another choice consists in requiring the image u to have zero normal derivative at the boundary, that is the case of *Neumann homogeneous* or *reflective* boundary conditions, leading to a blur matrix K which is a Toeplitz-plus-Hankel matrix with Toeplitz-plus-Hankel blocks and satisfying a property similar to the one in (2.9). Namely,

$$\mathrm{K} = \mathrm{C}\mathrm{H}\mathrm{C}^T \,,$$

where H, as before, is a diagonal matrix, and C is the 2D cosine transform matrix.

In Figures 2.6b-2.6d, we show how the north-west corner of the image in Figure 2.6a appears under the adoption of the mentioned boundary conditions. Besides the revised cases, we also cite other possible choices of boundary conditions more recently introduced, as the anti-reflective boundary conditions [38] and the synthetic boundary conditions [69]. We remark that the choice of suitable boundary conditions is far from being a negligible issue, since it may have a drastic influence on the quality of the restored image.

Chapter 3

The art of regularizing

In Chapter 2, the compactness of the blur kernel k, in continuous settings, and the ill-conditioning of the blur matrix K, in discrete settings, are two sides of the same coin explaining the instability of the image restoration problem. As a paradigm in this framework, we recall the linear inverse problem we are interested in:

$$\text{find } u \in \mathbb{R}^n \text{ such that} \quad b = b_0 + e\,, \qquad b_0 = \mathrm{K}u\,, \tag{3.1}$$

with $b, b_0, e \in \mathbb{R}^n$ and $\mathrm{K} \in \mathbb{R}^{n\times n}$. The process of replacing the original unstable and ill-posed problem with a nearby well-posed one goes under the name of *regularization*.
As already recalled in Section 2.3, the presence in the spectrum of K of relatively small singular values, may lead to a highly perturbed *naive* solution

$$u^* = \mathrm{K}^{-1}\, b = \mathrm{V}\,\Lambda^{-1}\,\mathrm{W}^T\, b = \sum_{i=1}^{n} \frac{w_i^T b}{\lambda_i} v_i\,.$$

To overcome this issue, the rank-1 matrices corresponding to the smallest singular values can be neglected, thus leading to

$$u^* = \tilde{\mathrm{K}}^{-1}\, b = \mathrm{V}\,\Lambda^{-1}\,\mathrm{W}^T\, b = \sum_{i=1}^{h} \frac{w_i^T b}{\lambda_i} v_i\,,$$

with $\lambda_1, \ldots, \lambda_h \in \mathbb{R}_+$ being the h larger singular values in the spectrum of K. This strategy goes under the name of truncated singular value decomposition (TSVD) [82] and it belongs to the wider class of *spectral filtering methods*. The approximated, or *filtered*, solution of the inverse problem in (3.1) via spectral filtering methods is given by

$$u^* = \sum_{i=1}^{n} \phi_i \frac{w_i^T b}{\lambda_i} v_i\,, \tag{3.2}$$

where ϕ_i are also referred to as *filter factors*. For the TSVD, they take the form

$$\phi_i = \begin{cases} 1 & \text{if} \quad i \le h \\ 0 & \text{otherwise} \end{cases}\,.$$

Another popular choice for the filter factors is given by

$$\phi_i = \frac{\lambda_i^2}{\lambda_i^2 + \alpha}\,, \quad \text{with} \quad \alpha > 0\,. \tag{3.3}$$

It is easy to show - see [83] for more details - that under the adoption of (3.3), the filtered solution (3.2) corresponds to the solution of the following minimization problem

$$u^* = \arg\min_{u\in\mathbb{R}^n} \{\|\mathrm{K}u - b\|_2^2 + \alpha\|u\|_2^2\}\,, \tag{3.4}$$

or, equivalently,

$$u^* = \arg\min_{u\in\mathbb{R}^n} \left\| \begin{pmatrix} \mathrm{K} \\ \sqrt{\alpha}\mathrm{I}_n \end{pmatrix} u - \begin{pmatrix} b \\ 0 \end{pmatrix} \right\|_2^2 .$$

What is interesting to remark here is that, when looking for the solution of (3.4) instead of computing (3.2), our perspective is slightly changing. In fact, spectral filtering methods lead eventually to a modification of the relationship between data and unknown, i.e. between effect and cause. However, one could alternatively solve the original linear system (3.1) by simultaneously penalizing the unwanted solutions. Observe that, when solving problem (3.4), we are trying to avoid solutions with large norm. It is possible to penalize different features of the final solution, by simply considering the more general model

$$u^* \in \arg\min_{u\in\mathbb{R}^n} \{\|\mathrm{K}u - b\|_2^2 + \alpha\|\mathrm{L}u\|_2^2\}\,, \tag{3.5}$$

where L is a linear operator, which is set based on the properties that we ask u^* to satisfy. A possible choice consists in considering, for instance, the image gradient, that is,

$$\mathrm{L} = \mathrm{D} = \begin{pmatrix} \mathrm{D}_h \\ \mathrm{D}_v \end{pmatrix} \in \mathbb{R}^{2n\times n},$$

where D_h, $\mathrm{D}_v \in \mathbb{R}^{n\times n}$ are the finite difference matrices discretizing the horizontal and vertical partial derivative operator, respectively. In case of Dirichlet boundary conditions they are defined as

$$\mathrm{D}_v = \mathrm{I}_{\sqrt{n}} \otimes \mathrm{B}\,, \quad \mathrm{D}_h = \mathrm{B} \otimes \mathrm{I}_{\sqrt{n}}\,, \quad \text{with} \quad \mathrm{B} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ -1 & 1 & \dots & 0 \\ & & \ddots & \\ 0 & \dots & -1 & 1 \end{pmatrix} \in \mathbb{R}^{\sqrt{n}\times\sqrt{n}}, \tag{3.6}$$

where "$\otimes$" denotes the Kronecker product - note that the matrix B slightly modifies when different boundary conditions are adopted. The action of the first order differential operator on u highlights image structures, such as edges or texture, that should be preserved. On the other hand, D also detects the spikes generated by the noise.

Imposing a penalization on the magnitude of Du will thus encourage solutions with few jumps, favoring the removal of the corrupting noise. Note that the greater the α, the stronger is the request of regularity.

Problem (3.5) is known as *Tikhonov* regularization model and, under certain assumptions, it admits an unique solution [156, 157, 155]. In particular, the null spaces of K and L must have trivial intersection, that is null(K) $\cap$ null(L) $= \{\mathbf{0}_n\}$. Tikhonov model belongs to the class of *variational methods* for determining a regularized solution of inverse problems, whose detailed analysis is addressed in this chapter.

3.1 Variational methods

At the core of variational methods there is the idea of substituting the ill-posed problem (3.1) with a well-posed one, consisting of the minimization of a cost functional $\mathcal{J} : \mathbb{R}^n \to \mathbb{R}_+$. In formula,

$$u^* \in \arg\min_{u\in\mathbb{R}^n} \{\mathcal{J}(u) := \mathcal{F}(u; \mathrm{K}, b) + \alpha\mathcal{R}(u)\}\,, \tag{3.7}$$

where u^* is an approximation of the solution of the original problem. The regularization parameter $\alpha > 0$ controls the trade-off between the two terms and its setting is usually a quite delicate issue; the functionals $\mathcal{F}$ and $\mathcal{R}$ are commonly referred to as the *data fidelity* and the *regularization* term, respectively. The $\mathcal{F}$ measures the 'distance' between the given image b and u after the action of the operator K with respect to some norm corresponding to the noise statistics in the data - see, e.g. [150] - while $\mathcal{R}$ encodes prior information on the desired image u (such as its regularity and its sparsity patterns). Here, we are mainly interested in the design of novel regularization terms with the purpose of driving the restoration taking into account the local properties of the image.

In the following we are adopting a Bayesian approach for the derivation of variational models, based on an interpretation of the data fidelity term and of the regularization term coming from well-defined pdfs. Clearly, we do not necessarily need to invoke a Bayesian framework in order to write down a generic variational model. That is, for instance, the case of Tikhonov model (3.5) that has been derived by means of a pure deterministic approach. Nevertheless, the Bayesian approach provides a rational basis for the choice of the functionals $\mathcal{R}$ and $\mathcal{F}$. Moreover, the adoption of a probabilistic framework allows to highlight the role of our beliefs on the solution in the design of $\mathcal{R}$. In other words, the art of regularization somehow involves our subjectivity.

3.2 Bayesian formulation

The first fruitful synergy between Bayesian formulation and inverse problems can be traced to applications in the geophysical field. In particular, a first explicit formalization of this idea can be found in the seminal works [151, 152]. As surveys extending the topic to other inverse problems, such as image processing, we mention [102, 101, 29] and [31], where the authors propose an interesting historical review of the relation between Bayesian approach and inverse problems.

Recasting the generic image restoration problem into a Bayesian perspective requires to interpret all unknown quantities involved in the model (3.1) as random variables. The goal here is to find the distribution of u, combining the information encoded in the observed data b with *a priori* beliefs or information that we may have on u.

More specifically, the discrete deterministic model (3.1) takes the following probabilistic form

$$B = \mathrm{K}\,U + E\,, \tag{3.8}$$

where B, U, and E are n-variate random vectors whose realizations are the ones denoted in the deterministic model as b, u and e, respectively. Observe that here only the realization b of the random variable B is actually available and we introduce the pdf related to E

$$\mathrm{P}(e) = \mathrm{P}(b - \mathrm{K}u) = \mathrm{P}(b \mid u). \tag{3.9}$$

We refer to $\mathrm{P}(b \mid u)$ as the *likelihood* pdf, since it expresses the likelihood of different measurement outcomes with $U = u$ given. In the same fashion, we denote by $\mathrm{P}(u)$ the *prior* pdf of the random variable U. The update of the prior based on the observed measurement b is given by the *posterior* pdf, that is the conditional distribution $\mathrm{P}(u \mid b)$ and is related to the prior and the likelihood via the *Bayes' formula*

$$\mathrm{P}(u \mid b) = \frac{\mathrm{P}(b \mid u)\,\mathrm{P}(u)}{\mathrm{P}(b)} \propto \mathrm{P}(b \mid u)\,\mathrm{P}(u)\,,$$

where

$$\mathrm{P}(b) = \int_{\mathbb{R}^n} \mathrm{P}(b \mid u)\,\mathrm{P}(u)\,du\,,$$

is called the *evidence* term.

In the following, we first review some popular and widespread choices for prior and likelihood terms in image restoration. Then, we will focus on how to extract handful information from the posterior pdf, once that its analytic expression is available.

3.2.1 Likelihoods

The choice of likelihood pdf is strictly connected to the statistical assumptions on the physical acquisition processes for the problem of interest, that is on the noise in the measurements. Its derivation is particularly straightforward in the case of additive noise - see (3.9).

Let E be a zero-mean Gaussian random variable, that is the original image is corrupted by a zero-mean Gaussian noise. In formula, $E \sim \mathcal{N}(0, \Sigma)$, where the covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$ is a symmetric positive definite with possibly non-zero off-diagonal entries. We denote by θ_{lkh} the possibly matricial parameters involved in the definition of the likelihood term, which in this case reads as

$$\theta_{\mathrm{lkh}} = \Sigma\,.$$

When U and E are assumed to be independent random variables, we have

$$\begin{aligned}
\mathrm{P}(b \mid u, \theta_{\mathrm{lkh}}) =& \mathrm{P}(b - \mathrm{K}u \mid \theta_{\mathrm{lkh}}) \\
=& \frac{1}{W} \exp\Big(- \frac{1}{2}(b - \mathrm{K}u)^T \Sigma^{-1} (b - \mathrm{K}u)\Big) \\
=& \frac{1}{W} \exp\Big(- \frac{1}{2}\|\mathrm{S}(b - \mathrm{K}u)\|_2^2\Big),
\end{aligned} \tag{3.10}$$

where W denotes the normalization constant and the matrix S in (3.10) is the lower-triangular Cholesky factor of Σ^{-1}, i.e. $\Sigma^{-1} = \mathrm{S}^T\mathrm{S}$.

In the specific case of an additive white Gaussian noise (AWGN), introduced in Section 2.2.1, the covariance matrix Σ is a scaled identity, $\Sigma = \sigma^2 \mathrm{I}_n$, and the pdf in (3.10) is of the form

$$\mathrm{P}(b \mid u, \theta_{\mathrm{lkh}}) = \frac{1}{W} \exp\Big(- \frac{1}{2\sigma^2}\|b - \mathrm{K}u\|_2^2\Big). \tag{3.11}$$

In this specific case, the parameter θ_{lkh} reduces to the standard deviation σ.
Let now e be a realization of a white Laplacian random variable, i.e. $E \sim Laplace(0, \Sigma)$, with $\Sigma = \sigma^2 \mathrm{I}_n$. The likelihood pdf now reads

$$\mathrm{P}(b \mid u, \theta_{\mathrm{lkh}}) = \frac{1}{W} \exp\left(- \frac{\|\mathrm{K}u - b\|_1}{\sigma} \right),$$

where, as before, W is the normalization constant and $\theta_{\mathrm{lkh}} = \sigma$.

3.2.2 Priors

A priori assumptions may concern different properties of the image. The challenging aspect of designing priors is the process of turning *qualitative* information into *quantitative* terms. For instance, if the processed image is known to be piece-wise constant, then, recalling the finite difference matrices definition in (3.6), we expect the vector with entries

$$\|(\mathrm{D}u)_i\|_2 = \sqrt{(\mathrm{D}_h u)_i^2 + (\mathrm{D}_v u)_i^2}\,,$$

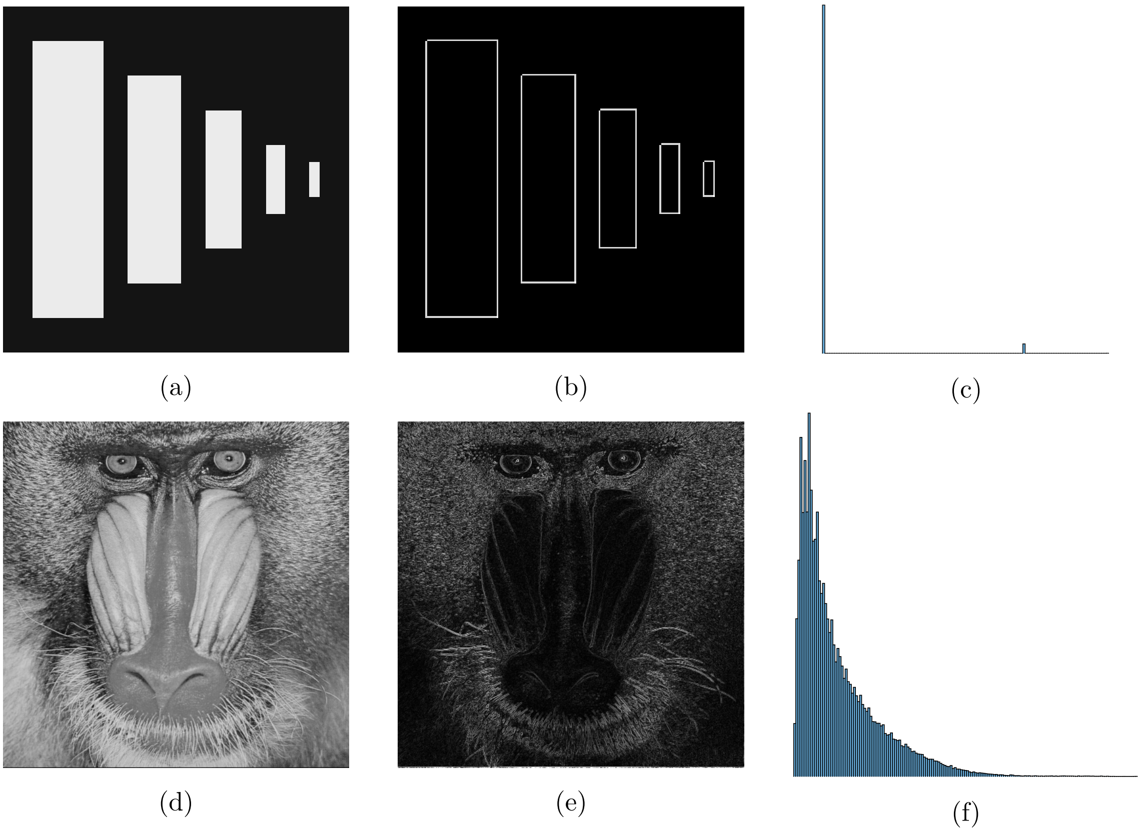
to be sparse, since it is reasonable to assume such an image to exhibit a few jumps. Consider the image `rectangles` in Figure 3.1a and the gradient magnitudes of the pixels represented in Figure 3.1b. As a further evidence of this, observe that the histogram in 3.1c exhibits a bi-modal distribution. Clearly, we do not expect the same behavior for the gradient magnitudes when considering a natural image presenting textures, as for image `mandrill` in Figure 3.1d. It is easy to see that the gradient structure is much richer than in the previous case - see Figure 3.1e and Figure 3.1f.
More generally, typical priors for image restoration problems encode information on the distribution of the gray levels within an image and the transition of gray-scale intensities between different areas of the image. As stated in [76], pixel-gray levels can be viewed as states of atoms in a lattice-like physical system. In this framework, it is very usual to model the unknown image as a Markov random field (MRF), that can be basically interpreted as an extension of Markov processes in more than one dimension. In other words, we ask that a selected feature at the generic pixel i of u only depends on the behavior of u at pixels belonging to $\mathcal{C}_i$, with $\mathcal{C}_i$ being the index set of neighbors of u_i. For instance, if the selected feature is the gray level, the 2-D Markovian property at pixel i reads

$$\mathbb{P}(U_i = u_i \mid U_j = u_j\,,\ j \neq i) = \mathbb{P}(U_i = u_i \mid U_j = u_j\,,\ j \in \mathcal{C}_i). \tag{3.12}$$

The prior distribution for a MRF is the so-called *Gibbs prior*

$$\mathrm{P}(u \mid \theta_{\mathrm{pr}}) = \frac{1}{Z} \exp\left(- \sum_{i=1}^{n} V_{\mathcal{C}_i^r}(u; \theta_{\mathrm{pr}}) \right),$$

where Z is a normalization constant, $V_{\mathcal{C}_i^r}$ is also referred to as the *Gibbs potential function* defined on a clique of pixels of radius r centered at pixel i, and θ_{pr} denotes the parameter involved in the expression of the prior.
A very natural choice is to design priors based on the properties that we expect the discrete gradient of the image $\mathrm{D}u$ to satisfy. When adopting a forward finite difference scheme, definition


(a) (b) (c)

(d) (e) (f)

Figure 3.1: *First row.* Test image `rectangles` (a), gradient magnitude (b), histogram of the gradient magnitudes (c). *Second row.* Test image `mandrill` (d), gradient magnitude (e), histogram of the gradient magnitudes (f).

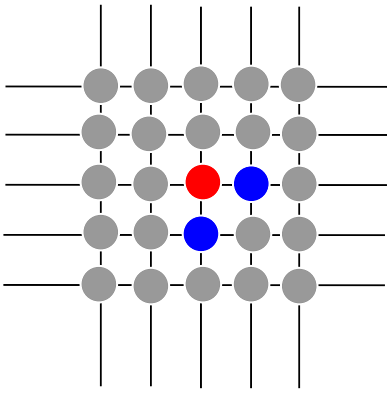


Figure 3.2: Pixels represented as atoms in a lattice. The colored ones belong to the clique related to red atom. In particular, the blue atoms are involved in the computation of the finite forward difference for the red atom.

(3.12) turns into

$$\mathbb{P}(U_i = u_i \mid U_j = u_j \,,\; j \neq i) = \mathbb{P}(U_i = u_i \mid U_{\text{right}} = u_{\text{right}} \,,\; U_{\text{down}} = u_{\text{down}}). \tag{3.13}$$

The configuration of the generic clique corresponding to this case is shown in Figure 3.2. Condition (3.13) states that the generic Gibbs potential function $V_{\mathcal{C}_i^r}$ is defined on a discrete set of cardinality 3, namely $\{u_i, u_{\text{right}}, u_{\text{down}}\}$, which are the values involved in the computation of the discrete gradient at pixel i. For instance, we may assume

$$DU = \Phi \,, \quad \text{with} \quad \Phi \sim \mathcal{N}(0, \Gamma) \,,\; \Gamma = \gamma^2 I_n. \tag{3.14}$$

Condition (3.14) is equivalent to assuming that the vertical and horizontal differences in the image behave as white Gaussian random variables with standard deviation γ. This can be expressed also as

$$\mathrm{P}(u \mid \theta_{\mathrm{pr}}) = \frac{1}{Z}\exp\left(- \sum_{i=1}^{n} \frac{1}{2\,\gamma^2}\|(\mathrm{D}u)_i\|_2^2 \right) = \frac{1}{Z}\exp\left(- \frac{1}{2\,\gamma^2}\|\mathrm{D}u\|_2^2 \right), \tag{3.15}$$

with $\theta_{\mathrm{pr}} = \gamma$. The pdf in (3.15), to which we can also refer as *first-order smoothness prior*, is a Gibbs prior with $V_{\mathcal{C}_i^r}(u) = \|(\mathrm{D}u)_i\|_2^2$ defined over a clique of radius $r = 1$. One can start noticing a connection, that will be made explicit in the next, with the above defined prior and the Tikhonov regularization term.

Another very popular prior for images modeled as MRF, is the *total variation prior*. For $u \in \mathbb{R}^n$, we define the *isotropic total variation* (TVI) as

$$\mathrm{TVI} = \sum_{j=1}^{n}\|(\mathrm{D}u)_j\|_2 = \sum_{j=1}^{n}\sqrt{(\mathrm{D}_h u)_j^2 + (D_v u)_j^2}\,. \tag{3.16}$$

Definition (3.16) can be slightly modified by considering the ℓ_1-norm instead of the euclidean norm, thus getting to the anisotropic total variation (TVA),

$$\mathrm{TVA} = \sum_{j=1}^{n}\|(Du)_j\|_1 = \sum_{j=1}^{n}|(\mathrm{D}_h u)_j| + |(\mathrm{D}_v u)_j|\,.$$

Setting the Gibbs potential functions as

$$V_{\mathcal{C}_j^r}(u) = \|(\mathrm{D}u)_j\|_q\,, \quad r = 1\,, \quad q \in \{1,2\}\,, \quad j = 1,...,n\,,$$

the corresponding TV prior can be expressed as

$$\mathrm{P}(u \mid \theta_{\mathrm{pr}}) = \left\{ \begin{array}{ll} \frac{1}{Z}\exp\left(-\beta\sum_{j=1}^{n}\|(\mathrm{D}u)_j\|_2 \right) & = \frac{1}{Z}\exp(-\beta\mathrm{TVI}(u)) \\ \frac{1}{Z}\exp\left(-\beta\sum_{j=1}^{n}\|(\mathrm{D}u)_j\|_1 \right) & = \frac{1}{Z}\exp(-\beta\mathrm{TVA}(u)) \end{array} \right. , \tag{3.17}$$

where $\theta_{\mathrm{pr}} = \beta > 0$ and the two definitions depend on the isotropic/anisotropic definition of TV above. In general, better performances are observed when TVI is adopted. Nonetheless, TVA is a more suitable choice in presence of images presenting edges oriented only along the x and the y axes. This is mainly due to the properties of the level curves of the two regularization terms, reported in Figure 3.3. In fact, while the diffusion produced by TVI is the same in every direction of the xy cartesian plane, the choice of TVA produces a more concentrated diffusion along the axes. A key observation for the following discussion is that, as already highlighted in [111], the adoption of a TVI prior and of a TVA prior, is equivalent to assuming that the ℓ_2-norm and the ℓ_1-norm of the gradients of the image, respectively, distribute according to an *exponential* or *half-Laplacian* distribution with *scale* parameter β. In fact, the pdf of an univariate half-Laplacian random variable X takes the form

$$\mathrm{P}(x) = \left\{ \begin{array}{ll} \beta\exp\left(-\beta x\right) & \text{if } x > 0\,, \\ 0 & \text{otherwise} \end{array} \right. . \tag{3.18}$$

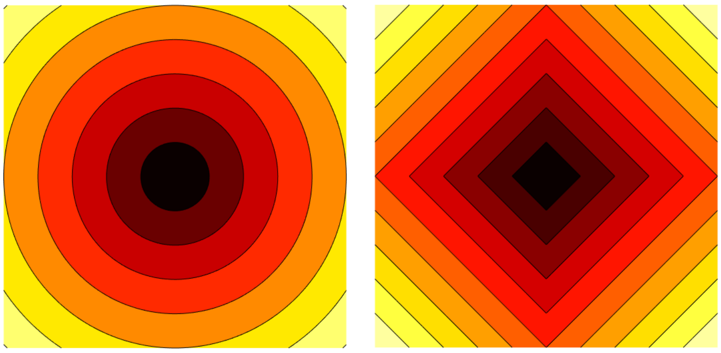


Figure 3.3: Level curve of TVI (a) and TVA (b) regularization term.

Clearly, one could have at disposal information about the image or about its representation in a given basis, rather than about its gradient structure. For instance, in the case of a sparse signal, the pixels are all assumed to be close to 0, hence they can be modeled as independent zero-mean Gaussian random variables. In particular, the potential functions $V_{\mathfrak{C}_i^r}$ takes the form

$$V_{\mathfrak{C}_i^r} = \frac{u_i^2}{2\gamma^2}\,, \quad \text{with} \quad r = 1\,,$$

and the corresponding Gibbs' prior becomes

$$\mathrm{P}(u \mid \theta_{\mathrm{pr}}) = \frac{1}{Z}\exp\left(-\sum_{i=1}^{n}\frac{u_i^2}{2\gamma^2}\right) = \frac{1}{Z}\exp\left(-\frac{1}{2\gamma^2}\|u\|_2^2\right)\,,$$

where $\theta_{\mathrm{pr}} = \gamma$ and Z is the normalization constant. Adopting a white Gaussian prior as the one in (3.2.2) is equivalent to set a Tikhonov prior on the image rather than on the magnitude of its gradient. Analogously, one could set

$$V_{\mathfrak{C}_i^r} = \beta|u_i|\,, \quad \text{with} \quad r = 1\,,$$

thus getting the ℓ_1-prior,

$$\mathrm{P}(u \mid \theta_{\mathrm{pr}}) = \frac{1}{Z}\exp\left(-\sum_{i=1}^{n}\beta|u_i|\right) = \frac{1}{Z}\exp\left(-\beta\|u\|_1\right)\,,$$

which is closely related to the TV prior.

3.2.3 Handling the posterior: MAP and CM single point estimates

Once both prior and likelihood terms have been set, based on reasonable assumptions on the image and on the information available on the noise, respectively, Bayes' formula is applied in order to recover an expression for the posterior distribution. It is worth emphasizing here that the goal of variational models, whose generic expression is given in (3.7), is to find an estimate u^* as accurate as possible of the original image u. In the Bayesian perspective, the solution is not a single point estimate but a posterior probability distribution $\mathrm{P}(u \mid b)$. Therefore, the main issue here is how to extract a single-point meaningful information from the posterior. To this purpose, a popular choice is to use the mode of the posterior as a representative of the distribution. In other words, we are interested in finding the image u maximizing the posterior, and thus presenting the highest occurrence probability. This procedure is known as Maximum A Posterior (MAP) estimation approach. In the following, we are dropping off the dependence from the parameters θ_{lkh} and θ_{pr} in order to avoid heavy notations. In formula,

$$u_{\text{MAP}} \in \arg\max_{u\in\mathbb{R}^n} \mathrm{P}(u \mid b) \propto \mathrm{P}(b \mid u)\mathrm{P}(u) \tag{3.19}$$

Note that the evidence term $\mathrm{P}(b)$ term in the Bayes' formula can be omitted since the minimization is with respect to u. Applying the negative logarithm to (3.19), we have

$$u_{\text{MAP}} \in \arg\min_{u\in\mathbb{R}^n} \{-\log \mathrm{P}(b \mid u) - \log \mathrm{P}(u)\}\ , \tag{3.20}$$

hence, the problem of finding the MAP single point estimate turns out to be an optimization problem that, in several cases, ends up to be of the form of the generic variational model (3.7), as we are going to show later. This is the reason why the Bayesian formulation is typically used as a supporting machinery for the generation of new variational models. It must be remarked that the solution of problem (3.20) may not exist or, if existing, it may not be unique. Nevertheless, when the solution exists and it is unique, fast and robust optimization algorithm can be adopted to address the minimization problem.

Instead of looking for the mode of the posterior distribution, a different single point estimate, which is in general more informative than the MAP estimate, can be considered, namely the *conditional mean* (CM), that is the mean of the posterior distribution $\mathrm{P}(u \mid b)$:

$$u_{\text{CM}} = \int_{\mathbb{R}^n} u\,\mathrm{P}(u \mid b)du\ = \int_{\mathbb{R}^n} u\,\frac{\mathrm{P}(b \mid u)\mathrm{P}(u)}{\mathrm{P}(b)}du\,. \tag{3.21}$$

In order to quantify the precision of such estimate, one can explore the spread around the CM value, by computing the covariance matrix

$$\Gamma_{\text{CM}} = \int_{\mathbb{R}^n} (u - u_{CM})(u - u_{CM})^T\,\frac{\mathrm{P}(b \mid u)\mathrm{P}(u)}{\mathrm{P}(b)}du. \tag{3.22}$$

In very high dimensional settings, quadrature rules to compute the integrals in (3.21)-(3.22) can not be applied, and one has to resort to other techniques, such as Markov Chain Monte Carlo (MCMC) methods, in order to sample from the posterior and then computing both mean and covariance matrix - notice that an additional issue is the computation of the evidence term $\mathrm{P}(b)$ term that here, unlike in the MAP case, can not be omitted. Nonetheless, sampling methods are usually very expensive in terms of computational costs.

In general, the MAP and the CM estimates are different. So, one has to make a choice between the level of trust we put in the estimate and the algorithmic issues to face, keeping always in mind that, independently from the final choice, approximating a whole distribution with a single point estimate unavoidably leads to a loss of information.

As suggested from the mentioned parallelism between Bayesian approach and variational methods, here we summarize the posterior in terms of the the MAP estimate.

3.2.4 Parameter dependence

It is clear that the choice of the parameters θ_{lkh} and θ_{pr} has a remarkable influence on the distribution used to describe either the information available on the data or the assumption stated on the unknown. Typically, the parameters θ_{lkh} involved in the analytical expression of the likelihood term are related to the noise level, that here is assumed to be known - see Section 3.4 - as in the case of the standard deviation σ in the definition of AWGN pdf (3.11). On the contrary, the presence of possibly unknown parameters in the expression of the prior pdf is a matter that is worth investigating. For instance, an interesting issue is how to set the parameters $\theta_{\mathrm{pr}} = \gamma$ and $\theta_{\mathrm{pr}} = \beta$ in the Tikhonov (3.15) and TV prior (3.17), respectively, in order to enforce the smoothness or the sparsity of the magnitude of the gradients in a way that is in agreement with the properties of the original and unknown image u. A substantial part of the thesis will be devoted to answering this question. For the moment, we start outlining the main strategies that will be adopted in the following. We remark that, in general, we are dealing with a vector of unknown parameters $\theta_{\mathrm{pr}} \in \mathbb{R}^q$.

(i) In the absence of any intuition about the meaning of the vector θ_{pr}, a *naive* strategy consists of tuning its entries manually. In general, this should be avoided, in particular for $q \gg 1$.

(ii) In the case θ_{pr} being a scalar, it can be automatically estimated by imposing a constraint on the set of the admissible solutions - see Section 3.3-3.4.

(iii) Alternatively, the unknown vector θ_{pr} can be modeled as a random variable Θ_{pr}, resorting to the same approach already adopted for the unknown image u. This layering in the process of knowledge acquisition is at the core of *hierarchical modeling*. We can thus introduce a prior pdf $\mathrm{P}(\theta_{\mathrm{pr}})$, also known as *hyperprior*, on the random variable Θ_{pr}. The unknown to be estimated is now the coupled vector $(u, \theta_{\mathrm{pr}})$, and the joint prior takes the form:

$$\mathrm{P}(u, \theta_{\mathrm{pr}}) = \mathrm{P}(u \mid \theta_{\mathrm{pr}}) \mathrm{P}(\theta_{\mathrm{pr}}) \, .$$

Consequently, the MAP estimation problem reads as

$$\begin{aligned}
(u^*, \theta_{\mathrm{pr}}^*) &\in \arg \max_{(u,\theta_{\mathrm{pr}}) \in \mathbb{R}^n \times \mathbb{R}^q} \mathrm{P}(u, \theta_{\mathrm{pr}} \mid b) \\
&= \arg \max_{(u,\theta_{\mathrm{pr}}) \in \mathbb{R}^n \times \mathbb{R}^q} \left\{ \frac{\mathrm{P}(u, \theta_{\mathrm{pr}}) \mathrm{P}(b \mid u, \theta_{\mathrm{pr}})}{\mathrm{P}(b)} \right\} \\
&= \arg \max_{(u,\theta_{\mathrm{pr}}) \in \mathbb{R}^n \times \mathbb{R}^q} \{ \mathrm{P}(\mathrm{u} \mid \theta_{\mathrm{pr}}) \mathrm{P}(\theta_{\mathrm{pr}}) \mathrm{P}(\mathrm{b} \mid \mathrm{u}) \} \\
&= \arg \min_{(u,\theta_{\mathrm{pr}}) \in \mathbb{R}^n \times \mathbb{R}^q} \{ -\log \mathrm{P}(u \mid \theta_{\mathrm{pr}}) - \log \mathrm{P}(\theta_{\mathrm{pr}}) - \log \mathrm{P}(b \mid u) \} \; , \qquad (3.23)
\end{aligned}$$

where

$$\mathrm{P}(b \mid u, \theta_{\mathrm{pr}}) = \mathrm{P}(b \mid u)$$

due to the conditional independence of the random variables B and Θ_{pr} given U. In this thesis, problem (3.23) is addressed by resorting to *iterative alternating sequential* (IAS) algorithm [28]. Setting $t \in \mathbb{N}$ and providing an initialization for $u^{(0)}$ at $t = 0$, we have

$$\theta_{\text{pr}}^{(t)} \in \arg\min_{\theta_{\text{pr}} \in C_{\text{pr}}} \left\{ -\log \mathrm{P}(u^{(t-1)} \mid \theta_{\text{pr}}) - \log \mathrm{P}(\theta_{\text{pr}}) \right\}, \tag{3.24}$$

$$u^{(t)} \in \arg\min_{u \in \mathbb{R}^n} \left\{ -\log \mathrm{P}(u \mid \theta_{\text{pr}}^{(t)}) - \log \mathrm{P}(b \mid u) \right\}, \tag{3.25}$$

where $C_{\text{pr}} \subseteq \mathbb{R}^q$ is the eventual constraint set for the minimization problem with respect to θ_{pr}. Observe that in case $C_{\text{pr}} = \mathbb{R}^q$, the minimization problem (3.24) is unconstrained. The IAS algorithm can be further detailed in two sub-cases:

(a) when an intuition on the behavior of θ_{pr} is not available, the hyperprior $\mathrm{P}(\theta_{\text{pr}})$ can be set in order to be *non-informative*. To this aim, an uniform distribution for the random variable Θ_{pr} is fixed:

$$\mathrm{P}(\theta_{\text{pr}}) = \begin{cases} \frac{1}{\mu(C_{\text{pr}})}, & \text{if } \theta_{\text{pr}} \in C_{\text{pr}} \\ 0, & \text{otherwise} \end{cases},$$

where $\mu(C_{\text{pr}})$ is the measure of the set C_{pr}. Moreover, assuming an uniform distribution, the constrained minimization problem simplifies to

$$\theta_{\text{pr}}^{(t)} \in \arg\min_{\theta_{\text{pr}} \in C_{\text{pr}}} \left\{ -\log \mathrm{P}(u^{(t-1)} \mid \theta_{\text{pr}}) \right\}.$$

(b) In presence of crucial and sensible information available on θ_{pr}, an *informative* prior can also be adopted. In this case, the alternating minimization scheme is implemented as it stands in (3.24)-(3.25).

3.3 Deriving Tikhonov-L$_2$ and TV-L$_2$ models via MAP

We now look into the derivation of three variational models via Bayesian approach and MAP estimation. Starting from the linear degradation model (3.8), let the random variable E model an AWGN, i.e. $E \sim \mathcal{N}(0, \sigma^2 I_n)$. Subsequently, as already observed in Section 3.2.1, the parameter in the expression of the likelihood term will simply be $\theta_{\text{lkh}} = \sigma$.

First, we set a first-order smoothness prior (3.15) for U. We recall that, in this case, $\theta_{\text{pr}} = \gamma$. Neglecting the normalization constants, we have

$$\begin{aligned} \mathrm{P}(u \mid b) &\propto \mathrm{P}(u \mid \theta_{\text{pr}})\, \mathrm{P}(b \mid u, \theta_{\text{lkh}}) \\ &= \exp\left(-\frac{1}{2\gamma^2} \sum_{i=1}^{n} \|(\mathrm{D}u)_i\|_2^2 \right) \exp\left(-\frac{1}{2\sigma^2} \|\mathrm{K}u - b\|_2^2 \right) \\ &= \exp\left(-\left\{ \frac{1}{2\gamma^2} \sum_{i=1}^{n} \|(\mathrm{D}u)_i\|_2^2 + \frac{1}{2\sigma^2} \|\mathrm{K}u - b\|_2^2 \right\} \right) \\ &= \exp\left(-\left\{ \alpha \sum_{i=1}^{n} \|(\mathrm{D}u)_i\|_2^2 + \frac{1}{2} \|\mathrm{K}u - b\|_2^2 \right\} \right), \end{aligned} \tag{3.26}$$

with $\alpha = \sigma^2/(2\,\gamma^2)$. By plugging (3.26) in (3.19) and taking the negative logarithm, it follows that

$$u^* \in \arg\min_{u \in \mathbb{R}^n} \left\{ \alpha \sum_{i=1}^{n} \|(\mathrm{D}u)_i\|_2^2 + \frac{1}{2} \|\mathrm{K}u - b\|_2^2 \right\}, \tag{3.27}$$

that is the solution of the Tikhonov model (3.5) introduced previously.
As a second example, we consider the TV prior. In this case, $\theta_{\mathrm{pr}} = \beta$, and the posterior density takes the form

$$\begin{aligned}
\mathrm{P}(u \mid b) &\propto \mathrm{P}(u \mid \theta_{\mathrm{pr}})\, \mathrm{P}(b \mid u, \theta_{\mathrm{lkh}}) \\
&= \exp\Big(-\beta \sum_{i=1}^{n} \|(\mathrm{D}u)_i\|_2 \Big) \exp\Big(-\frac{1}{2\sigma^2} \|\mathrm{K}u - b\|_2^2 \Big) \\
&= \exp\Big(-\Big\{ \beta \sum_{i=1}^{n} \|(\mathrm{D}u)_i\|_2 + \frac{1}{2\sigma^2} \|\mathrm{K}u - b\|_2^2 \Big\} \Big) \\
&= \exp\Big(-\Big\{ \alpha \sum_{i=1}^{n} \|(\mathrm{D}u)_i\|_2 + \frac{1}{2} \|\mathrm{K}u - b\|_2^2 \Big\} \Big), \qquad (3.28)
\end{aligned}$$

with $\alpha = \beta\, \sigma^2$. By plugging (3.28) in (3.19) and taking the negative logarithm, we end up with

$$u^* \in \arg\min_{u \in \mathbb{R}^n} \Big\{ \alpha \sum_{i=1}^{n} \|(\mathrm{D}u)_i\|_2 + \frac{1}{2} \|\mathrm{K}u - b\|_2^2 \Big\}, \qquad (3.29)$$

that is the popular TV-L_2 or TV-ROF model [137].
For the sake of completeness, we present the derivation of the variational model corresponding to the choice of the TVA prior. We have

$$\begin{aligned}
\mathrm{P}(u \mid b) &\propto \mathrm{P}(u \mid \theta_{\mathrm{pr}})\, \mathrm{P}(b \mid u, \theta_{\mathrm{lkh}}) \\
&= \exp\Big(-\beta \sum_{i=1}^{n} \|(\mathrm{D}u)_i\|_1 \Big) \exp\Big(-\frac{1}{2\sigma^2} \|\mathrm{K}u - b\|_2^2 \Big) \\
&= \exp\Big(-\Big\{ \beta \sum_{i=1}^{n} \|(\mathrm{D}u)_i\|_1 + \frac{1}{2\sigma^2} \|\mathrm{K}u - b\|_2^2 \Big\} \Big) \\
&= \exp\Big(-\Big\{ \alpha \sum_{i=1}^{n} \|(\mathrm{D}u)_i\|_1 + \frac{1}{2} \|\mathrm{K}u - b\|_2^2 \Big\} \Big),
\end{aligned}$$

and the TVA-L_2 variational model reads as

$$u^* \in \arg\min_{u \in \mathbb{R}^n} \left\{ \alpha \sum_{i=1}^{n} \|(\mathrm{D}u)_i\|_1 + \frac{1}{2} \|\mathrm{K}u - b\|_2^2 \right\},$$

where, again, $\alpha = \beta\sigma^2$. The three models derived in this section, are of the form

$$u^* \in \arg\min_{u \in \mathbb{R}^n} \Big\{ \alpha \mathcal{R}(u) + \frac{1}{2} \|\mathrm{K}u - b\|_2^2 \Big\}, \qquad (3.30)$$

with

$$\mathcal{R}(u) = \sum_{i=1}^{n} \|(\mathrm{D}u)_i\|_2^2, \quad \mathcal{R}(u) = \sum_{i=1}^{n} \|(\mathrm{D}u)_i\|_2, \quad \text{or} \quad \mathcal{R}(u) = \sum_{i=1}^{n} \|(\mathrm{D}u)_i\|_1,$$

depending on the case. Notice that model (3.30) can be reformulated equivalently as

$$u^* \in \arg\min_{u \in \mathbb{R}^n} \Big\{ \mathcal{R}(u) + \frac{\mu}{2} \|\mathrm{K}u - b\|_2^2 \Big\},$$

with $\mu = 1/\alpha$ playing, as well as α, the role of the regularization parameter, since it balances the contribution of fidelity and regularization term in the functional.

We remark that while the number of free parameters for each model is equal to 2, one coming from the likelihood term - namely, the standard deviation σ - and one coming from the prior term - i.e. γ for Tikhonov and β for TV, when setting

$$\begin{array}{llll} \alpha = \frac{\sigma^2}{2\gamma^2} & \text{or equivalently} & \mu = \frac{2\gamma^2}{\sigma^2} & \text{for Tikhonov,} \\ \alpha = \beta\sigma^2 & \text{or equivalently} & \mu = \frac{1}{\beta\sigma^2} & \text{for TV,} \end{array}$$

the 2 degrees of freedom reduces to 1. In Section 3.4, we will give some details on how the single global parameter α, or μ, can be estimated solely from the information available on the noise.

Example In order to explore the influence of the prior on the final restoration and to make clear how to set a reasonable prior based on the information available on the image, we compare the performance of the above derived models (3.27) and (3.29) on the restorations of two test images, namely `square` in Figure 3.4a and `sinusoid` in Figure 3.4e, with different properties: `square` is a typical blocky image, with a sparse gradient structure, while `sinusoid` is characterized by smooth gradients. The images have been corrupted by Gaussian blur and AWGN - see Figure 3.4b and Figure 3.4f, respectively. For the moment, we do not test the performance of the TVA-L_2 model because we want to highlight the influence of different assumptions on the gradients magnitude, while the TVI and TVA priors behave very similarly from this point of view.

The minimizers u^* of the Tikhonov and of the TVI variational models coupled with L_2 fidelity term are computed via Alternating Direction Method of multipliers (ADMM) [14] - see Appendix B. The regularization parameter α has been set so as to fulfill the *discrepancy principle* [117] - more details about this will be given in the next section.

The restorations of `square` for the Tikhonov and the TVI model are shown in Figure 3.4d and Figure 3.4c, respectively. The choice of a TVI prior is more reasonable and produces sharper edges, while the restorations via Tikhonov model appears still blurred since the corresponding prior does not naturally describe sparse gradient structures. On the other hand, the Tikhonov prior is more suitable to describe images such as the test image `sinusoid` - see Figure 3.4h - for which TVI is not capable of preserving the richness of the gradient structure. The main drawback of the TVI is the *staircase* effect, due to the tendency of TVI prior to promote piecewise constant images; this typical downside that TVI regularization brings along will be more deeply investigated in the next chapter.

3.4 Automatic estimation of the global regularization parameter

A substantial literature related to inverse problems in imaging has been devoted to the derivation of methods for automatically tuning the regularization parameter α - or, equivalently, μ. Overall, those methods can be divided into two classes:

1. methods relying on the noise level, that can be either known or accurately estimated;

2. methods exploiting only the information encoded in the data b.

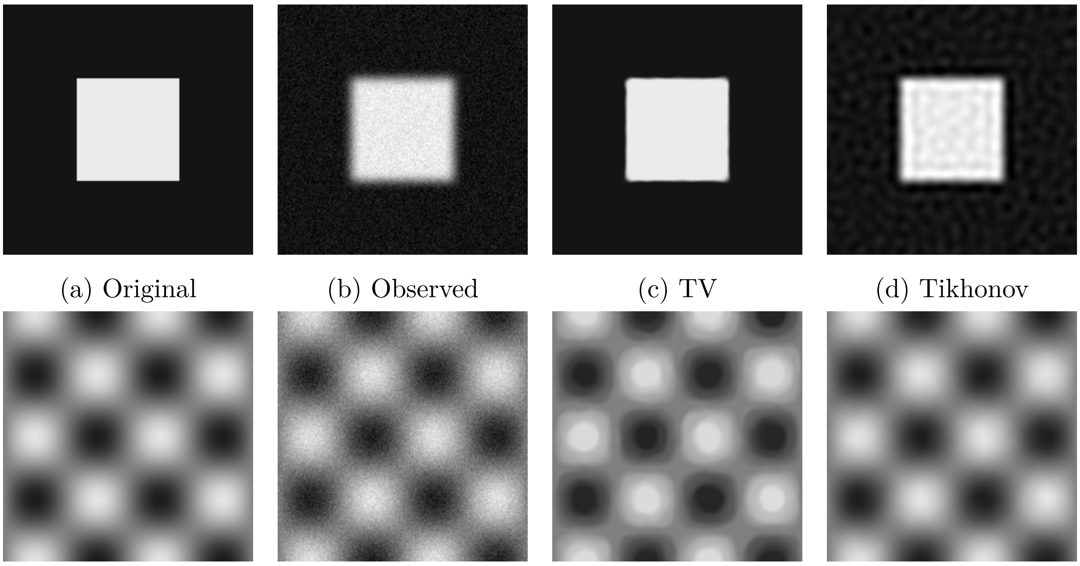


Figure 3.4: *First row.* Test image `square` (a), observed image (b), TVI restoration (c), Tikhonov restoration (d) *Second row.* Test image `sinusoid` (e), observed image (f), TV restoration (g), Tikhonov restoration (h).

Typically, we refer to methods belonging to the second class as *heuristic* methods. Among the heuristic methods, we mention the Generalized Cross Validation (GCV) [52] and the L-curve method [22, 84]. Denoting by u_α^* the solution of the variational model corresponding to a fixed choice of α, the GCV choses the values of the regularization parameter that satisfies

$$\alpha^* \in \arg\min_{\alpha\in\mathbb{R}_+} \|\mathrm{K}u_\alpha^* - b\|_2^2 \,.$$

Letting $\mathcal{R}$ denote the generic regularization term, the L-curve is defined as the plot of the norm of the regularized solution $\|\mathcal{R}\,(u_\alpha^*)\|_2$ versus the corresponding residual norm $\|\mathrm{K}u_\alpha^* - b\|_2$. The regularization parameter is thus set as the one realizing a *corner* in the L-curve.

Both methods bring along some downsides, such as, just to name a few, the existence of a corner in the L-curve and the feasibility of the computations in the GCV, especially when dealing with large-scale problems - see [68] for a more extensive discussion - and non-quadratic regularizers. A plethora of literature has been focused on developing strategies to overcome the limitations of these classical methods [70, 128, 95, 134, 16].

As far as the methods exploiting the information available on the noise level are concerned, when the corrupting noise is Gaussian, i.e. in the variational model a ℓ_2 fidelity term is adopted, a classical approach is Morozov discrepancy principle [117]. At the core of the discrepancy principle there is the idea that, typically, no benefit is expected when the problem is solved more accurately than the accuracy in the data. Let the corrupting noise vector $e \in \mathbb{R}^n$ be a realization drawn from a random variable $E \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_n)$. Observing that

$$\mathbb{E}\left(\|e\|_2^2\right) = \mathbb{E}\left(\sum_{i=1}^n e_i^2\right) = \mathrm{trace}\left(\mathbb{E}\left(ee^T\right)\right) = \mathrm{trace}\left(\sigma^2 \mathrm{I}_n\right) = \sigma^2 n\,,$$

the noise level can be estimated as

$$\sqrt{\mathbb{E}(\|e\|_2^2)} = \sigma\sqrt{n}\,.$$

The discrepancy principle thus consists in looking for a solution u belonging to the discrepancy set, defined as

$$\mathcal{D} := \{u \in \mathbb{R}^n \mid \|\mathrm{K}u - b\|_2 < \delta = \tau\sigma\sqrt{n}\} \ ,$$

where $\sigma\sqrt{n}$ is the noise level and $\tau \approx 1$ is a parameter avoiding the under-estimation - for $\tau > 1$ - and the over-estimation - for $\tau < 1$ - of the noise. Theoretically, δ is an upper bound for the noise level, hence one should only consider $\tau > 1$. Nevertheless, it has been observed that, in some circumstances, considering τ slightly less than 1 can improve the quality of the final restoration. More results in this direction will be given in Chapter 5.

In the numerical examples of Section 3.3 we have adopted the discrepancy principle with $\tau = 1$. More specifically, the estimation of the regularization parameter has been carried out via the adaptive parameter estimation (APE) procedure proposed in [89]. Due to its wide use in the computational aspects of Part II, APE strategy will be discussed in details in the following chapters.

3.5 Analysis versus synthesis

In Chapter 1, we mentioned that we will consider an analysis approach in Part II, while in Part III a synthesis framework will be adopted. Here we present a brief outline of the two approaches; further details and comparisons can be found in [61, 71].

In general, the priors introduced above encode information related to the action of an operator, namely the finite difference operator D on the image u. Mathematically,

$$\mathrm{P}(u \mid \theta_{\mathrm{pr}}) = f(\mathrm{D}u; \theta_{\mathrm{pr}}) \,,$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a scalar function parametrically depending on θ_{pr}. According to this notation, the MAP estimation can be written as

$$(u^*, \theta^*_{\mathrm{pr}}) \in \arg \min_{(u,\theta_{\mathrm{pr}}) \in \mathbb{R}^n \times \mathbb{R}^q} \{-\log \mathrm{P}(b \mid u) - \log \mathrm{P}(\theta_{\mathrm{pr}}) - \log f(\mathrm{L}u; \theta_{\mathrm{pr}})\} \ , \qquad (3.31)$$

for the generic operator L. In particular, L is known as an analysis operator, from which the name MAP-*analysis* estimation - see [67] - to refer to the problem stated in (3.31).

Alternatively, the image u can be modeled as a combination of atom signals, i.e.

$$u = \mathrm{T}z \,, \quad \mathrm{T} \in \mathbb{R}^{n \times h}, \, z \in \mathbb{R}^h \,,$$

where T is a chosen basis and z is the vector of coefficients representing u in the basis given by the columns of T. In this case, one can exploit the information available on z. In other words, the prior takes the form

$$\mathrm{P}(u \mid \theta_{\mathrm{pr}}) = g(z; \theta_{\mathrm{pr}}) \,,$$

with $g : \mathbb{R}^h \to \mathbb{R}$. Hence, the MAP estimate, also known as MAP-*synthesis* estimate, becomes

$$(z^*, \theta^*_{\mathrm{pr}}) \in \arg \min_{(z,\theta_{\mathrm{pr}}) \in \mathbb{R}^h \times \mathbb{R}^q} \{-\log \mathrm{P}(b \mid z, \theta_{\mathrm{pr}}) - \log \mathrm{P}(\theta_{\mathrm{pr}}) - \log g(z; \theta_{\mathrm{pr}})\} \ .$$

The two approaches are equivalent when L is invertible and $\mathrm{T} = \mathrm{L}^{-1}$. A detailed study of the equivalence between MAP-analysis and MAP-synthesis in the one-dimensional case in more general settings is given in [67], where the authors also remark that neither approach is better than the other, rather they are successful on *different* sets of signals.

Chapter 4

TV and sparse regularization: a closer look

In this chapter, we will discuss the motivations at the core of our proposal, that, as already remarked in the preamble, is twofold. First, we will review some of the downsides of TV, which make necessary the introduction of novel regularizers holding the potential to distinguish between local sparsity patterns of different nature in the gradient structure. TV regularization shares its conceptual limitations with all the other regularizers existing in literature exhibiting space-invariance and *blindness* to directionality, i.e. the inability to drive the regularization along dominant orientations in the image. Hence, the strategies proposed in the next chapters could be applied to improve the performances of other regularization terms, such as higher-order regularizers, e.g. total generalized variation (TGV) [15], infimal-convolution total variation [43] or the TV-TV^2 proposed in [123].

Then, we are giving a closer look to the classical literature on sparse recovery, that is heavily based on the adoption of ℓ_1-penalty terms. We will discuss the benefits and the limitations of this approach, the latter motivating the growing interest in non-convex regularization. Moreover, the advantages of considering a Bayesian framework instead of a strictly deterministic one will be also explored.

4.1 A TV anamnesis

Since its introduction in 1992, the TVI-L$_2$ model [137], which reads

$$u^* \in \arg\min_{u\in\mathbb{R}^n} \left\{ \alpha \sum_{i=1}^{n} \|(\mathrm{D}u)_i\|_2 + \frac{1}{2}\|\mathrm{K}u - b\|_2^2 \right\}, \tag{4.1}$$

has started a new very active field of research, aimed at studying both analytical properties and computational challenges, and at characterizing its solutions. Most of the results found in the literature concern the continuous model. Consider $u \in \mathrm{L}^1(\Omega)$, with Ω regular domain in $\mathbb{R}^2$. The total variation of the function u is defined as

$$\mathrm{TV}(u) = \sup \left\{ \int_\Omega u\, \nabla \cdot v \text{ such that } v \in \mathrm{C}_c^1(\Omega, \mathbb{R}^2) \text{ and } v(x) \in B_2(0)\,, \forall x \in \Omega \right\}, \tag{4.2}$$

where $B_2(0)$ is the closed Euclidean unit ball centered at the origin [5]. In particular, for $u \in W^{1,1}(\Omega)$, definition (4.2) can be expressed as

$$\mathrm{TV}(u) = \int_\Omega \|\nabla u\|_2 \, d\Omega \,.$$

It is worth remarking here that, when $\mathrm{TV}(u) < \infty$, u is a bounded variation function, thus belonging to $\mathrm{BV}(\Omega)$ and

$$\mathrm{TV}(u) = |Du|(\Omega) \,,$$

where $|Du|$ is the total variation measure of the distributional gradient of u. The continuous TV-L$_2$ model can be written as

$$u^* \in \arg \min_{u \in \mathrm{BV}(\Omega)} \left\{ \alpha \mathrm{TV}(u) + \frac{1}{2} \|\mathrm{K}u - b\|^2_{L^2(\Omega)} \right\} \,, \tag{4.3}$$

whose well-posedness has been proved by Acar and Vogel in [2]. The TV prior is well suited to describe discontinuities in images and it is particularly effective in the case of piece-wise constant structures. In fact, in [5] the authors show that the solution $u^* \in \mathrm{BV}(\Omega)$ of problem (4.3) is a bounded variation function whose gradient is the sum of an absolutely continuous part and of a singular part, the latter allowing for jumps in the final restoration.

Moving from the continuous to the discrete formulation, a discretization strategy for the gradient must be fixed. Many efforts have been made to propose finite difference schemes ensuring that the analytical properties of the continuous model are satisfied in discrete settings also - see, e.g, [160, 42, 51]. Here, we will not comment on the advantages of such sophisticated schemes and consider the forward finite difference scheme described in (3.6), that has been used in a plethora of works since it allows to extend the adjointness of the gradient operator with the divergence operator from continuous to discrete settings. In fact, the discrete divergence operator is given by the backward finite difference scheme [41].

Let us consider some typical artifacts encountered under the adoption of a TVI regularizer. To fix the ideas, we perform the TVI denoising on a one-dimensional signal. In Figure 4.1a, a test signal presenting both piece-wise constant and linear features is shown, while the observed noisy signal is displayed in Figure 4.1b. Finally, the signal u^* found by solving (4.1) is plotted in Figure 4.1c and compared with the original one. A common side effect of TVI is the mismatching between u^* and the baseline of the original signal, which here mainly arises in the flat region; in other words, TVI often leads to a *loss of contrast* in the final restoration. Moreover, the restoration of the smooth part of the signal exhibits a *staircasing* effect, which reflects the tendency of TVI to promote blocky structures; staircasing has been widely studied, for example, in [120, 39, 135, 99].

When moving from 1D signals to images, the issues to address are the same, and an additional one arises. Consider the piece-wise constant test image `rectangles` in Figure 4.2, which has been sythetically corrupted by Gaussian blur with parameters `band`=5 and `width`=1 and AWGN with standard deviation $\sigma = 0.15$ - see Figure 4.2b. Even if the distortion induced by the blur is rather mild, the restored image in Figure 4.2c presents *rounded corners*, which is a typical drawback when using TVI for the restoration of 2D signals. The effect is clearly visible in the absolute error image shown in Figure 4.2d.

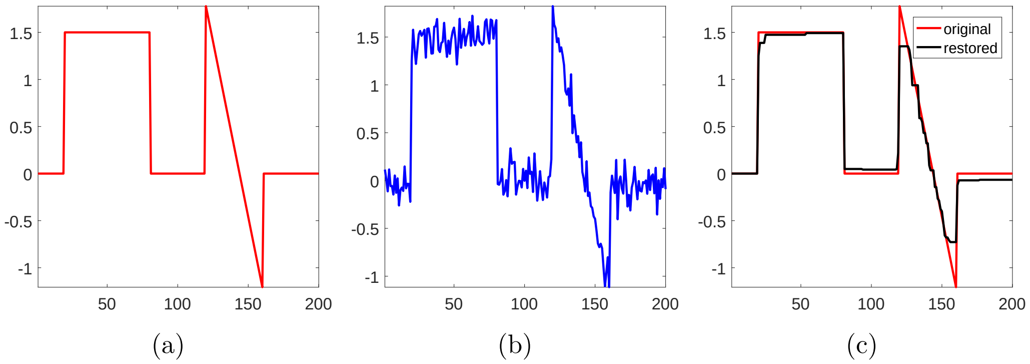


Figure 4.1: Original signal (a), noisy signal (b) and restored signal (c).

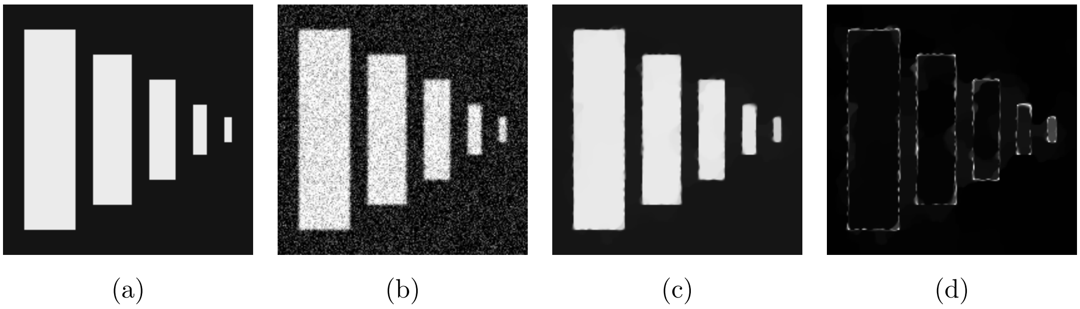


Figure 4.2: Original test image (a), observed image corrupted by Gaussian blur with `band`=5 and `width`=1 and AWGN with standard deviation $\sigma = 0.15$ (b), restoration via TVI-L$_2$ (c), absolute error (d).

In addition to the discussed unwanted effects already mentioned, the TVI regularizer suffers from additional shortcomings, including the fact that it is *global* or space-invariant, i.e. its functional shape and the local amount of regularization take the same form at each pixel, without matching local image properties and structures. Furthermore, it does not adapt easily to situations where clear local directional texture or edges may appear.

The visual and conceptual shortcomings of TVI reinforces the need for introducing regularization terms able to model in a more flexible way the image features and, consequently, to adopt a specific type of regularization for each feature. In order to give a glimpse of the possible benefits of this kind of approach, we consider the TVA regularization term and test its performance on a test image, namely `qrcode` in Figure 4.3a, presenting edges oriented along the x and y axes. As remarked in Section 3.2.2, the TVI regularizer does not present any preferred diffusion direction due to its circular level curves, while this is not the case of the TVA regularizer, from which we expect a particularly effective regularization along the axes. This property makes the TVA-L$_2$ model more suitable for the restoration of `qrcode`. Figure 4.3b shows the observed image b corrupted by Gaussian blur of `band` $= 13$ and `width = 3`, and AWGN with standard deviation $\sigma = 0.07$. We compute the solution of both models by means of the ADMM algorithm. The corresponding restorations are shown in Figure 4.3c and Figure 4.3d, respectively. Not surprisingly, the corners in the solution via TVI-L$_2$ are rounded, whereas the TVA-L$_2$ model returns an image with sharper corners and edges. The improvement achieved by considering the anisotropic model is clear even by visual inspection.

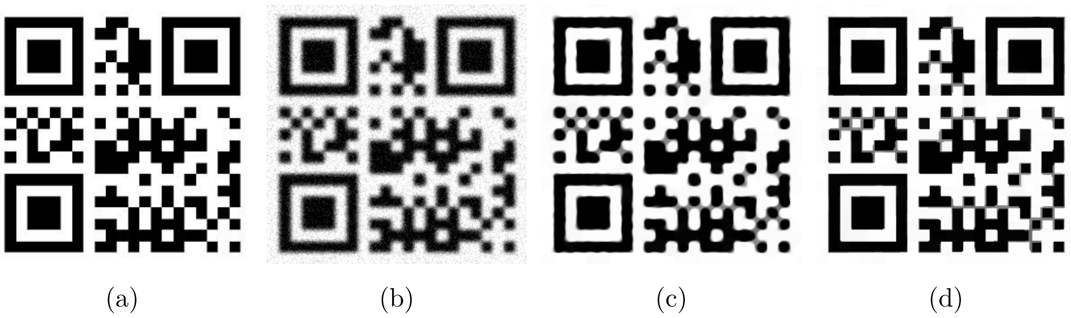


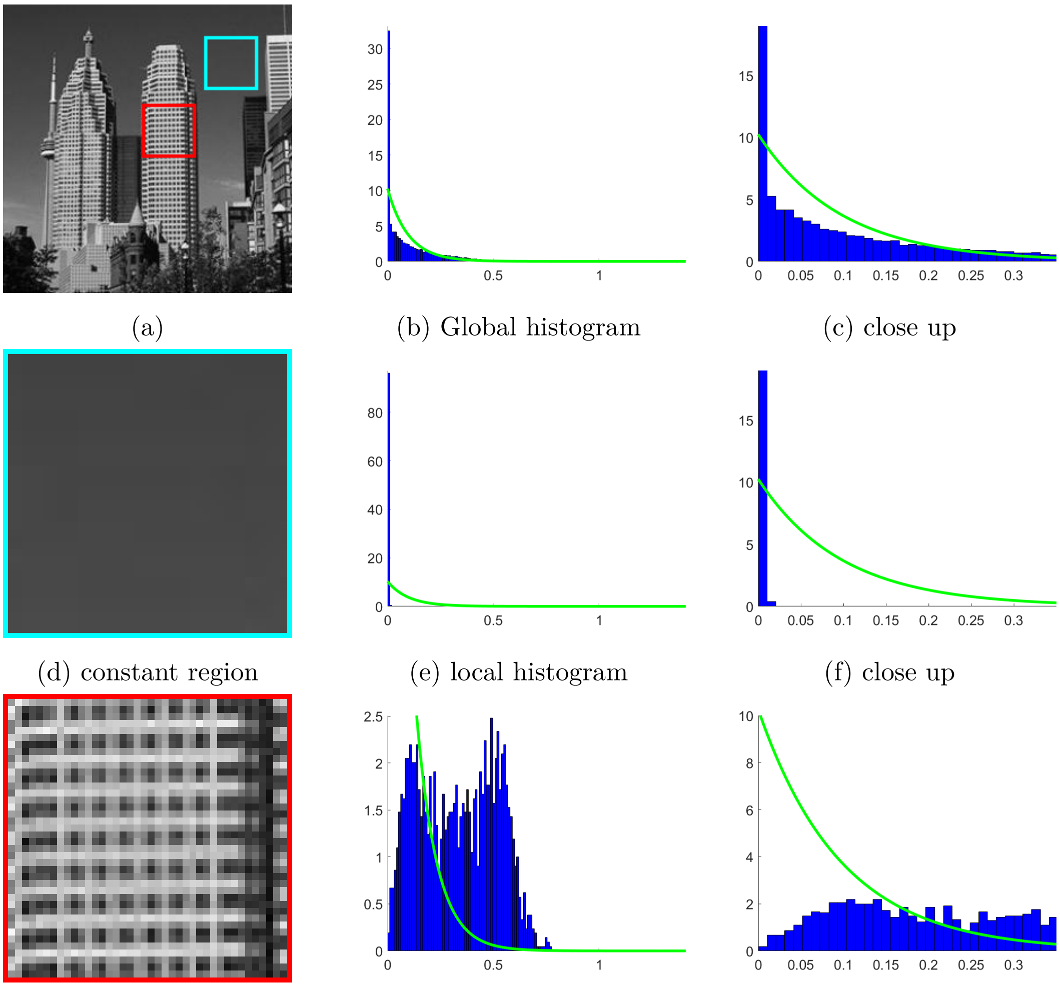
Figure 4.3: Original test image `qrcode` (a), image corrupted by Gaussian blur of parameters `band` = 13 and `width` = 3 and AWGN with standard deviation $\sigma = 0.07$ (b), restoration via TVI-L$_2$ (c), restoration via TVA-L$_2$ (d).

Nevertheless, it is worth remarking at this point that, in general, the limitations arising for the TVI regularizer, also arise for TVA. From now on, we are using the simplified acronym TV to denote the isotropic total variation regularization term.

The last example motivates the need of including directional information in the design of the regularizer. As we mentioned earlier, the global nature of TV can prevent from obtaining high quality restorations. In fact, the presence of a single parameter α in the TV-L$_2$ model produces a homogeneous regularization that is expected to favor some regions of the image while disadvantaging others. This can be clearly observed in Figure 4.4, where we compare the distribution of the ℓ_2-norm of the gradients in the whole test image `skyscraper` and in two sub-regions characterized by different properties, namely a constant region and a region presenting texture. We recall again that the adoption of the TV regularizer is equivalent to assume that the ℓ_2-norm of the gradients in the image follows a half-Laplacian (hL) distribution - see (3.18). The green solid line thus represents the half-Laplacian distribution that best fits the global histogram and it is super-imposed to the local histograms. It is clear how a global representation does not return a sufficient accurate approximation of the local distributions.

Remark 1. *In Figure 4.4, the pdfs fitting the histograms behaves as a zero-mean half Laplacian distribution. This is an admissible approximation when considering the global histogram in Figure 4.4b, since the samples are highly concentrated around zero. When going from a global to a local perspective, we do not lose any information when the selected region is a constant one - see Figure 4.4e. On the other hand, on texture regions, we do not expect the mass of the distribution to be close to zero and the approximation looks less accurate than in the previous cases. In spite of these observations, in the following we will only zero-mean distributions to model the global and local behavior of the gradients and of their magnitudes. Preliminary tests have shown that the heavy tail of the half-Laplacian distribution coupled with a non-zero mean may lead to inconsistent results. A more detailed analysis will be object of future research.*

Many contributions have been devoted to develop strategies for overcoming the limitations of TV regularization. As far as the staircasing problem is concerned, generalizations of TV based on higher order regularization have been proposed, e.g. TGV [15] and ICTV [43].
Keeping the focus on first order regularizers, in [111] the authors remark that the hL distribution, which is strictly related to the TV regularizer, is characterized by a single free-parameter, namely



(a) (b) Global histogram (c) close up

(d) constant region (e) local histogram (f) close up

(g) texture region (h) local histogram (i) close up

Figure 4.4: From top to bottom: histogram of the gradient magnitudes on the whole test image, on a constant region and on a texture region, with the corresponding close-up(s).

the scale parameter β, which, by itself, is not sufficient to model the gradient structure of a generic image. They thus assume that the ℓ_2-norm of the gradients distribute according to a two-parameter half-Generalized Gaussian (hGG) distribution, leading to the TV_p regularization term

$$\mathrm{TV}_p = \sum_{i=1}^{n} \beta \|(\mathrm{D}u)_i\|_2^p \, ,$$

where β is the scale parameter of the distribution and p is the additional *shape* parameter. Both of them can be automatically estimated [144].

When a dominant global direction is detected in the image, the *directional total variation* (DTV) is a natural choice. The discrete definition of DTV has been first proposed in [11] and subsequently in [103] it has been extended to continuous settings to generalize the definition of TV given in (4.2). Formally,

$$\mathrm{DTV}(u) = \sup \left\{ \int_\Omega u \, \nabla \cdot v \text{ such that } v \in \mathrm{C}_c^1(\Omega, \mathbb{R}^2) \text{ and } v(x) \in E^{a,\theta}(0) \, , \forall x \in \Omega \right\} ,$$

where $E^{a,\theta}(0)$ is an ellipse centered at 0 with minor semi-axes of length a and forming an angle

θ with the x-axis. In [125], the directional approach is also extended to TGV.

From the local perspective, many contributions have been proposed, especially in continuous settings. In [58], a TV-L$_2$ model equipped with space-variant regularization parameters estimated based on a local discrepancy principle, has been introduced. In the same fashion, local regularization parameters can be estimated via computationally expensive bilevel-optimization approaches, as suggested in [92, 93, 49].

Among the PDE approaches for image restoration, a reference model based on anisotropic diffusion can be formulated as the following Cauchy problem

$$\begin{cases} u_t = \operatorname{div}\left(\mathbf{W}_{\lambda,\theta}\nabla u\right) & \text{on } \Omega \times [0,\infty)\,, \\ u(x,0) = f(x) & \text{on } \Omega\,, \\ \langle \mathbf{W}_{\lambda,\theta}\nabla u, \mathbf{n}\rangle = 0 & \text{on } \partial\Omega \times [0,T) \end{cases}$$

where $\mathbf{n}$ stands for the outward normal vector on $\partial\Omega$, Ω is endowed with Neumann boundary conditions and $\mathbf{W}_{\lambda,\theta}$ is a symmetric and positive semidefinite anisotropic tensor. The tensor $\mathbf{W}_{\lambda,\theta}$ classically related to a structure-tensor modeling as in [159, 136, 140] or can stand for space-dependent diffusivity matrix which can introduce non-linearities in the model [161]. In search of additional benefits of a space-variant approach, in [48, 112] regularizers of the form

$$\mathcal{R}_{p(\cdot)}(u) := \int_\Omega \frac{1}{p(x)} |\nabla u(x)|^{p(x)}\, dx,$$

have been considered, where the exponent function $p : \Omega \to [1,2]$ is defined for every $x \in \Omega$ via the following explicit formula.

$$p(x) = 1 + \frac{1}{1 + k|G_\varsigma * \nabla g(x)|}, \quad \varsigma,\ k > 0,$$

where G_ς is a convolution kernel of parameter ς and g is the given corrupted image.

4.2 Sparse recovery: from ℓ_1 to non-convex regularization

Let $u \in \mathbb{R}^n$ be the unknown image in the restoration problem and consider the matrix $\mathrm{W} \in \mathbb{R}^{n\times n}$, whose columns form an orthonormal basis. We assume u to admit a sparse representation in the basis given by the columns of W, i.e.

$$u = \mathrm{W}\psi\,,\ \psi \in \mathbb{R}^n \text{ and } \psi \text{ is sparse}\,.$$

The sparse recovery problem is the task of restoring the signal u starting from a corrupted and possibly down-sampled observation. In the case of a linear inverse problem, it can be expressed as:

$$\text{find } u \in \mathbb{R}^n \text{ such that } b = \mathrm{A}\psi + e\,,\ \mathrm{A} = \mathrm{KW} \in \mathbb{R}^{m\times n}\,,\ m \le n$$

where A is the linear forward model operator. Moreover, in this framework, we refer to $\mathrm{K} \in \mathbb{R}^{m\times n}$, as the *measurement operator*. When W is orthogonal, the sparse recovery problem can be equivalently formulated in an analysis

$$u^* \in \arg\min_{u\in\mathbb{R}^n} \left\{ \|\mathrm{W}^{-1}u\|_0 + \mu\mathcal{F}\left(u;\mathrm{K}\right) \right\}\,, \tag{4.4}$$

or in a synthesis framework

$$\psi^* \in \arg\min_{\psi\in\mathbb{R}^n} \left\{\|\psi\|_0 + \mu\mathcal{F}(\psi;\mathrm{A})\right\} . \tag{4.5}$$

More specifically, in case of AWGN, the fidelity terms in (4.4) and (4.5) take the form

$$\mathcal{F}(u;\mathrm{K}) = \frac{1}{2}\|\mathrm{K}u - b\|_2^2 \quad \text{and} \quad \mathcal{F}(\psi;\mathrm{A}) = \frac{1}{2}\|\mathrm{A}\psi - b\|_2^2 \, ,$$

respectively. In this thesis, we will adopt a synthesis framework. In many applications, matrix W is not orthogonal and the sparsity constraint is rather expressed in terms of redundant and over-complete dictionaries, namely

$$u = \mathrm{W}\psi\, , \quad \text{with} \quad \mathrm{W}\in\mathbb{R}^{n\times N}\, ,\ \psi\in\mathbb{R}^N,\ N \gg n\, .$$

Clearly, an over-complete representation is helpful when the signal of interested can not be represented in an orthonormal basis and it also provides several benefits in terms of artifacts reduction [145, 147].

Efficient algorithms addressing the sparse recovery problem have been, and continue to be actively pursued. In fact, many problems arising in geophysics, statistics and signal processing can be formulated as (4.5) - see, e.g., [153, 154, 53, 54], and [98] for further references. In this sense, significant contributions were mainly due to the increasing interest in *compressed sensing*, which looks for the sparsest representation of an object and discards non-informative data without running into a perceptual loss [59].

A possible strategy to overcome the NP-hardness of problem (4.5) - see [118] - is to rather consider its convex relaxation

$$\psi^* \in \arg\min_{\psi\in\mathbb{R}^N} \left\{\|\psi\|_1 + \mathcal{F}(\psi;\mathrm{A})\right\} . \tag{4.6}$$

Notice that in problem (4.6), which is also known as Basis Pursuit (BP) [60], the more general over-complete formulation has been adopted. A substantial body of literature has been devoted to propose effective algorithms for the solution of (4.6) - see, e.g, [36, 72, 53, 73]. One of the reason for the popularity of the ℓ_1-penalty term is related to the following definition [37].

Definition 1. *The linear forward model operator* A *satisfies the restricted isoperimetry property (RIP) condition if, for each s-sparse signal ψ, i.e. ψ has at most s non-zero entries, the following inequalities chain holds,*

$$(1-\delta_s)\|\psi\|_2^2 \le \|\mathrm{A}\psi\|_2^2 \le (1+\delta_s)\|\psi\|_2^2\, ,$$

with $\delta_s > 0$ sufficiently small.

Among the matrices satisfying the RIP condition we mention sub-Gaussian matrices, partial bounded orthogonal matrices [47] and randomly generated circulant matrices [97]. In [35], the authors have shown that if the forward linear model operator satisfies the RIP condition, when $\|e\|_2 < \epsilon$, model (4.6) can recover an estimate ψ^* of the original representation vector which satisfies

$$\|\psi - \psi^*\|_2 \le C\left(\epsilon + \frac{\|\psi - \psi_s\|_1}{\sqrt{s}}\right)\, , \tag{4.7}$$

with ψ_s denoting the vector consisting of the s largest (in magnitude) coefficients of ψ and zeros otherwise. The error estimate in (4.7) is optimal. Moreover, when the representation vector is s-sparse and there is no noise in the measurements, model (4.6) recovers ψ exactly.

However, in most applications the RIP condition is not satisfied. As a consequence, more general ℓ_p regularization approaches, both with $1 \leq p < 2$ and $0 < p < 1$, have also been considered:

$$\psi^* \in \arg\min_{\psi \in \mathbb{R}^N} \left\{ \|\psi\|_p + \mathcal{F}(\psi; \mathrm{A}) \right\}, \, 0 < p \leq 2 \, . \tag{4.8}$$

While the regularizing properties of ℓ_p penalties, for $1 \leq p \leq 2$, are quite well understood from the theoretical and computational point of view - see e.g. [17, 81, 113, 133] - the design of efficient methods for computing the corresponding regularized solution when $p < 1$ continues to pose significant challenges and remains a very active field of research (4.6) - see e.g [45, 80, 166, 114] and [163] for a more detailed review. The challenges behind ℓ_p regularization with $p < 1$ are mostly the non-differentiability of the penalty term and the non-convexity of the functional in (4.8). A large class of works has focused on the introduction of smoothed version of the ℓ_p-norm in order to ensure the differentiability of the overall functional, such as [139, 109, 33] and the iterative reweighted algorithm in [55, 46]. Although the mentioned methods succeeded in outperforming the ℓ_1 regularization in terms of sparse recovery, the non-convexity is still an issue which is worth to be further investigated. Finally, we mention that, as highlighted in several works, some of them related to the optical imaging problem - see, e.g., [6] - the signal-to-noise-ratio (SNR), defined as

$$\mathrm{SNR} = \frac{\|b\|_2^2}{\|e\|_2^2},$$

produces a strong influence on the quality of the recovered signal. More specifically, under the same noise corruption, a less sparse signal is more likely to be poorly recovered with respect to a more sparse signal. One can easily notice that, as a matter of fact, the SNR does not play a significant role in any of the mentioned regularization schemes. In order to take the information encoded in the SNR into account in the restoration machinery, one could rather adopt a Bayesian approach. A great contribution in this direction is given in [32], where the authors present the sparse recovery as an inverse problem in the Bayesian framework, and express the sparsity assumption via a suitable hierarchical modeling.

4.3 Contribution

The main contribution of this thesis is the proposal of novel space-variant regularization or penalty terms motivated by a strong statistical rational. In light of the connection between the classical variational framework and the Bayesian formulation, we will focus on the design of highly flexible priors characterized by a large number of unknown parameters. The latter will be *automatically* estimated by setting up a hierarchical modeling, i.e. introducing informative or non-informative hyperpriors depending on the information at hand on the parameters.

From Chapter 5 to Chapter 7, the problem of restoring natural images will be addressed. Starting from the half-Laplacian distribution modeling the behavior of the gradient magnitudes and corresponding to the TV prior, we will consider more highly parameterized distributions, modeling the local behavior of the gradients and performing a pixel-wise tuning of the strength, of the type and of the orientation of the regularization. We will highlight the contribution of each space-variance, by detecting the one that eventually returns the more decisive improvement in terms of the quality of the final restoration.

As far as the sparse recovery problem is concerned, in Chapter 8 we will first extend the study proposed in [32] to a wider class of hyperpriors modeling the sparsity of the signal of reference. In this context, the space-variance follows directly from the independence of the components in the signal, each following its own parametrized distribution. The results on the analytical properties of the energy functionals corresponding to the class of hyperpriors considered, will motivate the introduction, in Chapter 9, of two hybrid algorithms. The first one is strictly related to the definition of a novel convex penalty term holding the potential to outperform the classic ℓ_1-penalty term in terms of sparsity promotion. The second scheme addresses a possibly non-convex minimization problem, exploiting the benefits of a more suitably designed initial guess. Finally, in Chapter 10, we will show how the outlined Bayesian framework can be applied to the recovery of signals sparsely represented in an over-complete basis. This last contribution is also aimed at providing a further interpretation of *space-variance* in a synthesis framework. In fact, we will highlight how different local features of an image can be naturally and sparsely represented by different bases.

Part II

Non-informative hyperpriors

Chapter 5

Space-variant strength regularization tuning

In this chapter, we start our investigation aiming to design more and more flexible priors that, in this part of the thesis, will be coupled with non-informative hyperpriors. The automatic estimation of the parameters involved in the proposed priors will also be a matter of discussion.

We recall the discrete linear degradation model for the image restoration problem,

$$b = \mathrm{K}u + e\,, \quad E \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_n)\,, \tag{5.1}$$

where $\mathrm{K} \in \mathbb{R}^{n\times n}$ is the blur matrix and $u, b, e \in \mathbb{R}^n$ are respectively the vectorized unknown image, observed image and corrupting additive noise, the latter being drawn from a zero-mean white Gaussian distribution with standard deviation σ. We have already remarked the equivalence between

$$u^* \in \arg\min_{u\in\mathbb{R}^n} \left\{ \alpha \sum_{i=1}^{n} \|(\mathrm{D}u)_i\|_2 + \frac{1}{2}\|\mathrm{K}u - b\|_2^2 \right\} \tag{5.2}$$

and

$$u^* \in \arg\min_{u\in\mathbb{R}^n} \left\{ \sum_{i=1}^{n} \|(\mathrm{D}u)_i\|_2 + \frac{\mu}{2}\|\mathrm{K}u - b\|_2^2 \right\}\,. \tag{5.3}$$

We will indistinctly refer to μ and to α as *regularization parameter*. In fact, for small μ, or equivalently, for large α, edges tend to be preserved at the expense of the smoothing of the noise. On the other hand, the larger the μ, or the smaller the α, the stronger is the effort of the regularization in removing the noise. As already discussed in Section 3.4, the setting of the regularization parameter is a very delicate issue. In literature, several strategies have been considered. When the noise level is known, a classical approach is based on the use of the discrepancy principle [68], or of its adaptive version [89]. Besides the already mentioned *a posteriori* estimation criteria, such as GCV and L-curve, more recently, in blind scenarios, optimization techniques learning the optimal amount of regularization from training data have also been used - see, e.g., [121, 94, 18].

Nonetheless, the presence in model (5.2) of the global parameter α prevents from tuning the strength of the regularization on images presenting both fine-scale details, as textures, and piece-wise constant or smooth regions, whereas the global parameter μ in model (5.3) is not suited for non-homogeneous noise corruption. To overcome these two issues, weighted versions of the TV-L$_2$ model have been proposed. In continuous settings they read

$$u^* \in \arg \min_{u \in \mathrm{BV}(\Omega)} \left\{ \int_\Omega \alpha(x)|\nabla u| + \frac{1}{2} \int_\Omega |\mathrm{K}u - b|^2 dx \right\}, \tag{5.4}$$

with $\alpha \in C(\bar{\Omega})$, and

$$u^* \in \arg \min_{u \in \mathrm{BV}(\Omega)} \left\{ \int_\Omega |\nabla u| + \frac{1}{2} \int_\Omega \mu(x)|\mathrm{K}u - b|^2 dx \right\}, \tag{5.5}$$

with $\mu \in L^\infty(\Omega)$, $\mu \geq 0$. Here, α and μ are two weight functions determining the local strength of the regularization and of the fidelity term, respectively. The estimation of the α weight function in (5.4) has been addressed by inferring local geometries [149] or by means of computationally expensive bilevel-optimization approaches [92, 93], whereas in [58] the weight function μ in (5.5) of the *spatially adapted total variation* (SATV) have been estimated based on the use of a local discrepancy principle. We remark that the two locally-weighted (5.4) and (5.5) models do show significant differences when used for image reconstruction problems, as it has been rigorously studied in [91] in continuous settings.

In the following, we will address the problem of tuning the strength of the regularization in a purely discrete setting. We thus introduce the weighted versions of the discrete models (5.2) and (5.3), which read

$$\min_{u \in \mathbb{R}^n} \left\{ \mathrm{WTV}(u) + \mathrm{L}_2(u) \right\}, \qquad \mathrm{WTV}(u) := \sum_{i=1}^n \alpha_i \|(\mathrm{D}u)_i\|_2, \qquad \alpha_i > 0\,,\ i = 1, \ldots, n\,, \tag{5.6}$$

$$\min_{u \in \mathbb{R}^n} \left\{ \mathrm{TV}(u) + \mathrm{WL}_2(u) \right\}, \qquad \mathrm{WL}_2(u) := \frac{1}{2} \sum_{i=1}^n \mu_i (\mathrm{K}u - b)_i^2, \qquad \mu_i > 0\,,\ i = 1, \ldots, n\,. \tag{5.7}$$

5.1 The HWTV-L$_2$ model

Our first contribution in a space-variant perspective has been introduced in [19] and consists of a hybrid version of the two space-variant variational models (5.6) and (5.7) variational models, with variable regularization parameters α_i and global fidelity parameter μ, referred to as HWTV-L$_2$:

$$u^* \in \arg \min_{u \in \mathbb{R}^n} \left\{ \mathcal{J}(u) := \mathrm{WTV}(u) + \frac{\mu}{2} \|\mathrm{K}u - b\|_2^2 \right\}, \tag{5.8}$$

where HWTV stands for ‘hybrid weighted total variation’. We remark that the local parameters α_i describe local image scales in a statistical sense, as it will be explained in detail in the following, while the global parameter μ codifies the discrepancy with respect to the given AWGN level. The redundancy of such parameter is therefore only apparent in (5.8) as its value is computed depending on the global noise statistics in comparison with the local regularization strength encoded by the parameters α_i.

5.2 Model derivation via MAP

As already stated in (5.1), we assume the corrupting noise to be AWG with standard deviation $\sigma > 0$. Hence, the expression for the likelihood term is given by

$$\mathrm{P}(b \mid u) = \frac{1}{W} \exp\left(-\frac{1}{2\sigma^2} \|\mathrm{K}u - b\|_2^2 \right) .$$

Notice that, since $\theta_{\mathrm{lkh}} = \sigma$ is supposed to be known, it has been omitted in the likelihood term. For what concerns the prior, we recall that the underlying assumption is that the unknown image u behaves as a Markov Random Field, thus admitting the following general expression,

$$\mathrm{P}(u \mid \theta_{\mathrm{pr}}) = \frac{1}{Z} \exp\left(-\sum_{i=1}^{n} V_{\mathcal{C}_i^r}(u; \theta_{\mathrm{pr}}) \right) , \tag{5.9}$$

with $V_{\mathcal{C}_i^r}$ being the i-th potential function defining on the clique $\mathcal{C}_i^r$ of radius r centered at pixel i - see Section 3.2.2. Here, we consider the case in which the analytic form of the potential functions is the same for any pixel i, whereas the set of parameters defining each $V_{\mathcal{C}_i^r}$ changes as the current pixel changes. In other words, (5.9) turns into

$$\mathrm{P}(u \mid \theta_{\mathrm{pr}}) = \frac{1}{Z} \exp\left(-\sum_{i=1}^{n} V_{\mathcal{C}_i^r}(u; (\theta_{\mathrm{pr}})_i) \right) . \tag{5.10}$$

The adoption of a Gibbs prior of the form (5.10) implicitly states that we are now dealing with a *non-stationary* Markov Random Field. Before going on with these statistical considerations, let us go back for a moment to the expression of the TV prior,

$$\mathrm{P}(u \mid \theta_{\mathrm{pr}}) = \frac{1}{Z} \exp\left(-\alpha \sum_{i=1}^{n} \|(\mathrm{D}u)_i\|_2 \right) = \frac{1}{Z} \exp(-\alpha\, \mathrm{TV}(u)) , \quad \text{with } \theta_{\mathrm{pr}} = \alpha > 0 .$$

In Section 3.2.2, we have already pointed out that the adoption of the TV is equivalent to assume that the ℓ_2-norm of the image gradients distribute according to a half-Laplacian or exponential distribution

$$\mathrm{P}\left(\|(\mathrm{D}u)_i\|_2 \mid \alpha\right) = \begin{cases} \alpha \exp\left(-\alpha \|(\mathrm{D}u)_i\|_2\right) & \text{if } \|(\mathrm{D}u)_i\|_2 > 0 \\ 0 & \text{otherwise} \end{cases} ,$$

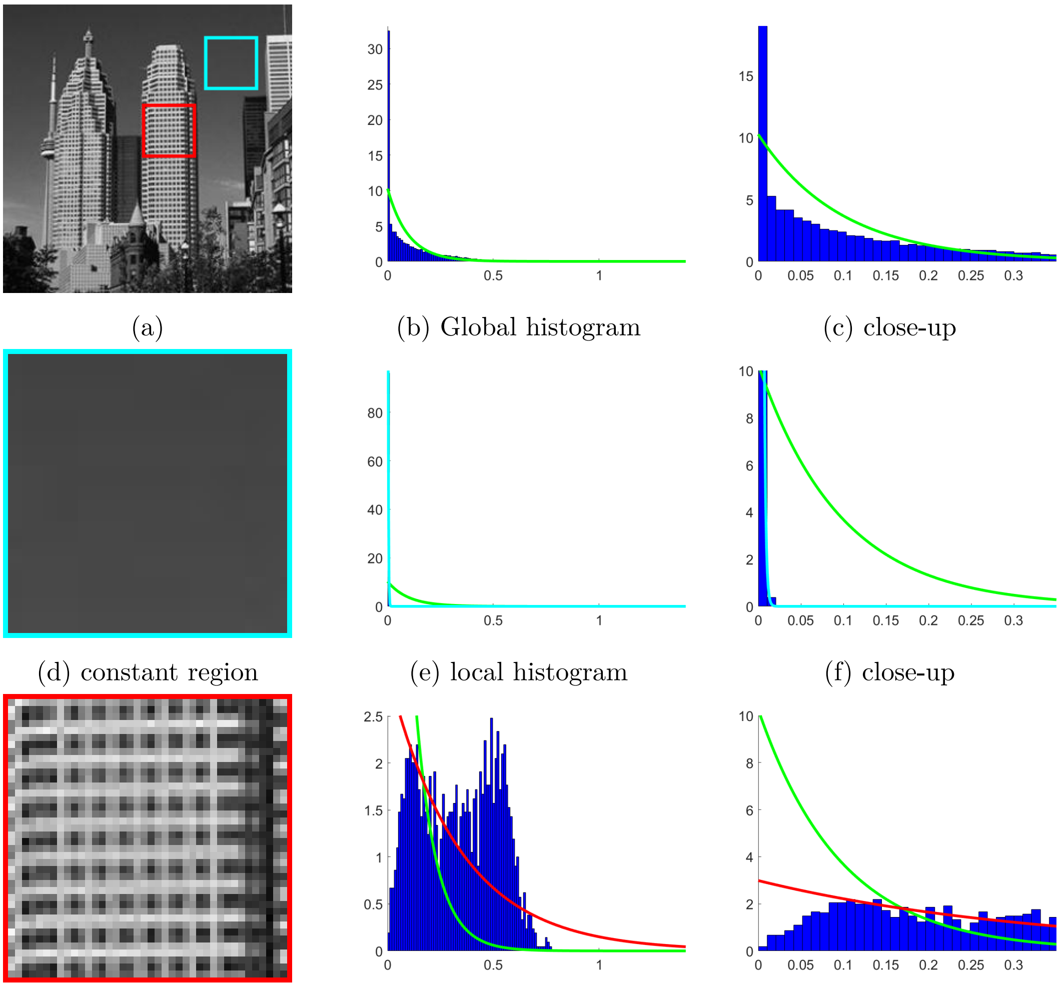
where $\alpha > 0$ is the *scale* parameter of the distribution.

To allow more flexibility, we propose a space-variant model where gradient norms distribute according to a half-Laplacian distribution with *locally varying* scale parameter $\alpha_i > 0$. The prior associated to such choice is then

$$\mathrm{P}(u \mid \theta_{\mathrm{pr}}) = \frac{1}{Z} \exp\left(-\sum_{i=1}^{n} \alpha_i \|(\mathrm{D}u)_i\|_2 \right) = \frac{1}{Z} \exp\left(-\mathrm{WTV}(u)\right) , \tag{5.11}$$

with $Z = \left(\prod_{i=1}^{n} \alpha_i \right)^{-1}$. The WTV term in (5.11) is the regularizer defined in (5.6) and the entries of $\theta_{\mathrm{pr}} \in \mathbb{R}_+^n$ are the scale parameters of the local distribution, i.e.,

$$\theta_{\mathrm{pr}} = \alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} \in \mathbb{R}_+^n .$$



(a) (b) Global histogram (c) close-up

(d) constant region (e) local histogram (f) close-up

(g) texture region (h) local histogram (i) close-up

Figure 5.1: From top to bottom: histogram of the gradient magnitudes on the whole test image, on a constant region and on a texture region, with the corresponding close-up(s).

The introduction of the WTV prior in (5.11) is motivated by the behavior of the ℓ_2-norm of the gradients in the whole image and in local regions, which shows significant different properties. In Figure 5.1b, the green solid line represents the pdf of the half-Laplacian distribution that best fits the histogram of the gradient of the image. Then, two small regions of the image have been selected, namely an almost constant region, in the red box in Figure 5.1d, and a region characterized by texture, in the cyan box in Figure 5.1g. The pdf returning the best approximation of the magnitude of the gradients in the two sub-regions, namely the cyan solid line for the constant region and the red line for the textured region, have thus been computed and compared with the global green pdf, which has been super-imposed to the local histograms. In Figures 5.1e-5.1h, and in the respective close-ups in Figure 5.1f-5.1i, it is easy to appreciate the potential of a space-variant approach in terms of a more faithful and accurate description of the local image features.

Once that the likelihood and, more importantly, the prior term have been set, they can be plugged in the MAP estimation formula for u, thus leading to the HWTV-L$_2$ model

$$u^* \in \arg\min_{u\in\mathbb{R}^n} \left\{ \mathcal{J}(u) := \sum_{i=1}^{n} \alpha_i \|(\nabla u)_i\|_2 + \frac{\mu}{2}\|\mathrm{K}u - b\|_2^2 \right\} ,$$

with $\mu = 1/\sigma^2$ being the global regularization parameter. The above minimization problem can be solved by means of the iterative scheme already introduced in Chapter 3 and based on the alternating update of the image u and of the vector of unknown parameters $\theta_{\mathrm{pr}} = \alpha$, in presence of non-informative hyperprior. We can also refer to this procedure, outlined in Algorithm 1 in the case of non-informative hyperprior, as *outer scheme.*

Algorithm 1: IAS with non-informative hypeprior for the HWTV-L$_2$ model

```
input: observed image $b \in \mathbb{R}^n$
output: restored image $u^*$
initialize: set $u^{(0)} = b$
for $t = 1, 2, \ldots$ until convergence do:
    · update $\alpha^{(t)} \in \arg\min_{\alpha\in\mathbb{R}^n_+} \left\{ -\log \mathrm{P}(u^{(t-1)} \mid \alpha) \right\}$
    · update $u^{(t)} \in \arg\min_{u\in\mathbb{R}^n} \left\{ -\log \mathrm{P}(u \mid \alpha^{(t)}) - \log \mathrm{P}(b \mid u, \alpha^{(t)}) \right\}$
end for
return: $u^* = u^{(t)}$
```

In the next sections, we will detail how the two update steps in the IAS Algorithm 1 are performed.

5.3 Parameter estimation via non-informative hyperpriors

In order for the space-variant approach to be effective, an automatic way of estimating the parameters defining the prior must be introduced as well. We remark that for the proposed model (5.2) the number of unknown parameters equals the number of pixels. Thus, our focus is first set on the solution of the α-update:

$$\alpha^{(t)} \in \arg\min_{\alpha\in\mathbb{R}^n_+} \left\{ -\log \mathrm{P}(u^{(t-1)} \mid \alpha) \right\} . \tag{5.12}$$

A key observation here is that, in order to favor the space-variant nature of the prior and, at the same time, to avoid the destroying action of the noise and blur corruption, the estimation of the entries of α must be carried out by exploiting local information.
In fact, considering the expression for the prior given in (5.11), problem (5.12) can be reformu-

lated as

$$\begin{aligned}\alpha^{(t)} &\in \arg\min_{\alpha\in\mathbb{R}_+^n} \left\{ -\log\prod_{i=1}^{n}\alpha_i + \sum_{i=1}^{n}\alpha_i\|\left(\mathrm{D}u^{(t-1)}\right)_i\|_2 \right\} \\ &= \arg\min_{\alpha\in\mathbb{R}_+^n} \left\{ \sum_{i=1}^{n}\left(-\log\alpha_i + \alpha_i\|\left(\mathrm{D}u^{(t-1)}\right)_i\|_2\right) \right\} .\end{aligned}$$

Hence, problem (5.12), can be decomposed in n one-dimensional minimization problem of the form

$$\alpha_i^{(t)} \in \arg\min_{\alpha_i\in\mathbb{R}_+} \left\{ f(\alpha_i) := -\log\alpha_i + \alpha_i\|\left(\mathrm{D}u^{(t-1)}\right)_i\|_2 \right\} . \tag{5.13}$$

The following result holds true.

Proposition 1. *The function $f : \mathbb{R}_+ \to \mathbb{R}$ in (5.13) is continuous and convex, hence it admits a unique global minimizer.*

In particular, since f is differentiable in $\mathbb{R}_+$, the solution of the i-th minimization problem (5.13) can be found by imposing a first order optimality condition:

$$f'(\alpha_i) = -\frac{1}{\alpha_i} + \|\left(\mathrm{D}u^{(t-1)}\right)_i\|_2 , \quad \text{hence} \quad \alpha_i^{(t)} = \frac{1}{\|(\mathrm{D}u^{(t-1)})_i\|_2} . \tag{5.14}$$

The update in (5.14) for the i-th scale parameter only involves the i-th pixel. As a consequence, if the information at pixel i is damaged by the action of noise and blur, the low degree of confidence in the estimate of the local scale parameters does not justify their use.

In order to overcome this issue, an ensemble of pixels close to pixel i is involved in the update of α_i. More in details, for any $i = 1, \ldots, n$, we consider the set $\mathcal{S}_i := \{x_{i,j}\}_{j=1}^N$, with $x_{i,j} = \|(\mathrm{D}u^{(t-1)})_j\|_2$. The gradients $(\mathrm{D}u^{(t-1)})_j$ are computed in the pixels belonging to the square neighborhood $\mathcal{C}_i^r$ centered at pixel i with side $2r+1$. We recall that the norm of the generic gradient of pixels in $\mathcal{C}_i^r$ is here assumed to be drawn from a half-Laplacian distribution with scale parameter α_i. Therefore, we are interested in solving n minimization problems of the form

$$\alpha_i^{(t)} \in \arg\min_{\alpha_i\in\mathbb{R}^+} \left\{ -\log \mathrm{P}(\mathcal{S}_i \mid \alpha_i) \right\} , \tag{5.15}$$

where the i-th conditioned pdf $\mathrm{P}\left(\mathcal{S}_i \mid \alpha_i\right)$ reads as

$$\mathrm{P}(\mathcal{S} \mid \alpha_i) \;=\; \prod_{x_{i,j}\in\mathcal{S}_i} \mathrm{P}(x_{i,j} \mid \alpha_i) = \prod_{j=1}^{N} \mathrm{P}(x_{i,j} \mid \alpha_i) \;=\; \alpha_i^N \exp\left(-\sum_{j=1}^{N}\alpha_i x_{i,j} \right). \tag{5.16}$$

First note that the existence and uniqueness result in Proposition 1 still holds. Moreover, observe that in order to allow the factorization of the pdf in (5.16), we need to assume the independence of the samples in $\mathcal{S}_i$. The minimization of $-\log\mathrm{P}\left(\mathcal{S}_i \mid \alpha_i\right)$ in (5.15) is equivalent to the maximization of $\mathrm{P}(\mathcal{S}_i \mid \alpha_i)$, that is also known as *likelihood* function. In fact, the hierarchical approach via non-informative hyperpriors is equivalent to the classical *maximum likelihood* (ML) procedure. We recall that the ML returns the value for α_i that most likely produces the observed samples $x_{i,j}$. From (5.16), we have

$$-\log\mathrm{P}(\mathcal{S}_i \mid \alpha_i) = -N\log\alpha_i + \sum_{j=1}^{N}\alpha_i x_{i,j} .$$

By imposing a first order optimality condition on the objective function in problem (5.15) with respect to α_i, we obtain the closed formula

$$\alpha_i = \left(\frac{1}{N} \sum_{j=1}^{N} x_{i,j} \right)^{-1} , \tag{5.17}$$

which can be handily updated along the iterations $t \geq 0$ to estimate the local regularization parameters $\alpha_i^{(t)}$ at each pixel $i = 1, \ldots, n$ by taking as samples $x_{i,j} = \|(\mathrm{D}u^{(t-1)})_j\|_2$, $j = 1, \ldots, N$ i.e. the norms of the image gradients in the neighborhood $\mathcal{C}_i^r$. Note that, in order to avoid degenerate configurations in the case of a neighborhood with null gradients, a small $\epsilon > 0$ is added to the local means in (5.17). We also remark that the selection of the pixels involved in the estimates $\alpha_i^{(t)}$ in (5.17) can be efficiently carried out based on 2D convolution (realized by a fast 2D discrete transform) of the map of gradient norms with a square $(2r+1) \times (2r+1)$ averaging kernel.

In Figure 5.2, the α-map corresponding to different test images are shown. As expected, the scale parameters assume higher values on smooth or piece-wise constant regions, whereas lower values are obtained in correspondence of edges and texture. In those areas, a weaker regularization is preferable in order to preserve details. Note also that α-maps are sensitive to the choice of radius r. When considering small values of r - see, for instance, Figure 5.2f - possibly small artifacts due to image compression or resolution may appear. We do expect the same effect in presence of noise. On the other hand, setting a large radius r, could make some details or finer structures in the image less detectable, as in the case of Figure 5.2d, where some inner edges are not visible in the map. A future work could certainly be focused on the design of an automatic procedure for the selection of r, that, at the moment, in our experiments is hand-tuned.

5.4 Existence and uniqueness of solutions

In this section, we provide an existence and uniqueness result for the solution of the proposed discrete HWTV-L$_2$ variational model (5.8). We have that the following Proposition holds true.

Proposition 2. *The HWTV-*L$_2$ *functional* $\mathcal{J} : \mathbb{R}^n \to \mathbb{R}$ *defined in* (5.8) *is continuous, bounded from below by zero and strongly convex, hence it admits a unique global minimizer.*

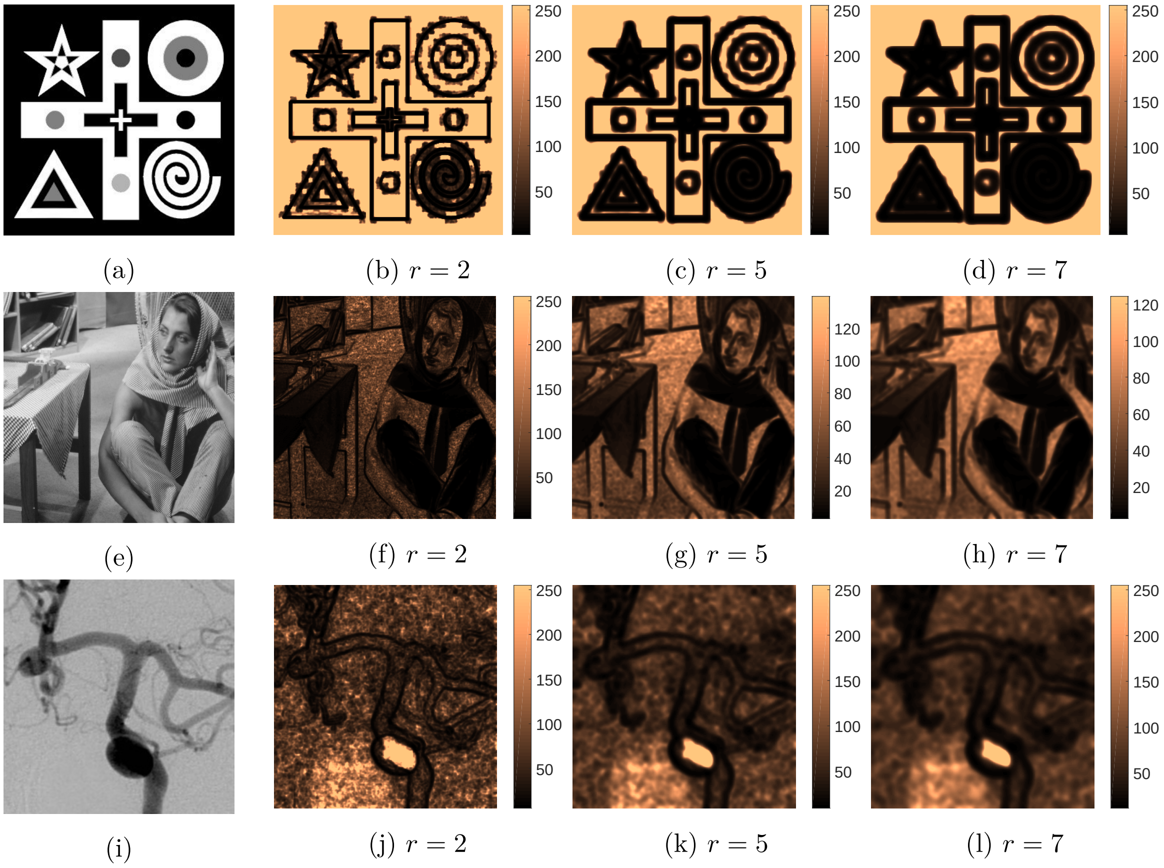
Proof. The proof comes from considering that the fidelity term, under the assumption of analytically non-singular blur matrix K, is strongly convex. □

5.5 ADMM optimization

Once that an existence and uniqueness result for the u-update has been given, we now focus on how to solve it.

The u-update step in Algorithm 1 is equivalent to the minimization problem in (5.8) reported below,

$$u^* \in \arg\min_{u \in \mathbb{R}^n} \left\{ \sum_{i=1}^{n} \alpha_i \|(\mathrm{D}u)_i\|_2 + \frac{\mu}{2} \|\mathrm{K}u - b\|_2^2 \right\} , \tag{5.18}$$



(a) (b) $r = 2$ (c) $r = 5$ (d) $r = 7$

(e) (f) $r = 2$ (g) $r = 5$ (h) $r = 7$

(i) (j) $r = 2$ (k) $r = 5$ (l) $r = 7$

Figure 5.2: Test images with the corresponding map of scale parameters α for different values of radius r.

where the entries α_i of vector α are completely determined by the closed formula in (5.17). Resorting to the ADMM algorithm, we start writing problem (5.18) in its equivalent constrained form

$$\{u^*, w^*, v^*\} \in \arg\min_{u,w,v} \left\{ \sum_{i=1}^{n} \alpha_i \, \|v_i\|_2 + \frac{\mu}{2} \, \|w\|_2^2 \right\} \qquad (5.19)$$
$$\text{subject to} \quad w = \mathrm{K}u - b, \;\; v = \mathrm{D}u.$$

with $w \in \mathbb{R}^n$ and $v \in \mathbb{R}^{2n}$ being two auxiliary variables that are introduced to transfer the discrete gradient operators $(\mathrm{D}\,\cdot)_i$ and the ill-conditioned blur operator $\mathrm{K}\,\cdot$ out of the non-smooth regularization terms $\|\cdot\|_2$ and fidelity term $\|\cdot\|_2^2$, respectively. We define the augmented Lagrangian functional:

$$\begin{aligned}\mathcal{L}(u, w, v; \rho_w, \rho_v; \alpha, \mu) := \sum_{i=1}^{n} \alpha_i \, \|v_i\|_2 + \frac{\mu}{2}\|w\|_2^2 - \rho_v^T(v - \mathrm{D}u) + \frac{\gamma_v}{2}\|v - \mathrm{D}u\|_2^2 \\ - \rho_w^T(w - (\mathrm{K}u - b)) + \frac{\gamma_w}{2}\|w - (\mathrm{K}u - b)\|_2^2, \qquad (5.20)\end{aligned}$$

where $\gamma_w, \gamma_v > 0$ are scalar penalty parameters and $\rho_w \in \mathbb{R}^n$, $\rho_v \in \mathbb{R}^{2n}$ are the vectors of Lagrange multipliers. The solution (u^*, w^*, v^*) of problem (5.19) is a saddle point for $\mathcal{L}$ in

(5.20). Hence, the original problem can be recast as follows:

$$\begin{aligned} &\text{Find } (x^*;\rho^*) \;\in\; X \times R \\ &\text{such that } \mathcal{L}(x^*;\rho;\alpha,\mu) \;\leq\; \mathcal{L}(x^*;\rho^*;\alpha,\mu) \;\leq\; \mathcal{L}(x;\rho^*;\alpha,\mu) \quad \forall\,(x;\rho) \;\in\; X \times R\,, \end{aligned} \tag{5.21}$$

where, for simplicity of notations, we set $x := (u, w, v)$, $\rho := (\rho_w, \rho_v)$, $X := \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^{2n}$ and $R := \mathbb{R}^n \times \mathbb{R}^{2n}$. Given an initialization for the vectors $u^{(0)}$, $\rho_w^{(0)}$ and $\rho_v^{(0)}$, the t-th iteration of the proposed ADMM-based iterative scheme applied to the solution of the saddle-point problem (5.21) - minimization for the primal variables u, w, v, maximization for the dual variables ρ_w, ρ_v - reads as follows:

$$v^{(t)} \in \;\arg\min_{v\in\mathbb{R}^n}\; \mathcal{L}(u^{(t-1)}, w^{(t-1)}, v; \rho_w^{(t-1)}, \rho_v^{(t-1)}; \alpha, \mu)\,, \tag{5.22}$$

$$w^{(t)} \in \;\arg\min_{w\in\mathbb{R}^{2n}}\; \mathcal{L}(u^{(t-1)}, w, v^{(t)}; \rho_w^{(t-1)}, \rho_v^{(t-1)}; \alpha, \mu)\,, \tag{5.23}$$

$$u^{(t)} \in \;\arg\min_{u\in\mathbb{R}^n}\; \mathcal{L}(u, w^{(t)}, v^{(t)}; \rho_w^{(t-1)}, \rho_v^{(t-1)}; \alpha, \mu)\,, \tag{5.24}$$

$$\rho_w^{(t)} \in \;\rho_w^{(t-1)} - \gamma_w\left(w^{(t)} - (\mathrm{K}u^{(t)} - b)\right), \tag{5.25}$$

$$\rho_v^{(t)} \in \;\rho_v^{(t-1)} - \gamma_v\left(v^{(t)} - \mathrm{D}u^{(t)}\right). \tag{5.26}$$

Hence, we can alternate a minimization step with respect to the primal variables w, v, u with a maximization step with respect to the dual variables ρ_v, ρ_w, in combination with an iterative update of the space variant entries of α and μ, which hence will be denoted by $\alpha^{(t)}$ and $\mu^{(t)}$. In particular, for what concerns $\alpha^{(t)}$ we use the easy estimation strategy described in Section 5.3, whereas for $\mu^{(t)}$ we will rely on a global discrepancy principle.

5.5.1 Primal variables update.

The three primal sub-problems can be solved efficiently and in closed-form by simple projection operators and linear system solvers. We are giving more details below.

Sub-problem for v Given the definition of the augmented Lagrangian functional in (5.20), the minimization sub-problem for the primal variable v in (5.23) can be written as follows:

$$\begin{aligned} v^{(t)} &\in \arg\min_{v\in\mathbb{R}^{2n}} \left\{ \sum_{i=1}^{n} \alpha_i \left\|v_i\right\|_2 - \langle \rho_v^{(t-1)}, v - \mathrm{D}u^{(t-1)} \rangle + \frac{\gamma_v}{2}\left\| v - \mathrm{D}u^{(t-1)} \right\|_2^2 \right\} \\ &= \arg\min_{v\in\mathbb{R}^{2n}} \left\{ \sum_{i=1}^{n} \alpha_i \left\|v_i\right\|_2 + \frac{\gamma_v}{2}\left\| v - \left(\mathrm{D}u^{(t-1)} + \frac{1}{\gamma_v}\rho_v^{(t-1)}\right) \right\|_2^2 \right\} \\ &= \arg\min_{v\in\mathbb{R}^{2n}} \sum_{i=1}^{n} \left\{ \alpha_i \left\|v_i\right\|_2 + \frac{\gamma_v}{2}\left\| v_i - \left(\left(\mathrm{D}u^{(t-1)}\right)_i + \frac{1}{\gamma_v}\left(\rho_v^{(t-1)}\right)_i\right) \right\|_2^2 \right\}. \end{aligned} \tag{5.27}$$

Note that in (5.27) the minimized functional is written in explicit component-wise (or pixel-wise) form, with $\left(\mathrm{D}u^{(t-1)}\right)_i, \left(\rho_v^{(t-1)}\right)_i \in \mathbb{R}^2$ denoting the discrete gradient and the Lagrange multipliers at pixel i, respectively. Solving the $2n$-dimensional minimization problem in (5.27) is thus equivalent to solve the n following independent 2-dimensional problems:

$$v_i^{(t)} \in \arg\min_{v_i\in\mathbb{R}^2} \left\{ \left\|v_i\right\|_2 + \frac{(\gamma_v/\alpha_i)}{2}\left\| v_i - q_i^{(t-1)} \right\|_2^2 \right\}, \quad i = 1, \ldots, n\,, \tag{5.28}$$

with the constant vectors $q_i^{(t-1)} \in \mathbb{R}^2$ defined by

$$q_i^{(t-1)} \;:=\; \left(\mathrm{D}u^{(t-1)}\right)_i + \frac{1}{\gamma_v}\left(\rho_v^{(t-1)}\right)_i \;\;, \quad i = 1, \ldots, n \,.$$

Moreover, it holds,

$$v_i^{(t)} = \mathrm{prox}_{\frac{\alpha_i}{\gamma_v}\|\cdot\|_2}\left(q_i^{(t-1)}\right) ,$$

with 'prox' being the proximity operator defined as follows [124],

Definition 2. *Let $f : \mathbb{R} \to]-\infty, +\infty]$ be a lower semi-continuous convex function. For every $x \in \mathbb{R}$*

$$\mathrm{prox}_f(x) := \arg\min_{y\in\mathbb{R}} \left\{ f(y) + \frac{1}{2}\|x - y\|_2^2 \right\} .$$

We refer to the operator

$$\mathrm{prox}_f : \mathbb{R} \to \mathbb{R} ,$$

as the proximity operator *of f.*

Hence, the solution of each one-dimensional separable problem is given by

$$v_i^{(t)} = q_i \max\left(1 - \frac{\alpha_i^{(t)}}{\gamma_v \|q_i^{(t-1)}\|_2}, 0\right) , \quad i = 1, \ldots, n. \tag{5.29}$$

The overall computational cost of this subproblem is linear in the number of pixels n.

Sub-problem for w Recalling the definition of the augmented Lagrangian functional in (5.20) and carrying out some simple algebraic manipulations, the minimization sub-problem (5.22) for the primal variable w can be written as

$$w^{(t)} \in \arg\min_{w\in\mathbb{R}^n} \left\{ \frac{\mu}{2}\|w\|_2^2 + \frac{\gamma_w}{2}\left\|w - z^{(t-1)}\right\|_2^2 \right\} , \tag{5.30}$$

with the constant (with respect to the optimization variable w) vector $z^{(t-1)} \in \mathbb{R}^n$ given by

$$z^{(t-1)} \;=\; \mathrm{K}u^{(t-1)} \,-\, b \,+\, \frac{1}{\gamma_w}\,\rho_w^{(t-1)} \,. \tag{5.31}$$

Since $\mu \geq 0$, $\gamma_w > 0$, the cost function in (5.30) is strongly convex, hence it admits a unique global minimizer. In particular, the unique solution $w^{(t)}$ of (5.30) can be computed, depending on z, by means of the following closed-form formula:

$$w^{(t)} = \left(\frac{\gamma_w}{\gamma_w + \mu}\right) z^{(t-1)} , \tag{5.32}$$

which depends on the previous update of z.

Sub-problem for u For the solution of (5.24), imposing a first order optimality condition on the augmented Lagrangian with respect to the primal variable u, leads to the following linear system

$$\left(\mathrm{D}^T\mathrm{D} + \frac{\gamma_w}{\gamma_v}\mathrm{K}^T\mathrm{K}\right) u = \mathrm{D}^T\left(v^{(t)} - \frac{1}{\gamma_v}\rho_v^{(t-1)}\right) + \frac{\gamma_w}{\gamma_v}\mathrm{K}^T\left(w^{(t)} - \frac{1}{\gamma_w}\rho_w^{(t-1)} + b\right) , \tag{5.33}$$

that can be solved, since

$$\mathrm{null}\left(\mathrm{D}^T\mathrm{D} + \frac{\gamma_w}{\gamma_v}\mathrm{K}^T\mathrm{K}\right) = \mathrm{null}\left(\mathrm{D}^T\mathrm{D}\right) \cap \mathrm{null}\left(\mathrm{K}^T\mathrm{K}\right) = \mathrm{null}\left(\mathrm{D}\right) \cap \mathrm{null}\left(\mathrm{K}\right) = \{\mathbf{0}_n\}, \qquad (5.34)$$

i.e. the coefficient matrix has full rank. In our case, condition (5.34) is satisfied; in fact, K represents a blurring operator, which is a low-pass filter, whereas the regularization matrix D is a first-order difference operator and, hence, is a high-pass filter. Moreover, since $\gamma_v, \gamma_w > 0$, the coefficient matrix in (5.33) is symmetric positive definite and typically highly sparse. Hence, the linear system in (5.33) can be solved quite efficiently by the iterative (eventually preconditioned) conjugate gradient method. Moreover, under appropriate assumptions about the solution u near the image boundary, the linear system can be solved even more efficiently. We assume *periodic* boundary conditions for u, so that both $\mathrm{D}^T\mathrm{D}$ and $\mathrm{K}^T\mathrm{K}$ are block circulant matrices with circulant blocks and, hence, the coefficient matrix in (5.33) can be diagonalized by the 2D discrete Fourier transform (FFT implementation). Provided that the penalty parameters γ_v, γ_w are kept fixed during the ADMM iterations, the coefficient matrix in (5.33) does not change and it can be diagonalized once for all at the beginning. Therefore, at any ADMM iteration the linear system (5.33) can be solved by one forward 2D FFT and one inverse 2D FFT, each at a cost of $O(n \log n)$.

5.5.2 Estimation of the global parameter μ

We remark that in Section 5.3, that was devoted to detail the parameter estimation procedure, we did not mention how to deal with the estimation of the global parameter μ. The reason why this topic has been postponed is that it is strictly connected to the update of u.

In fact, μ is updated along the iterations so as to fulfill the global discrepancy principle as described in [88]: we ask each iterate $u^{(t)}$ to satisfy the condition

$$||\mathrm{K}u^{(t)} - b||_2 \leq \delta := \tau\sigma\sqrt{n}\,,$$

where σ is the noise standard deviation and the parameter $\tau \approx 1$ is set a priori. In order to avoid under-estimation or over-estimation of the noise level, τ can be set slightly grater or less than 1, respectively. Moreover, as already mentioned in Section 3.4, here the μ-update is carried out via APE [89]. Recalling the definition of $z^{(t-1)}$ given in (5.31), the update reads:

$$\begin{aligned} \|z^{(t-1)}\|_2 \leq \delta \quad &\Longrightarrow \quad \mu^{(t)} = 0, \\ \|z^{(t-1)}\|_2 > \delta \quad &\Longrightarrow \quad \mu^{(t)} = \gamma_w\big(\|\, z^{(t-1)}\|_2/\delta - 1\big). \end{aligned} \qquad (5.35)$$

Once that both the α and u update step in the IAS outer scheme in Algorithm 1 have been detailed, the pseudo-code of the ADMM-based procedure for the solution of HWTV-L$_2$ model and the embedded estimation of the space-variant and global parameters, is reported in Algorithm 2.

In Algorithm 2 both the local space-variant parameters α_i and the global parameter μ are updated along the iterations. This is a standard strategy for this type of optimization problems (see, e.g., [88]), especially in the case of a cheap update of parameters adapting to the image quality improvement. Despite the iterative change in the expression of the cost functional corresponding to such update, we remark that we observed empirical convergence of the algorithm. We are showing some results in this sense in the experimental section.

Algorithm 2: ADMM-based algorithm for the HWTV-L$_2$ model

```
input:  $b \in \mathbb{R}^n$, $r > 0$, $\tau \approx 1$, $\gamma_v > 0$,$\gamma_w > 0$
output:  restored image $u^*$
initialize:  set $u^{(0)} = b$, $\rho_w^{(0)} = \mathbf{0}_n$, $\rho_v^{(0)} = \mathbf{0}_{2n}$
for  $t$ = 1, 2, ... until convergence  do:
    update parameters
        · $\alpha_i^{(t)}$ by (5.17) for every $i = 1, \ldots, n$
        · $\mu^{(t)}$ by (5.35)
    update primal variables
        · $v^{(t)}$ by (5.29)
        · $w^{(t)}$ by (5.32)
        · $u^{(t)}$ by solving (5.33)
    update dual variables
        · $\rho_w^{(t)}$ by (5.25)
        · $\rho_v^{(t)}$ by (5.26)
end for
return:  $u^* = u^{(t)}$
```

Remark 2. *The prior (5.11) leading to the HWTV-L$_2$ model in (5.8) has been introduced as a space-variant modification of the isotropic TV prior. Clearly, the anisotropic version of (5.11) can be introduced as well. More in general, we have*

$$\mathrm{P}(u \mid \theta_{\mathrm{pr}}) = \frac{1}{Z} \exp\left(-\sum_{i=1}^{n} \alpha_i \|(\mathrm{D}u)_i\|_d \right) , \quad \text{with } \ d = 1, 2 ,$$

and the corresponding model reads as

$$u^* \in \arg\min_{u\in\mathbb{R}^n} \left\{ \sum_{i=1}^{n} \alpha_i \|(\mathrm{D}u)_i\|_d + \frac{\mu}{2}\|\mathrm{K}u - b\|_2 \right\} , \quad \text{with } \ d = 1, 2 . \tag{5.36}$$

We also remark that, from the computational point of view, when considering $d = 1$, the ADMM-based algorithm is not modified, except for the sub-problem for v. In fact, (5.28) turns into

$$v_i^{(t)} \in \arg\min_{v_i\in\mathbb{R}^2} \left\{ \|v_i\|_1 \ + \frac{(\gamma_v/\alpha_i)}{2} \left\| v_i - q_i^{(t-1)} \right\|_2^2 \right\}, \quad i = 1, \dots, n .$$

Also in this case, $v_i^{(t)}$ is related to a proximity operator, namely,

$$v_i^{(t)} = \mathrm{prox}_{\frac{\alpha_i}{\gamma_v}\|\cdot\|_1} \left(q_i^{(t-1)} \right) ,$$

and again a closed-formula is available,

$$v_i^{(t)} = q_i \max\left(1 - \frac{\alpha_i^{(t)}}{\gamma_v \|q_i^{(t-1)}\|_1}, 0 \right) , \quad i = 1, \dots, n.$$

5.6 Computed examples

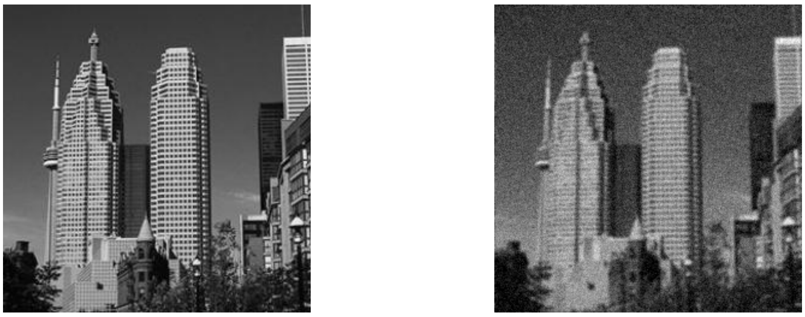
In this section we report some numerical results obtained by solving the image restoration model (5.8) via the ADMM-based Algorithm 2, with stopping criteria based on the number of iterations as well as on the iterates relative change, i.e. we stop iterating as soon as

$$t \geq 1000 , \quad \text{or} \quad \delta_u = \frac{\|u^{(t)} - u^{(t-1)}\|_2}{\|u^{(t-1)}\|_2} \leq 10^{-4} . \tag{5.37}$$

Due to the automatic estimation procedure of the space-variant parameters α_i and global parameter μ, the only parameters that need to be fixed in Algorithm 2 are the penalty parameters γ_v and γ_w. We set $\gamma_v = 20$ and $\gamma_w = 100$, observing in our experiments that the convergence properties of the algorithm are not affected by this choice, if not in terms of convergence speed. The value $r > 0$ denotes the radius of the neighborhoods $\mathcal{C}_i^r$ defined in Section 5.3 and used to estimate the space-variant parameters α_i. Denoting by $u \in [0, 1]^n$ the ground-truth image, we assess the quality of the reconstruction u^* by means of the Improved Signal-to-Noise Ratio

$$\mathrm{ISNR}(b, u, u^*) := 10 \log_{10} \frac{\|b - u\|_2^2}{\|u^* - u\|_2^2} ,$$

and in terms of the Structural Similarity Index (SSIM) [167]. We compare our results with the ones obtained by the standard TV-L$_2$ model in (4.1) model, the SATV approach [58] based on coupling the (5.7) model with a local discrepancy-based procedure for automatically selecting the parameters μ_i and the bilevel learning strategy used in [92, 93] to estimate the parameters α_i of (5.6) model via a nested optimization procedure.



(a) Original u.

(b) Corrupted b.

Figure 5.3: Original image (a) and observed image corrupted with AWGN with $\sigma = 0.05$ (b).

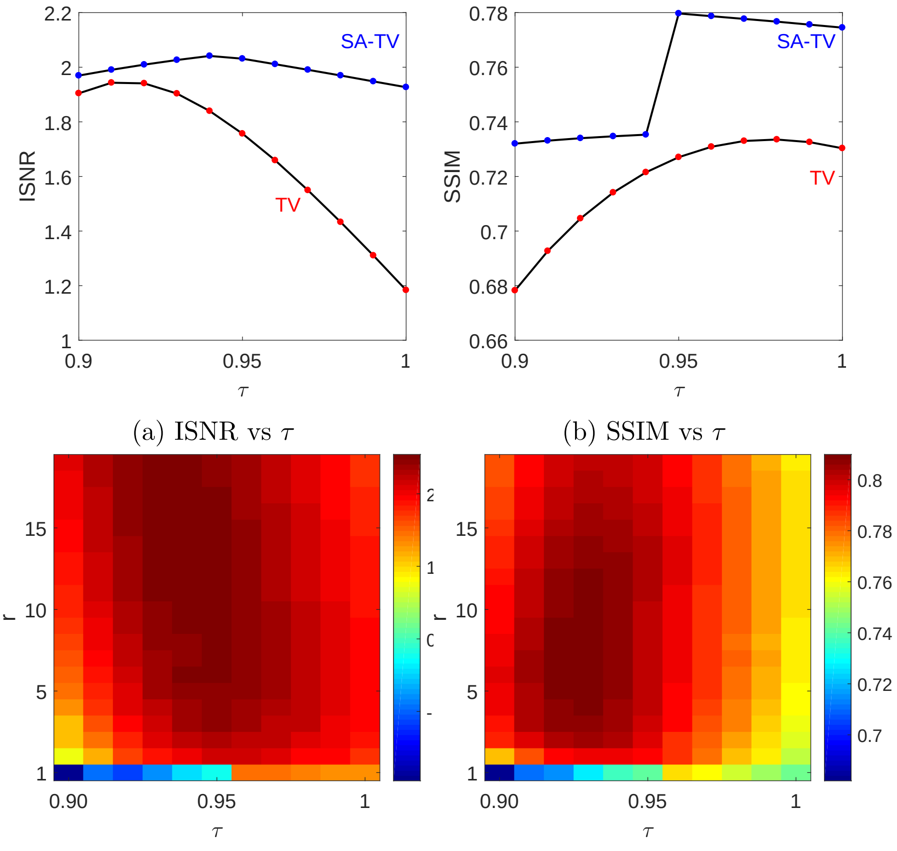
Image deblurring. We consider the `skyscraper` test image (256×256) corrupted by AWGN of two levels $\sigma = 0.02, 0.05$, and Gaussian blur of `band` $= 5$ and `width` $= 1$. The ground-truth image u and the observed image b for $\sigma = 0.05$ are shown in Figure 5.3a-5.3b, respectively. In this test we highlight the improvements obtained by Algorithm 2 for our model (5.8) in comparison to the TV-L$_2$ model, solved by means of ADMM, and the SATV[1]. As mentioned above, for the automatic adaption of the parameter μ along the iterations, a value for the parameter τ needs to be chosen. For the three models considered, we observed that the value of τ maximizing the ISNR does not necessarily correspond to the value maximizing the SSIM, see Figure 5.4a. For the TV-L$_2$ model the maximum SSIM is reached for $\tau \approx 1$, while the ISNR achieves its maximum when $\tau \approx 0.9$, the latter being the case in which texture is better preserved but noise is not completely removed. For SATV, the maximum ISNR and SSIM values are reached approximately for the same τ. As remarked in [58], the SATV method is robust with respect to the choice of the radius r of the neighborhoods used for the estimation. Thus, we set such parameter as the default value $r = 5$ in our tests. We performed similar sensitivity tests for our HWTV-L$_2$ model for different (τ, r) values. Results are shown in Figures 5.4c-5.4d. Larger values of ISNR are observed as r increases, while SSIM reaches its maximum for $r \approx 5$. In both cases, considering $\tau < 1$ helps in improving the quality of the final restoration. For each method, we then selected the parameter(s) yielding the maximum ISNR/SSIM values and compared the results obtained. In Table 5.1 we report the achieved ISNR/SSIM values, whereas in Figures 5.5-5.6 we show the associated restored images for the case of AWGN with $\sigma = 0.05$. We observe that our HWTV-L$_2$ method results in higher quality reconstructions if compared to TV-L$_2$ and SATV. Visual inspection confirms the effectiveness of our approach in distinguishing between textured and homogeneous regions, see Figures 5.5i-5.6i.

The output α-maps in the two cases proposed are shown in Figure 5.7.
In Figure 5.8 and Figure 5.9, for the two cases proposed, we report the behavior along the iterations of the relative changes in α and in u,

$$\delta_\alpha = \frac{\|\alpha^{(t)} - \alpha^{(t-1)}\|_2}{\|\alpha^{(t-1)}\|_2}, \qquad \delta_u = \frac{\|u^{(t)} - u^{(t-1)}\|_2}{\|u^{(t-1)}\|_2},$$

and of the objective function $\mathcal{J}$. We also plot in red solid line the estimated discrepancy value

[1] We used the MATLAB code available at: `https://www.math.hu-berlin.de/~hp_hint/software/satv.html`.



(a) ISNR vs τ

(b) SSIM vs τ

(c) ISNR(τ, r)

(d) SSIM(τ, r)

Figure 5.4: ISNR 5.4a and SSIM 5.4b values reached for different values of τ by applying TV-L$_2$ and SATV to the restoration of `skyscraper` test image in Figure 5.3a. For the same image, ISNR 5.4c and SSIM 5.4d values achieved by HWTV-L$_2$ method for different values of τ and r.

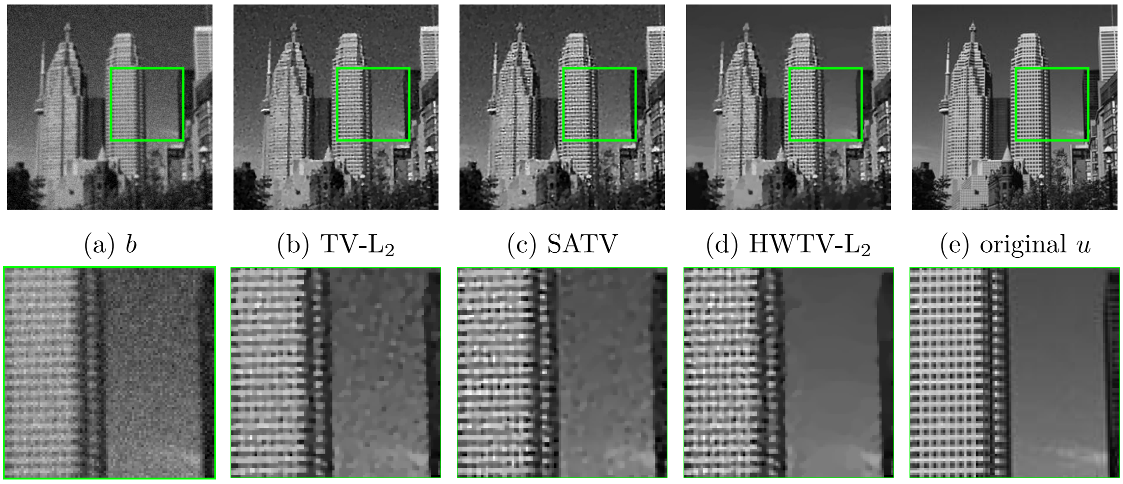
	$\sigma = 0.02$			$\sigma = 0.05$		
	TV	SATV	HWTV	TV	SATV	HWTV
ISNR	3.4701	3.6625	**4.3331**	1.9433	2.0414	**2.5408**
SSIM	0.8733	0.8966	**0.9007**	0.7335	0.7797	**0.8099**

Table 5.1: Maximum ISNR/SSIM values achieved by TV-L$_2$, SATV and HWTV-L$_2$ on the `skyscraper` test image in Figure 5.3a(top) corrupted by AWGN of two different levels.

$\tilde{\delta}$ at each iteration,

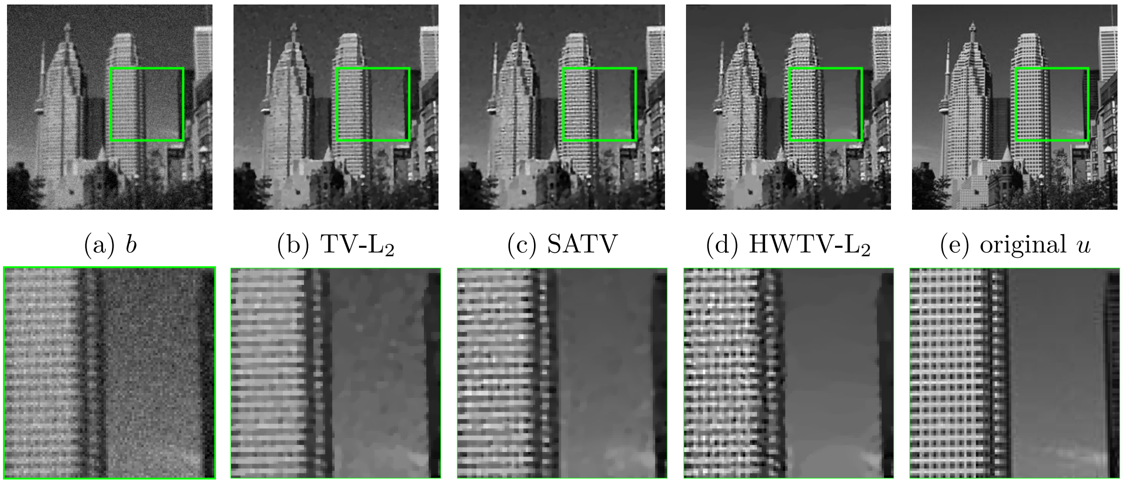
$$\tilde{\delta} = \frac{\|\mathrm{K}u - b\|_2}{\sqrt{n}},$$

compared with the true discrepancy value, i.e. $\delta/\sqrt{n} = \tau\sigma$, in black dashed line. In Figure 5.8, one can notice that, after the first iterations, the relative changes δ_α and δ_u decrease monotonically, while the behavior of the objective function $\mathcal{J}$ is characterized by a global and smooth decrease. For what concerns the plots in Figure 5.9, observe that the behavior are approximately the same for all the quantities monitored, except for the relative change in α, δ_α. In fact, when considering small radius, the update of the parameters can be more unstable, even if this choice helps preserving finer details in the image.


(a) b (b) TV-L$_2$ (c) SATV (d) HWTV-L$_2$ (e) original u

(f) close-up of (a) (g) close-up of (b) (h) close-up of (c) (i) close-up of (d) (j) close-up of (e)

Figure 5.5: **ISNR optimization**. *First row*: Corrupted image b (a), reconstruction of image in Figure 5.3a by TV-L$_2$ ($\tau = 0.91$) (b), SATV ($\tau = 0.94$) (c), HWTV-L$_2$ ($\tau = 0.94$, $r = 14$) (d) and original image u (e). *Second row*: close-up(s).

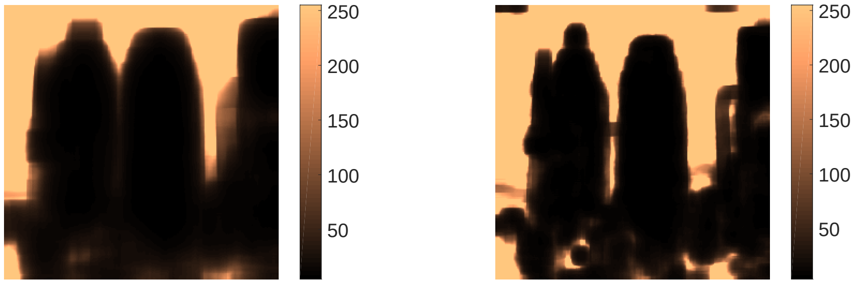

(a) b (b) TV-L$_2$ (c) SATV (d) HWTV-L$_2$ (e) original u

(f) close-up of (a) (g) close-up of (b) (h) close-up of (c) (i) close-up of (d) (j) close-up of (e)

Figure 5.6: **SSIM optimization**. *First row*: Corrupted image (a), reconstruction of image in Figure 5.3a by TV-L$_2$ ($\tau = 0.98$) (b), SATV ($\tau = 0.95$) (c), HWTV-L$_2$ ($\tau = 0.93$, $r = 6$) (d) and original image u (e). *Second row*: close-up(s).

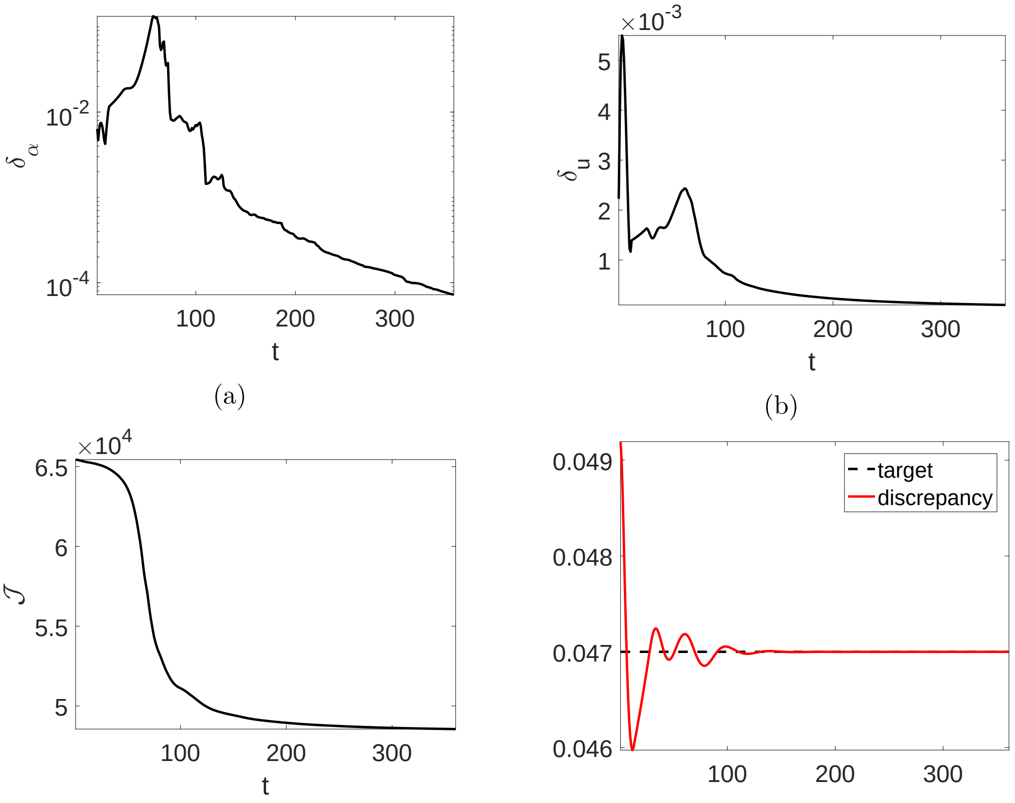
Image denoising. We now consider the test image `turtle`[2] (150×200) corrupted by AWGN of level $\sigma = 0.1$ (see Figures 5.10a -5.10b) and focus on the quality and computational improvements of our HWTV-L$_2$ method in the case of anisotropic TV (TVA) - i.e. $d = 1$ in (5.36) - in comparison to the alternative bilevel optimization strategy used [92, 93] for estimating the space-variant parameters α_i. After optimizing the HWTV-L$_2$ method over τ as discussed above, the maximum achieved value is SSIM = 0.7708 (for $r = 40$, $\tau = 0.86$), while the restoration via bilevel optimization strategy reaches SSIM = 0.7602. The reconstructions are shown in Figures 5.10c-5.10d. We remark that, in addition to the obtained SSIM and visual improvements, our

[2] Photo courtesy of K. Papafitsoros.



(a) Output α for 5.5d (b) Output α for 5.6d

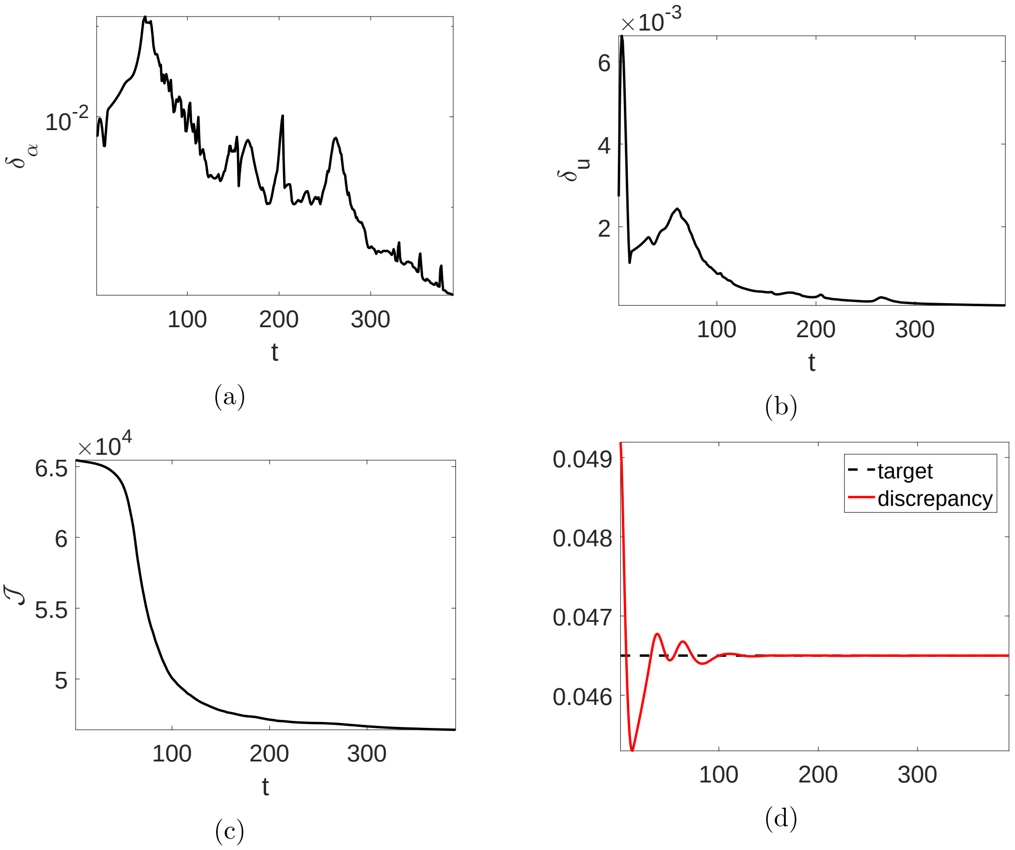
Figure 5.7: Output map of parameters α for the restoration via HWTV-L$_2$ maximizing the ISNR (a) and the SSIM (b).



(a) (b) (c) (d)

Figure 5.8: **ISNR optimization.** Relative change δ_α (a), relative change δ_u (b) and objective function $\mathcal{J}$ (c), estimated discrepancy (d) along the iterations.

approach exhibits a very high computational efficiency, whereas bilevel codes are known to be computational expensive and hardly applicable to high-resolution images. For instance, in this experiments the proposed ADMM Algorithm 2 for the HWTV-L$_2$ model required only 40 seconds on a standard laptop, compared to the 1429 seconds required by the bilevel algorithm [93].



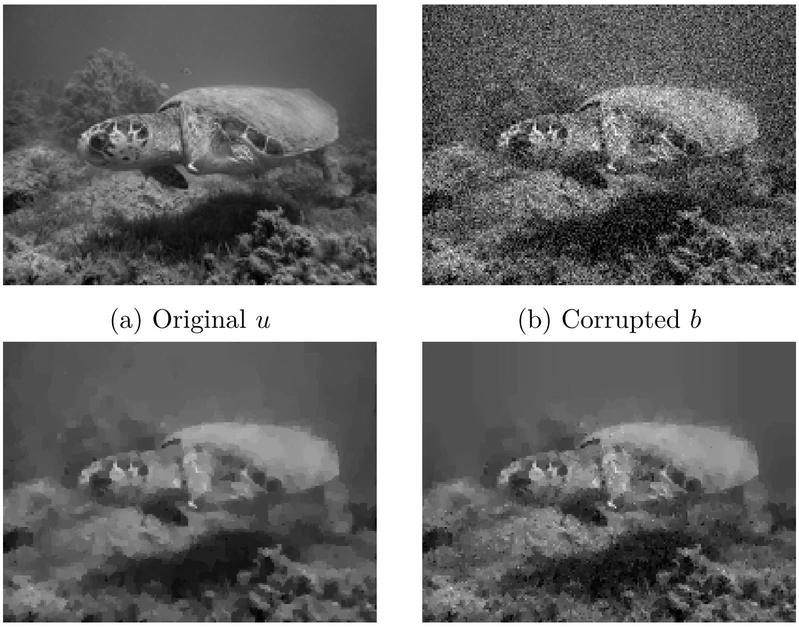
(a)

(b)

(c)

(d)

Figure 5.9: **SSIM optimization.** Relative change δ_α (e), relative change δ_u (f) and objective function $\mathcal{J}$ (g), estimated discrepancy (h) along the iterations.



(a) Original u

(b) Corrupted b

(c) Bilevel WTV

(d) HWTV

Figure 5.10: **SSIM optimization**. Original image (a), observed image corrupted by AWGN with $\sigma = 0.1$ (b), WTV reconstruction obtained by bilevel optimization of parameters α_i as in [92, 93] (SSIM = 0.7602) (c) and HWTV reconstruction ($\tau = 0.86$, $r = 40$, SSIM = **0.7708**) (d).

Chapter 6

Space-variant type regularization tuning

In the previous chapter, the problem of tuning the amount of regularization over the image in order to take into account local features has been tackled. We now focus on the qualitative way in which the regularization is carried out. Our goal is to adapt the type of the regularization to different local structures in the image. In order to do that, we will consider a further space-variant parameter to the already designed WTV regularizer in (5.6), with the purpose of exploiting the additional flexibility provided by a second local degree of freedom.

We are still referring to the image restoration problem in (5.1). As already highlighted, the adoption of the TVI prior is equivalent to implicitly assuming the gradient magnitudes to follow a one-parameter half Laplacian distribution. Based on the observation that such distribution is not sufficiently flexible, the authors in [111] proposed a generalization of the TV regularizer, referred to as TV_p. The TV_p regularizer relies on the adoption of a space-invariant, two-parameters half-Generalized Gaussian (hGG) distribution for modeling the distribution of the ℓ_2-norm of the gradients:

$$\mathrm{P}(\|(\mathrm{D}u)_i\|_2; p, \alpha) \;=\; \begin{cases} \frac{\alpha p}{\Gamma(1/p)} \exp\left(-(\alpha\|(\mathrm{D}u)_i\|_2)^p\right) & \text{for} \quad \|(\mathrm{D}u)_i\|_2 \geq 0 \\ \\ 0 & \text{for} \quad \|(\mathrm{D}u)_i\|_2 < 0 \end{cases},$$

where p denotes the additional parameter, namely the *shape* parameter of the hGG distribution, and Γ denotes the Gamma function. This family of distributions covers a wider spectrum of pdfs including half-Laplacian ($p = 1$) and half-Gaussian ($p = 2$). Some of them, corresponding to different values of the shape parameter p and fixed scale parameter $\alpha = 1$, are plotted in Figure 6.1.

The presence of a second parameter p allows, in principle, for a better approximation of the gradient magnitude distribution depending on the image at hand, and leads to the introduction of the TV_p prior [111]:

$$\mathrm{P}(u) = \frac{1}{Z} \exp\left(-\alpha \sum_{i=1}^{n} \|(\nabla u)_i\|_2^p\right) = \frac{1}{Z} \exp\left(-\alpha\, \mathrm{TV}_p(u)\right). \tag{6.1}$$

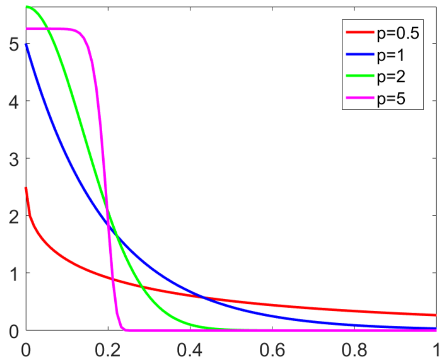


Figure 6.1: The hGG pdfs for different values of p and fixed $\alpha = 1$.

The TV_p prior has been proven to outperform the TV prior especially on images presenting global properties resembling the local ones, such as piece-wise constant images. Nevertheless, despite the flexibility gained with the introduction of a further parameter p, we do expect the TV_p regularizer to experience all the limitations that typically characterize global regularization terms. In this chapter, we are investigating the motivation and the benefits of the introduction of a new prior based on a space-variant modification of the TV_p prior in (6.1).

6.1 The $\mathrm{TV}^{\mathrm{sv}}_{p,\alpha}$-$\mathrm{L}_2$ model

In [110], we proposed a further generalization of the TV-L_2 model reading as,

$$u^* \in \arg\min_{u\in\mathbb{R}^n} \left\{ \mathrm{TV}^{\mathrm{sv}}_{p,\alpha}(u) + \frac{\mu}{2}\|\mathrm{K}u - b\|_2^2 \right\}, \tag{6.2}$$

where the new space-variant $\mathrm{TV}^{\mathrm{sv}}_{p,\alpha}$ regularizer defined as

$$\mathrm{TV}^{\mathrm{sv}}_{p,\alpha}(u) := \sum_{i=1}^{n} \alpha_i \|(\nabla u)_i\|_2^{p_i}, \quad \alpha_i \in\,]0,+\infty[, \;\; p_i \in\,]0,2] \;\; \forall\, i \in \Omega\,, \tag{6.3}$$

is coupled with a L_2 fidelity term, since the noise is still known to be AWG. The proposed regularizer in (6.3) is highly flexible as it is characterized by two per-pixel free parameters p_i, α_i, such that local, space-variant properties of the target clean image u can be potentially addressed. As in the case of the WTV regularization term, the usefulness of such a great flexibility in the proposed regularizer is conditioned to the existence of effective procedures for the automatic estimation of the p_i and α_i parameters. Hence, we also propose a suitable method for the automatic estimation of such parameters from the observed image based, as before, on a statistical procedure involving non-informative hyperpriors.

Besides the space-variant parameters, a global regularization parameter μ appears in model (6.2), thus making also the $\mathrm{TV}^{\mathrm{sv}}_{p,\alpha}$-$\mathrm{L}_2$ a hybrid model. The estimation of μ is still based on a discrepancy principle, similarly as described in Section 5.5.2.

6.2 Model derivation via MAP

The rational of our proposal is that the distribution of the gradient magnitudes of the unknown clean image is space-variant and it is well modeled locally by a two-parameters Generalized

Gaussian distribution. In Section 5.2, we remarked the connection between a space-variant approach and the assumption according to which u can be modeled as a non-stationary Markov Random Field, with prior

$$\mathrm{P}(u \mid \theta_{\mathrm{pr}}) = \frac{1}{Z} \exp\left(- \sum_{i=1}^{n} V_{\mathcal{N}_i^r}(u; \theta_{\mathrm{pr}_i}) \right) ,$$

where, as before, $V_{\mathcal{N}_i^r}$ is the i-th potential function defined on the clique $\mathcal{N}_i^r$ of radius r centered at pixel i. Here, we consider a modified version of the TV_p prior in (6.1) with the shape and scale parameters of the hGG distribution changing at any pixel, thus leading to the following prior

$$\Pr(u \mid \theta_{\mathrm{pr}}) = \frac{1}{Z} \exp\left(- \sum_{i=1}^{n} \alpha_i \|(\nabla u)_i\|_2^{p_i} \right) = \frac{1}{Z} \exp\left(- \mathrm{TV}_{p,\alpha}^{\mathrm{sv}}(u) \right) , \tag{6.4}$$

with

$$Z = \left(\prod_{i=1}^{n} \left(\frac{\alpha_i p_i}{\Gamma\left(\frac{1}{p_i}\right)} \right) \right)^{-1} ,$$

and

$$\theta_{\mathrm{pr}} = \begin{pmatrix} \theta_{\mathrm{pr}_1} \\ \theta_{\mathrm{pr}_2} \\ \vdots \\ \theta_{\mathrm{pr}_n} \end{pmatrix} \in \mathbb{R}^{n\times 2} , \quad \theta_{\mathrm{pr}_i} = (\alpha_i, p_i) \in \mathbb{R}^2 .$$

In order to shed some light on the motivation behind the introduction of a second space-variant parameter, we propose the same analysis carried out in Section 5.2. More specifically, we consider the histogram of the gradient magnitudes of the global image, plotting with a solid green line the hGG pdf returning the best approximation of the histogram - see Figure 6.2b and the close up in Figure 6.2c. We then choose the same two regions selected in Figure 5.1, namely a cyan-bordered constant one and a red-bordered textured one. In Figure 6.2e, the cyan solid line represents the hGG pdf that best fits the local histogram for the constant region, whereas the super-imposed dashed cyan line and green solid line represent the local half-Laplacian pdf already plotted in Figure 5.1e and the global hGG pdf, respectively. The interpretation of the lines in Figure 6.2h is analogous. In the case of the constant region, we can observe a slight improvement in terms of a better modeling of the histogram behavior. The improvement is more remarkable on the textured region, where the rich gradient structure is poorly approximated by a one-parameter distribution.

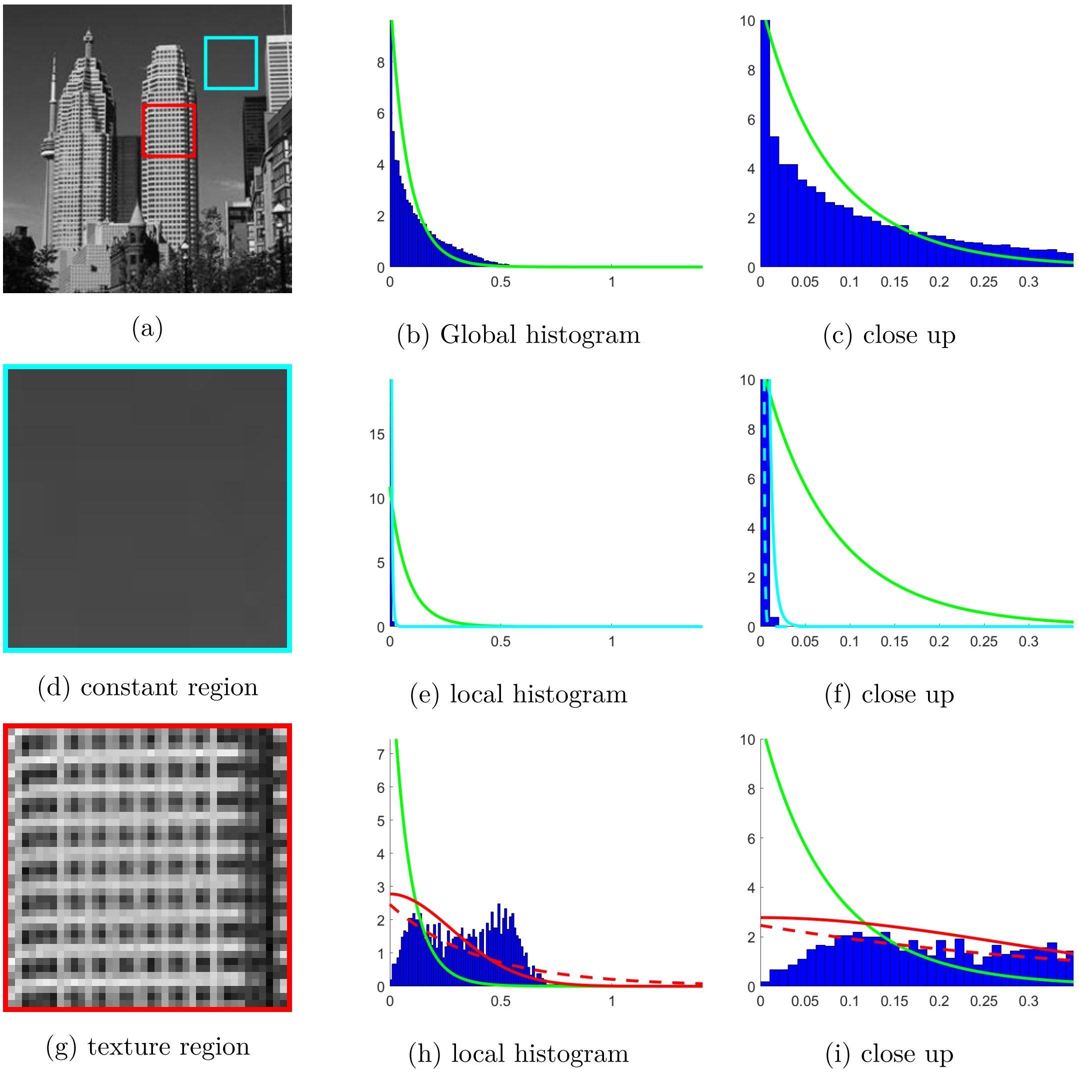
After recalling the expression of the Gaussian likelihood

$$\mathrm{P}(b \mid u) = \frac{1}{W} \exp\left(- \frac{1}{2\sigma^2} \|\mathrm{K}u - b\|_2^2 \right) ,$$

we replace the latter and the proposed prior (6.4) in the MAP inference formula; dropping the constant terms, we obtain our $\mathrm{TV}_{p,\alpha}^{\mathrm{sv}}$-$\mathrm{L}_2$ model in (6.5), that is in extended form:

$$u^* \in \arg\min_{u\in\mathbb{R}^n} \left\{ \sum_{i=1}^{n} \alpha_i \|(\nabla u)_i\|_2^{p_i} + \frac{\mu}{2} \|\mathrm{K}u - b\|_2^2 \right\} , \tag{6.5}$$

where $\mu = 1/\sigma^2$.



(a)

(b) Global histogram

(c) close up

(d) constant region

(e) local histogram

(f) close up

(g) texture region

(h) local histogram

(i) close up

Figure 6.2: From top to bottom: histogram of the gradient magnitudes on the whole test image, on a constant region and on a texture region, with the corresponding close-up(s).

6.3 Parameter estimation via non-informative hyperpriors

As already remarked, the effectiveness of a space-variant approach is also strictly related to the derivation of an automatic estimation procedure for the unknown parameters involved in the expression of the $\mathrm{TV}_{p,\alpha}^{\mathrm{sv}}$ prior. Notice that here the number of parameters to set is twice the number of pixels. Hence, the minimization problem in (6.5) represents only one of the two steps of the outer scheme outlined in Algorithm 3.

In this section, we first focus on the parameter estimation step, adopting the same strategy and notation already introduced in Chapter 5. For any pixel i, in order to estimate the parameters defining the local hGG distribution from which the magnitude of the gradient $\|(\mathrm{D}u)_i\|_2$ is assumed to be drawn, one has to solve

$$(\alpha_i^{(t)}, p_i^{(t)}) \in \arg \min_{\alpha_i \in \mathbb{R}_+,\, p_i \in]0,2]} \left\{ -\log \mathrm{P}(\mathcal{S}_i \mid \alpha_i, p_i) \right\}. \tag{6.6}$$

Algorithm 3: IAS with non-informative hypeprior for the $\mathrm{TV}^{\mathrm{sv}}_{p,\alpha}$-$\mathrm{L}_2$ model

```
input: observed image $b \in \mathbb{R}^n$
output: restored image $u^*$
1. initialize: set $u^{(0)} = b$
2. for $t = 1, 2, \ldots$ until convergence do:
3.     · update $(\alpha^{(t)}, p^{(t)}) \in \arg \min_{\alpha \in \mathbb{R}^n_+,\, p \in ]0,2]^n} \left\{ -\log \mathrm{P}(u^{(t-1)} \mid \alpha, p) \right\}$ ,
4.     · update $u^{(t)} \in \arg \min_{u \in \mathbb{R}^n} \left\{ -\log \mathrm{P}(u \mid \alpha^{(t)}, p^{(t)}) - \log \mathrm{P}(b \mid u, \alpha^{(t)}, p^{(t)}) \right\}$
5. end for
6. return: $u^* = u^{(t)}$
```

We recall that we are denoting by

$$\mathcal{S}_i = \{x_{i,j}\}_{j=1}^N \,, \quad \text{with} \quad x_{i,j} = \|(\mathrm{D}u^{(t-1)})_j\|_2 \,,$$

the ensemble of gradient magnitudes computed at pixels belonging to the clique $\mathcal{N}_i^r$ of radius r and side $2r+1$. For what concerns the constrain set in the minimization problem (6.6), observe that the local scale parameter α_i is only required to be strictly positive, whereas we ask the local shape parameter p_i to belong to the interval $]0,2]$. From a theoretical point of view, we could just impose $p_i > 0$, but we do not expect any significant benefit in letting p_i be greater than 2.

Assuming the independence of the samples, the conditioned pdf $\mathrm{P}(\mathcal{S}_i \mid \alpha_i, p_i)$, which has already been recognized as the local likelihood function, takes the form

$$\mathrm{P}(\mathcal{S}_i \mid \alpha_i, p_i) = \prod_{j=1}^N \mathrm{P}(x_{i,j} \mid \alpha_i, p_i) = \left(\frac{\alpha_i p_i}{\Gamma(1/p_i)} \right)^N \exp\left(-\sum_{j=1}^N (\alpha_i x_{i,j})^{p_i} \right) .$$

We thus have

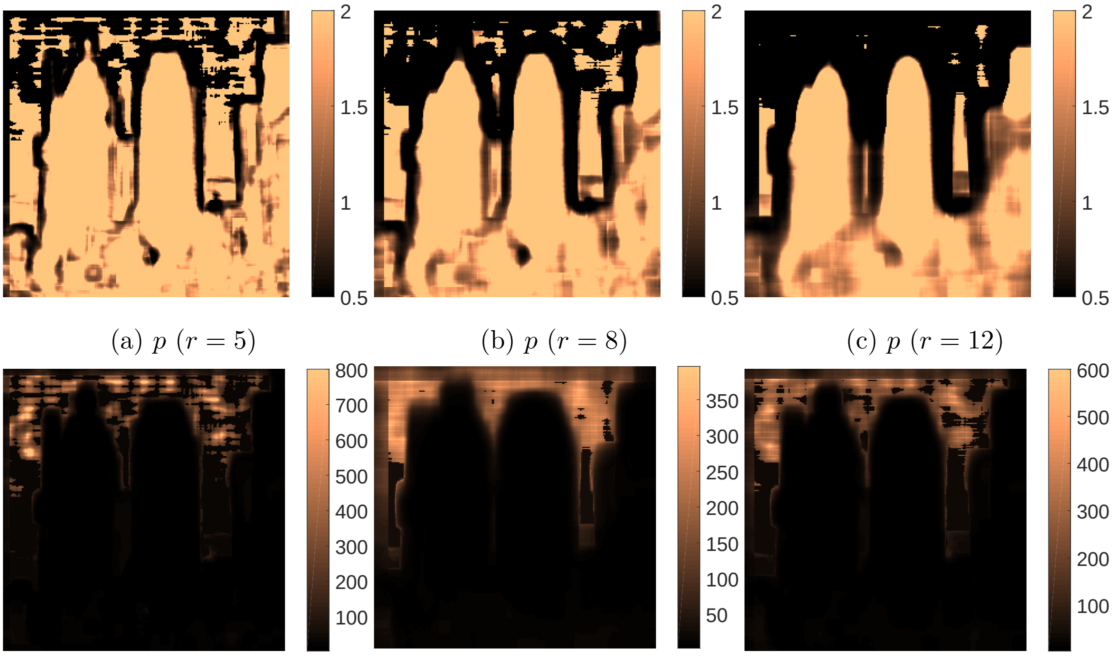
$$\begin{aligned} -\log \mathrm{P}(\mathcal{S}_i \mid \alpha_i, p_i) &= -N \log \alpha_i + N \log \frac{1}{p_i} + N \log \Gamma\left(\frac{1}{p_i}\right) + \alpha_i^{p_i} \sum_{j=1}^N x_{i,j}^{p_i} \\ &= -N \log \alpha_i + N \log \Gamma\left(1 + \frac{1}{p_i}\right) + \alpha_i^{p_i} \sum_{j=1}^N x_{i,j}^{p_i} \,. \end{aligned} \tag{6.7}$$

Imposing a first order optimality condition on $-\log \mathrm{P}(\mathcal{S}_i|\alpha_i, p_i)$ with respect to α_i, we thus get

$$\frac{\partial}{\partial \alpha_i} \left(-\log \mathrm{P}(\mathcal{S}_i|\alpha_i, p_i) \right) = -\frac{N}{\alpha} + p_i \alpha_i^{p_i - 1} \sum_{j=1}^N x_{i,j}^{p_i} = 0 \,,$$

from which a closed formula for the update of the i-th scale parameter α_i is derived:

$$\alpha_i = \left(\frac{p_i}{N} \sum_{j=1}^N x_{i,j}^{p_i} \right)^{-\frac{1}{p_i}} . \tag{6.8}$$



(a) p $(r = 5)$ (b) p $(r = 8)$ (c) p $(r = 12)$

(d) α $(r = 5)$ (e) α $(r = 8)$ (f) α $(r = 12)$

Figure 6.3: Maps of shape parameter p and scale parameter α computed on test image `skyscraper` for different values of r.

After substituting the expression for α_i in (6.7), the following minimization problem for the update of p_i is obtained

$$p_i^{(t)} \in \arg\min_{p_i \in (0,2]} \left\{ \frac{N}{p} \log\left(\frac{p_i}{N} \sum_j x_{i,j}^{p_i} \right) + N \log \Gamma\left(1 + \frac{1}{p_i}\right) + \frac{N}{p_i} \right\} \tag{6.9}$$

Problem (6.9) can be reformulated as a minimization problem over a compact constrain set,

$$p_i^{(t)} \in \arg\min_{p_i \in [\epsilon,2]} \left\{ f(p_i) := \frac{N}{p_i} \log\left(\frac{p_i}{N} \sum_j x_{i,j}^{p_i} \right) + N \log \Gamma\left(1 + \frac{1}{p_i}\right) + \frac{N}{p_i} \right\}. \tag{6.10}$$

with ϵ being slightly greater than 0.

Proposition 3. *Function $f : [\epsilon, 2] \to \mathbb{R}$ is continuous and admits a minimum in its compact domain.*

In Figure 6.3 the maps of local p and α values, obtained with neighborhoods of different sizes starting from the original test image `skyscraper` are shown. The p-maps have been computed by setting $\epsilon = 0.5$, discretizing the compact interval $[0.5, 2]$ with 50 grid values and performing a grid-search. As the radius r increases, image features of increasing scale are highlighted, but in any case the method associates very low p values with flat regions and higher values with texture, while the strength of the regularization is expected to be again weaker on regions presenting many details. It is worth remarking that in Section 6.6 numerical experiments have been carried out by computing the p-map starting from the corrupted images. Before going on with the numerical solution of model (6.3), we want to draw a connection between the p value

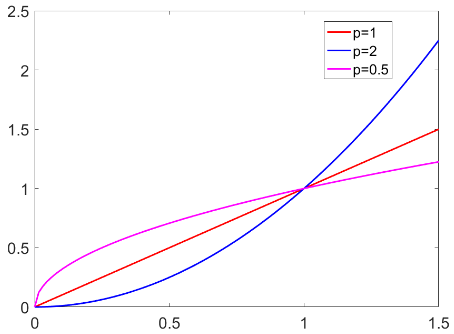


Figure 6.4: Behavior of function $|\cdot|^p$ for different values of p.

and the sparsity promotion properties of the regularization. The i-th term in the sum defining the regularizer $\mathrm{TV}^{\mathrm{sv}}_{p,\alpha}$ can be thought as a function of the form

$$\phi(y) = |y|^{p_i}\,, \quad \text{where } y = \|(\mathrm{D}u)_i\|_2 \text{ and } p_i \in (0,2].$$

Function ϕ for different values of p_i is graphed in Figure 6.4. It is clear how the estimation of a value $p_i < 1$ leads to a stronger sparsity promotion. On the other hand, letting the shape parameter to assume value less than 1 can possibly make the model (6.3) non-convex.

6.4 Existence and uniqueness of solutions

In this section, we provide an existence and uniqueness - under certain condition - result for the solution of the proposed $\mathrm{TV}^{\mathrm{sv}}_{p,\alpha}$-$\mathrm{L}_2$ variational model (6.3). We start recalling a general lemma whose proof can be found in [50, Lemma 2.7.1], guaranteeing the existence of global minimizers.

Lemma 1. *Let* $\mathrm{A}_1 \in \mathbb{R}^{m\times n}$*,* $m \geq n$*,* $\mathrm{A}_2 \in \mathbb{R}^{q\times n}$*,* $q \geq n$*, be two linear operators satisfying*

$$\mathrm{null}(\mathrm{A}_1) \,\cap\, \mathrm{null}(\mathrm{A}_2) \;=\; \{\mathbf{0}_n\}\,,$$

and let $f_1 : \mathbb{R}^m \to [-\infty,+\infty]$ *and* $f_2 : \mathbb{R}^q \to [-\infty,+\infty]$ *be two proper, lower semicontinuous and coercive functions. Then, the function* $h : \mathbb{R}^n \to [-\infty,+\infty]$ *defined by*

$$h(x) := f_1(\mathrm{A}_1 x) + f_2(\mathrm{A}_2 x)$$

is lower semicontinuous and coercive.

We now apply this result to the $\mathrm{TV}^{\mathrm{sv}}_{p,\alpha}$-$\mathrm{L}_2$ model.

Proposition 4. *The* $TV^{\mathrm{sv}}_{p,\alpha}$-$\mathrm{L}_2$ *functional* $\mathcal{J} : \mathbb{R}^n \to \mathbb{R}$ *defined in* (6.3) *is continuous, bounded from below by zero and coercive, hence it admits global minimizers.*

Proof. Let $\mathrm{D} \in \mathbb{R}^{2n\times n}$ be the finite difference operator discretizing the image gradient, K be the blur operator, and let $f_1 : \mathbb{R}^{2n} \to \mathbb{R}$, $f_2 : \mathbb{R}^n \to \mathbb{R}$ be the functions defined by

$$\begin{aligned} f_1(y) \;&:=\; \sum_{i=1}^{n} \alpha_i \,\|(y_{2i-1}, y_{2i})\|_2^{p_i}, \qquad && y \in \mathbb{R}^{2n}, \\ f_2(z) \;&:=\; \frac{\mu}{2}\,\|z-b\|_2^2\,, && z \in \mathbb{R}^n. \end{aligned} \tag{6.11}$$

Then, the $\mathrm{TV}^{\mathrm{sv}}_{p,\alpha}$-$\mathrm{L}_2$ energy functional in (5.8) can be written as

$$\mathcal{J}(u) = f_1(\mathrm{D}u) + f_2(\mathrm{K}u).$$

It holds:

$$\left(\mathrm{null}(\mathrm{D}) = \mathrm{span}(\mathrm{I}_n)\right) \cap \left(\mathrm{null}(\mathrm{K})\right) \;=\; \{\mathbf{0}_n\}\,,$$

in fact constant images do not belong to the null space of the linear blur operator K. Furthermore, functions f_1 and f_2 in (6.11) are clearly continuous, bounded from below by zero and coercive. It thus follows from Lemma 1 that the $\mathrm{TV}^{\mathrm{sv}}_{p,\alpha}$-$\mathrm{L}_2$ functional $\mathcal{J}$ in (5.8) is continuous, bounded from below by zero and coercive, hence it admits at least one global minimizer. □

In general, uniqueness of solution is not guaranteed. However, if the regularizer is convex, assuming that K is analytically non-singular, then the following result holds true.

Corollary 1. *Let $\mathcal{J} : \mathbb{R}^n \to \mathbb{R}$ be the $\mathrm{TV}^{\mathrm{sv}}_{p,\alpha}$-$\mathrm{L}_2$ functional defined in (6.3). If $p_i \geq 1$ for every $i = 1, \dots, n$, then $\mathcal{J}$ is strongly convex. Hence it admits a unique global minimizer.*

6.5 ADMM optimization

We now focus on the u-update step in Algorithm 3, namely,

$$u^* \;\in\; \arg\min_{u \in \mathbb{R}^n} \left\{ \sum_{i=1}^{n} \alpha_i \|(\nabla u)_i\|_2^{p_i} + \frac{\mu}{2}\|\mathrm{K}u - b\|_2^2 \right\}, \tag{6.12}$$

where α_i and p_i are fixed. Similarly as before, we consider a variable splitting technique [3] and introduce two auxiliary variables $w \in \mathbb{R}^n$ and $t \in \mathbb{R}^{2n}$, such that model (6.12) is rewritten in the following equivalent constrained form:

$$\begin{aligned} \{\, u^*, v^*, w^* \} \;\in\; \arg\min_{u,v,w} \left\{ \sum_{i=1}^{n} \alpha_i \|v_i\|_2^{p_i} + \frac{\mu}{2}\, \|w\|_2^2 \right\}, , \\ \text{subject to} : w = \mathrm{K}u - b\,, \quad v = Du\,, \end{aligned} \tag{6.13}$$

where $\mathrm{D} := (\mathrm{D}_h^T, \mathrm{D}_v^T)^T \in \mathbb{R}^{2n \times n}$ and $v_i := \left((\mathrm{D}_h u)_i\,, (\mathrm{D}_v u)_i\right)^T \in \mathbb{R}^2$ represents the discrete gradient of image u at pixel i. We consider the augmented Lagrangian functional introduced in Chapter 5

$$\begin{aligned} \mathcal{L}(u, v, w; \rho_v, \rho_w) \;=\; & \sum_{i=1}^{n} \alpha_i \|v_i\|_2^{p_i} + \frac{\mu}{2}\, \|w\|_2^2 - \langle\, \rho_v, v - \mathrm{D}u \,\rangle + \frac{\gamma_v}{2}\, \|v - \mathrm{D}u\|_2^2 \\ & - \langle\, \rho_w, w - (\mathrm{K}u - b) \,\rangle \;+ \frac{\gamma_w}{2}\, \|\, w - (\mathrm{K}u - b)\|_2^2\,, \end{aligned} \tag{6.14}$$

where $\gamma_w, \gamma_v > 0$ are scalar penalty parameters and $\rho_w \in \mathbb{R}^n$, $\rho_v \in \mathbb{R}^{2n}$ are the vectors of Lagrange multipliers associated with the linear constraints $w = \mathrm{K}u - b$ and $v = \mathrm{D}u$ in (6.13), respectively. As in the previous chapter, we look for the saddle point of the augmented Lagrangian. In other words, we set up an alternating chain of minimization problems for the primal variables u, w, v, and maximization problems for the dual variables ρ_w, ρ_v:

$$\begin{aligned}
v^{(t)} &\in \; \arg \min_{v \in \mathbb{R}^n} \; \mathcal{L}(u^{(t-1)}, w^{(t-1)}, v; \rho_w^{(t-1)}, \rho_v^{(t-1)})\,, \\
w^{(t)} &\in \; \arg \min_{w \in \mathbb{R}^{2n}} \; \mathcal{L}(u^{(t-1)}, w, v^{(t)}; \rho_w^{(t-1)}, \rho_v^{(t-1)})\,, \\
u^{(t)} &\in \; \arg \min_{u \in \mathbb{R}^n} \; \mathcal{L}(u, w^{(t)}, v^{(t)}; \rho_w^{(t-1)}, \rho_v^{(t-1)})\,, \\
\rho_w^{(t)} &\in \; \rho_w^{(t-1)} - \gamma_w \left(w^{(t)} - (\mathrm{K}u^{(t)} - b) \right), \\
\rho_v^{(t)} &\in \; \rho_v^{(t-1)} - \gamma_v \left(v^{(t)} - \mathrm{D}u^{(t)} \right).
\end{aligned} \tag{6.15}$$

The minimization problem with respect to w and u can be addressed as already done in Section 5.5. The global regularization parameter μ is updated along the iterations according to the discrepancy principle, as detailed in Section 5.5.2. Thus we only need to give a closer look to the minimization step with respect to the auxiliary variable v.

Minimization sub-problem for the primal variable v Given the definition of the augmented Lagrangian functional in (6.14), the minimization sub-problem for the primal variable v in (6.15) can be written as follows:

$$\begin{aligned}
v^{(t)} &\in \arg \min_{v \in \mathbb{R}^{2n}} \left\{ \sum_{i=1}^{n} \alpha_i \left\| v_i \right\|_2^{p_i} \, - \langle \rho_v^{(t-1)}, v - \mathrm{D}u^{(t-1)} \rangle + \frac{\gamma_v}{2} \left\| v - \mathrm{D}u^{(t-1)} \right\|_2^2 \right\} \\
&\in \arg \min_{v \in \mathbb{R}^{2n}} \left\{ \sum_{i=1}^{n} \alpha_i \left\| v_i \right\|_2^{p_i} \, + \frac{\gamma_v}{2} \left\| v - \left(\mathrm{D}u^{(t-1)} + \frac{1}{\gamma_v} \rho_v^{(t-1)} \right) \right\|_2^2 \right\} \\
&\in \arg \min_{v \in \mathbb{R}^{2n}} \sum_{i=1}^{n} \left\{ \alpha_i \left\| v_i \right\|_2^{p_i} \, + \frac{\gamma_v}{2} \left\| v_i - \left(\left(\mathrm{D}u^{(t-1)} \right)_i + \frac{1}{\gamma_v} \left(\rho_v^{(t-1)} \right)_i \right) \right\|_2^2 \right\}.
\end{aligned} \tag{6.16}$$

Note that in (6.16) the minimized functional is written in explicit component-wise (or pixel-wise) form, with $\left(\mathrm{D}u^{(t-1)} \right)_i, \left(\rho_v^{(t-1)} \right)_i \in \mathbb{R}^2$ denoting the discrete gradient and the Lagrange multipliers at pixel i, respectively. Solving the $2n$-dimensional minimization problem in (6.16) is thus equivalent to solve the n following independent 2-dimensional problems:

$$v_i^{(t)} \in \arg \min_{v_i \in \mathbb{R}^2} \left\{ \left\| v_i \right\|_2^{p_i} \, + \frac{(\gamma/\alpha_i)}{2} \left\| v_i - q_i^{(t-1)} \right\|_2^2 \right\}, \quad i = 1, \ldots, n\,,$$

with the constant vectors $q_i^{(t-1)} \in \mathbb{R}^2$ defined by

$$q_i^{(t-1)} \; := \; \left(\mathrm{D}u^{(t-1)} \right)_i + \frac{1}{\gamma_v} \left(\rho_v^{(t-1)} \right)_i \;, \quad i = 1, \ldots, n\,. \tag{6.17}$$

The solutions of the n optimization problems in (6.17) can be obtained as:

$$v_i^{(t)} \; = \; \xi_i \, q_i^{(t-1)} \;, \quad i = 1, \ldots, n\,, \tag{6.18}$$

In particular, the shrinkage coefficients $\xi_i \in [0, 1]$, $i = 1, \ldots, n$, are given by the following proposition that has been proven in [111].

Proposition 5. *Let $\beta > 0$, $0 < p < 2$ and $q \in \mathbb{R}^m$ with $m \geq 1$ be given constants. Then, the proximal operator of the m-variate function $f(x) = \|x\|_2^p$, $x \in \mathbb{R}^m$ defined as the m-dimensional minimization problem*

$$x^* \in \mathrm{prox}_{\beta f}(q) = \arg \min_{x \in \mathbb{R}^m} \left\{ \|x\|_2^p + \frac{\beta}{2} \|x - q\|_2^2 \right\}$$

is given by

$$x^* = \xi^* q\,, \;\; \textit{where } \xi^* \in [0,1]$$

with

$$\begin{array}{lll}
\textit{(a) } \xi^* = 0 & & \textit{if } \|q\|_2 = 0\,,\ \forall p \\
\textit{(b) } \xi^* = \max\{1 - 1/\gamma, 0\} & & \textit{if } \|q\|_2 > 0\,,\ p = 1 \\
\textit{(c) } \xi^* \textit{ unique solution in }]0,1[\textit{ of}: & & \\
\quad p\xi^{p-1} + \gamma(\xi - 1) & & \textit{if } \|q\|_2 = 0\,,\ 1 < p < 2 \\
\textit{(d) } \begin{cases} \xi^* = 0 & \textit{if } \gamma < \bar{\gamma} \\ \xi^* \in \left\{0, \bar{\xi}\right\} & \textit{if } \gamma = \bar{\gamma} \\ \xi^* \textit{ unique solution in }]\bar{\xi}, 1[\textit{ of} & \\ p\xi^{p-1} + \gamma(\xi - 1) & \textit{if } \gamma > \bar{\gamma} \end{cases} & & \textit{if } \|q\|_2 > 0\,,\ 0 < p < 1
\end{array}$$

where we set

$$\begin{aligned}
\gamma =& \beta \|q\|_2^{2-p} \\
\bar{\gamma} =& \frac{(2-p)^{2-p}}{(2-2p)^{1-p}}\,, \quad \bar{\xi} = 2\frac{1-p}{2-p}\,.
\end{aligned}$$

To summarize previous results, in Algorithm 4 we report the main steps of the proposed ADMM-based iterative scheme used to solve the saddle-point problem and, at the same time, compute the space-variant parameters.

As far as convergence of this algorithm is concerned, we remark that in convex settings numerous convergence results have been established for ADMM as well as its varieties, see for example [87] and references therein. In particular, when $p_i \geq 1\ \forall\, i$ and step 3-4 in Algorithm 4 are only performed for $t = 1$, i.e. α and p are not updated along the ADMM iterations, the convergence results hold for the proposed $\mathrm{TV}^{\mathrm{sv}}_{p,\alpha}$-$\mathrm{L}_2$ model. However, in case that one or more $p_i < 1$, the ADMM may be under non-convex settings, where there have been a few studies on the convergence properties. To the best of our knowledge, existing convergence results of ADMM for non-convex problems is very limited to particular classes of problems [158] and under certain conditions of the dual step size [96]. Nevertheless, the ADMM works extremely well for various applications involving non-convex optimization problems, and this is a practical justification of its wide use.

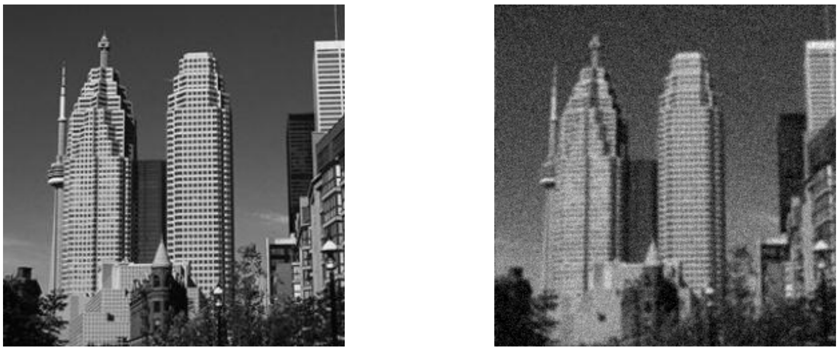
6.6 Computed examples

In this section, we evaluate experimentally the performance of the proposed model $\mathrm{TV}^{\mathrm{sv}}_{p,\alpha}$-$\mathrm{L}_2$ in (6.3), when applied to the restoration of gray level images synthetically corrupted by known blur and by AWGN. In particular, the proposed model is compared with:

- the TV-L_2 model;
- the TV_p-L_2 model, with $0 < p \leq 2$ [111];
- the HWTV-L_2, introduced in Chapter 5.

Algorithm 4: ADMM-based algorithm for the $\mathrm{TV}^{\mathrm{sv}}_{p,\alpha}$-$\mathrm{L}_2$ model

```
input:  $b \in \mathbb{R}^n$, $r > 0$, $\tau \approx 1$, $\gamma_v > 0$,$\gamma_w > 0$
output: restored image $u^*$
initialize:  set $u^{(0)} = b$, $\rho_w^{(0)} = \mathbf{0}_n$, $\rho_v^{(0)} = \mathbf{0}_{2n}$
for  $t$ = 1, 2, ... until convergence  do:
    update parameters
        · $\alpha_i^{(t)}$ by (6.8) for every $i = 1, \ldots, n$
        · $p_i^{(t)}$ by (6.9) for every $i = 1, \ldots, n$
        · $\mu^{(t)}$ by (5.35)
    update primal variables
        · $v^{(t)}$ by (6.18)
        · $w^{(t)}$ by (5.32)
        · $u^{(t)}$ by solving (5.33)
    update dual variables
        · $\rho_w^{(t)}$ by (5.25)
        · $\rho_v^{(t)}$ by (5.26)
end for
return:  $u^* = u^{(t)}$
```



(a) Original u.

(b) Corrupted b.

Figure 6.5: Original test image `skyscraper` (a) and observed image b (b) corrupted by Gaussian blur and AWGN with $\sigma = 0.05$.

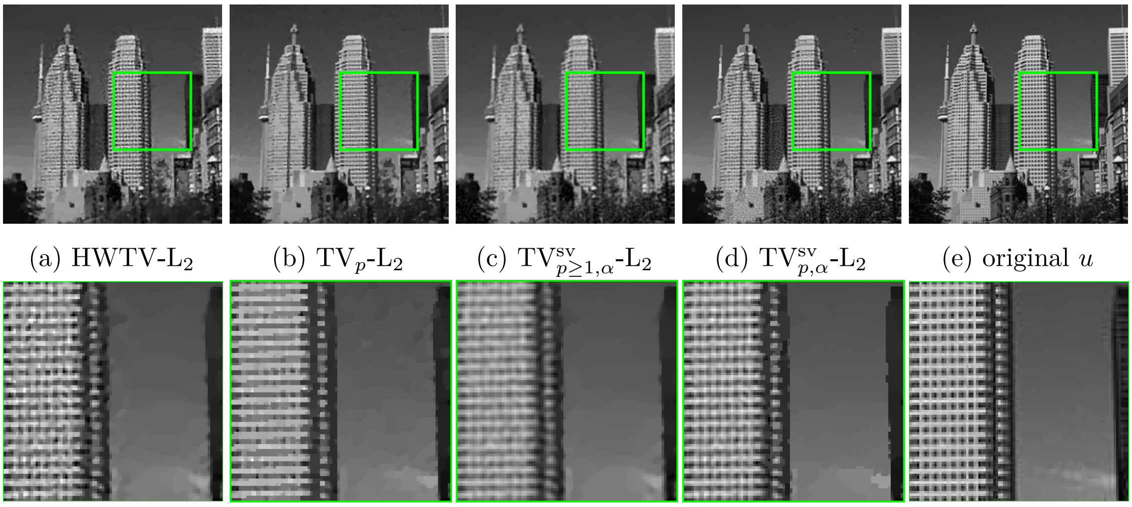
The alternating update of the maps of the parameters α and p as well as of the image u is carried out via Algorithm 4. We point out that, the local scale parameters α_i are updated by means of the closed formula in (6.8), while the local shape parameters p_i are sought as minimizers of the objective function $f(p_i)$ in (6.10) over the compact set $[\epsilon, 2]$. For each i, problem (6.10) is solved by considering a k-points discretization of $[\epsilon, 2]$ - usually $k = 5, 6$ - and computing the minimum of $f(p_i)$ over the grid. Thus, in order to improve the efficiency of Algorithm 4, the maps of the parameters are updated at each 10 or 20 iterations.
Moreover, for all the ADMM-based minimization algorithms and for all the tests, the stopping criteria detailed in (5.37) are adopted and the penalty parameters γ_v and γ_w are suitably set.

Example 1 We start proposing the same example reported in Section 9.2, namely the restoration of the test image `skyscraper` in Figure 6.5a corrupted by AWGN with standard deviation $\sigma = 0.05$ and Gaussian blur of `band`=4 and `width`=1 - see Figure 6.5b. In order to highlight the effective contribution of the space-variance of the shape parameter p, we are first performing the $\mathrm{TV}^{\mathrm{sv}}_{p,\alpha}$-$\mathrm{L}_2$ model in convex settings, that is $1 \leq p_i \leq 2$. Then, we are extending our analysis to the case $0.5 \leq p_i \leq 2$.

In Figure 6.6, we compare the proposed model with the HWTV-L_2 and the TV_p-L_2 models. We remark that here the parameter τ as well as the radius r for the space-variant models have been tuned in order to maximize the ISNR values, that are reported in Table 6.1, together with the SSIM values.

It is worth noticing that, when $p_i \geq 1$, we do not observe significant improvements with respect to the restoration via HWTV-L_2 model. This is mainly due to the fact that on the region where α assumes low values, namely the textured one, the contribution of $p > 1$ is not relevant - see Figure 6.7a-6.7b. On the other hand, ISNR and SSIM values increase when we let p_i be less than 1. The quality of the restoration improves because lower value of p are preferable on constant - or almost constant - regions, as in the case of the background in `skyscraper`.

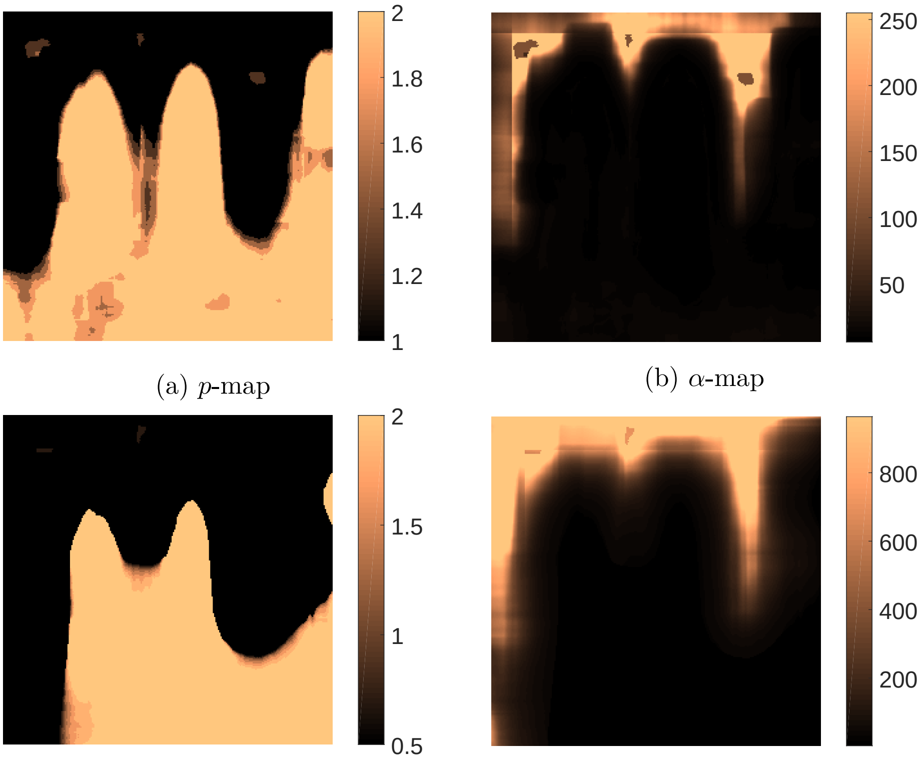
Example 2 As a second example, we consider the restoration of test image `bridge` in Figure 6.8a corrupted by Gaussian blur of `band`=4 and `width`=1 and by AWGN with standard deviation $\sigma = 0.05$ - see Figure 6.8b. The restorations obtained via the compared models are shown in



(a) HWTV-L$_2$ (b) TV$_p$-L$_2$ (c) TV$^{\text{sv}}_{p\geq 1,\alpha}$-L$_2$ (d) TV$^{\text{sv}}_{p,\alpha}$-L$_2$ (e) original u

(f) close-up of (a) (g) close-up of (b) (h) close-up of (c) (i) close-up of (d) (j) close-up of (e)

Figure 6.6: **ISNR optimization**. *First row*: reconstruction of image in Fig.6.5a by HWTV-L$_2$ ($\tau = 0.91$) (a), TV$_p$-L$_2$ ($\tau = 0.94$) (b), TV$^{\text{sv}}_{p,\alpha}$-L$_2$ with local $p_i \geq 1$ ($\tau = 0.94$, $r = 14$) (c), TV$_p, \alpha^{\text{sv}}$-L$_2$ with local $p_i \geq [0.5, 2]$ ($\tau = 0.95$, $r = 22$) (d) and original image u (e). *Second row*: close-up(s).



(a) p-map (b) α-map

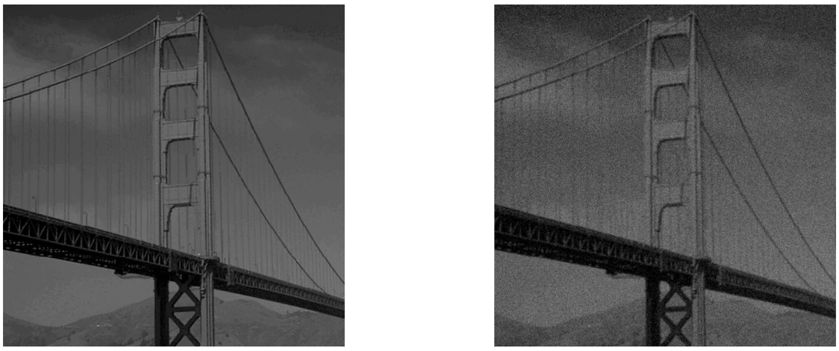
(c) p-map (d) α-map

Figure 6.7: *First row*: p-map (a) and α-map (b) with radius $r = 14$ and $p_i \in [1, 2]$. *Second row*: p-map (c) and α-map (d) with radius $r = 22$ and $p_i \in [0.5, 2]$.

Figure 6.9 and the corresponding ISNR and SSIM values are reported in Table 6.1. For all the models, the parameter τ has been set equal to 1, while for the space-variant models, namely

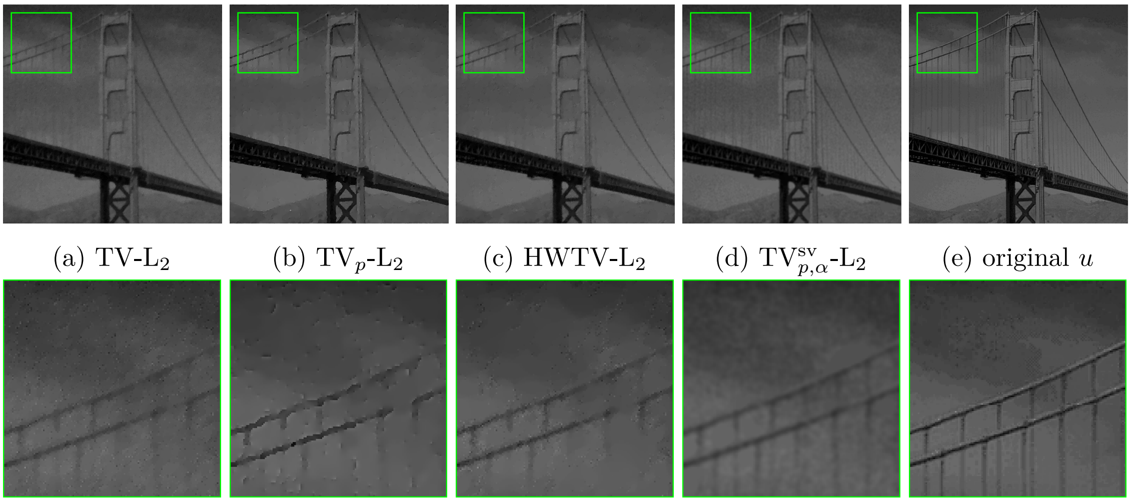
	skyscraper				bridge			
	HWTV	TV_p	$\mathrm{TV}^{\mathrm{sv}}_{p\geq 1,\alpha}$	$\mathrm{TV}^{\mathrm{sv}}_{p,\alpha}$	Tv	HWTV	TV_p	$\mathrm{TV}^{\mathrm{sv}}_{p,\alpha}$
ISNR	2.54	2.42	**2.57**	**2.98**	4.54	5.17	5.09	**5.50**
SSIM	0.79	0.81	**0.80**	**0.84**	0.71	0.77	0.78	**0.80**

Table 6.1: ISNR and SSIM values achieved by the compared models on `skyscraper` and `bridge` test images.



(a) Original u (b) Corrupted b

Figure 6.8: Original test image `bridge` (a) and observed image b corrupted by Gaussian blur and AWGN with $\sigma = 0.05$.



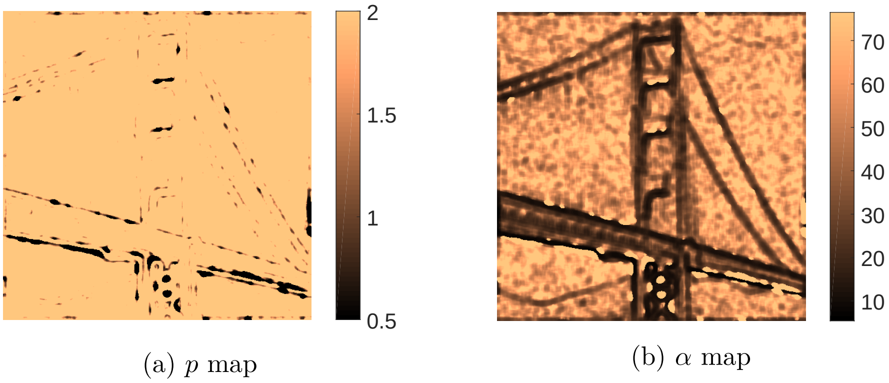
(a) TV-L$_2$ (b) TV$_p$-L$_2$ (c) HWTV-L$_2$ (d) $\mathrm{TV}^{\mathrm{sv}}_{p,\alpha}$-L$_2$ (e) original u

(f) close-up of (a) (g) close-up of (b) (h) close-up of (c) (i) close-up of (d) (j) close-up of (e)

Figure 6.9: *First row*: Restoration of `bridge` by TV-L$_2$ (a), TV$_p$-L$_2$ (b), HWTV-L$_2$ ($r = 4$) (c), $\mathrm{TV}^{\mathrm{sv}}_{p,\alpha}$-L$_2$ ($r = 4$) (d) and original u (e). The tests are all performed with $\tau = 1$. *Second row*: close-up(s).

the HWTV-L$_2$ and the $\mathrm{TV}^{\mathrm{sv}}_{p,\alpha}$-L$_2$ models, a radius $r = 4$ has been fixed. Moreover, for what concerns the p-map estimation in the $\mathrm{TV}^{\mathrm{sv}}_{p,\alpha}$-L$_2$ model, we look for $p_i \in [0.5, 2]$.

The test image is characterized by smooth regions, hence the $\mathrm{TV}^{\mathrm{sv}}_{p,\alpha}$ mainly performs a Tikhonov-type regularization, as also suggested by the output p-map in Figure 6.10a. In fact, in this case the classical TV regularization, even if locally weighted, produces staircasing in



(a) p map

(b) α map

Figure 6.10: Output maps of parameters with radius $r = 4$.

the background, since it not suited for the restoration of smooth regions.

Chapter 7

Space-variant direction regularization tuning

The space-variant regularizers WTV (5.6) and $\mathrm{TV}_{p,\alpha}^{\mathrm{sv}}$ (6.3) have been designed in order to overcome some of the shortcomings related to the adoption of a TV regularizer, mainly related to both its global weighting at any pixel as well as its global smoothness, and furthermore to its convexity, which is well-known to be easier to handle in practice, but on the other hand has been shown to be uncapable to favor sparsity as non-convex regularizers may do. Nevertheless, the regularization term proposed in Chapter 5 and Chapter 6, as well as TV, share a further limitation that is the blindness to dominant local directionality in the image. In other words, making an assumption on the distribution of the ℓ_2-norm of the gradients does not allow to include in the model information concerning the spatial distribution of the gradients themselves. In literature, several methods addressing this issue have been considered. Many of those are based on a structure-tensor modeling [159, 148, 136, 140]. Directional information have also been encoded in higher-order regularizer, such as TGV, see, e.g., [104, 126, 127]. Also, a plethora of literature addressed the problem of detecting and exploiting directional information in a multi-scale framework, where the restoration is driven by curvelets and shearlets [100, 108, 64, 86]. Nonetheless, the novelty of our contribution is represented by the statistical assumptions motivating the introduction of the proposed regularizer.

7.1 The $\mathrm{DTV}_p^{\mathrm{sv}}$-$\mathrm{L}_2$ model

In this Chapter, in order to address the linear inverse problem in (5.1), we propose a novel regularization term that inherits the space-variant contributions presented by WTV and $\mathrm{TV}_{p,\alpha}^{\mathrm{sv}}$ regularization terms, and adds one more space-variant free parameter encoding the local orientation of the gradients. We are referring to this regularizer as $\mathrm{DTV}_p^{\mathrm{sv}}$. We define it as

$$\mathrm{DTV}_p^{\mathrm{sv}}(u) := \sum_{i=1}^{n} \|\Lambda_i \mathrm{R}_{\zeta_i} (\mathrm{D}u)_i\|_2^{p_i}\,, \quad p_i > 0\,,\ \forall\, i = 1, \ldots, n. \tag{7.1}$$

For every $i = 1, 2, \ldots, n$, the weighting and rotation matrices $\Lambda_i, \mathrm{R}_{\zeta_i} \in \mathbb{R}^{2\times 2}$ are defined respectively by:

$$\Lambda_i := \begin{pmatrix} \lambda_i^{(1)} & 0 \\ 0 & \lambda_i^{(2)} \end{pmatrix}, \quad \lambda_i^{(1)} \geq \lambda_i^{(2)} > 0, \qquad \mathrm{R}_{\zeta_i} := \begin{pmatrix} \cos\zeta_i & -\sin\zeta_i \\ \sin\zeta_i & \cos\zeta_i \end{pmatrix}, \quad \zeta_i \in [0, 2\pi), \tag{7.2}$$

so that ζ_i has to be understood as the local image orientation, while the parameters $\lambda_i^{(1)}$ and $\lambda_i^{(2)}$ weight at any point the TV-like smoothing along the direction ζ_i and its orthogonal $\zeta_i^{\perp}$, respectively.

We thus define the space-variant, anisotropic (or directional) and possibly non-convex $\mathrm{DTV}_p^{\mathrm{sv}}$-$\mathrm{L}_2$ variational model for image restoration:

$$u^* \in \arg\min_{u \in \mathbb{R}^n} \left\{ \mathcal{J}(u) := \mathrm{DTV}_p^{\mathrm{sv}}(u) + \frac{\mu}{2}\|\mathrm{K}u - b\|_2^2 \right\}, . \tag{7.3}$$

The HWTV-L_2 model (5.8) and the $\mathrm{TV}_{p,\alpha}^{\mathrm{sv}}$-$\mathrm{L}_2$ model (6.2) can be also interpreted as sub-cases of the more general $\mathrm{DTV}_p^{\mathrm{sv}}$-$\mathrm{L}_2$. In fact, the generic Λ_i matrix can be rewritten as

$$\Lambda_i = \lambda_i^{(2)} \tilde{\Lambda}_i\,, \text{ where } \tilde{\Lambda}_i = \begin{pmatrix} \frac{\lambda_i^{(1)}}{\lambda_i^{(2)}} & 0 \\ 0 & 1 \end{pmatrix} .$$

The normalized matrix $\tilde{\Lambda}_i$ encodes the relative strength of the regularization along ζ_i and $\zeta_i^{\perp}$, while $\lambda_i^{(2)}$ plays the role of a scaling factor. Hence, the $\mathrm{DTV}_p^{\mathrm{sv}}$ regularizer in (7.1) takes the equivalent form

$$\mathrm{DTV}_p^{\mathrm{sv}}(u) := \sum_{i=1}^{n} \alpha_i \left\| \tilde{\Lambda}_i \mathrm{R}_{\zeta_i} (\mathrm{D}u)_i \right\|_2^{p_i}, \text{ with } \alpha_i = \left(\lambda_i^{(2)}\right)^{p_i} . \tag{7.4}$$

When $\zeta_i = 0$, $\lambda_i^{(1)} = \lambda_i^{(2)}$ for $i = 1, \ldots, n$, (7.4) reduces to the $\mathrm{TV}_{p,\alpha}^{\mathrm{sv}}$ regularizer; if, in addition, $p_i = 1$ for $i = 1, \ldots, n$, the WTV regularizer is obtained.

The $\mathrm{DTV}_p^{\mathrm{sv}}$-$\mathrm{L}_2$ can be also classified as a hybrid model, due to the presence of a global regularization parameter μ in its formulation. Moreover, in analogy with the $\mathrm{TV}_{p,\alpha}^{\mathrm{sv}}$-$\mathrm{L}_2$ model, in model (7.3) the non-convexity can possibly arise whenever $0 < p_i < 1$.

The proposed $\mathrm{DTV}_p^{\mathrm{sv}}$ regularizer (7.1) is highly flexible as it potentially adapts to local smoothness and directional properties of the image at hand, provided that a reliable estimation of the parameters $\lambda_i^{(1)}, \lambda_i^{(2)}, \zeta_i$ and p_i is given. In comparison to the previously introduced regularization terms, the $\mathrm{DTV}_p^{\mathrm{sv}}$ accommodates further local directional information, which can significantly improve the restoration results in the case, for instance, of textured and/or high-detailed images. The statistical rational of our approach relies on a prior assumption on the distribution of the gradient magnitudes of the desired image u which we assume to be space-variant and locally drawn from a Bivariate Generalized Gaussian Distribution (BGGD) [12, 144, 143].

As in the previous chapter, an effective estimation procedure for the estimation of the unknown parameters involved in the expression of the $\mathrm{DTV}_p^{\mathrm{sv}}$ regularizer (7.1) will be discussed as well. More specifically, we will propose a robust parameter estimation procedure based on a hierarchical modeling and on non-informative hyperpriors, reducing dramatically the number of parameters to estimate. As far as the estimation of the global parameter μ is concerned, we are still resorting to the discrepancy principle.

7.2 Model derivation via MAP

We start modeling the joint distribution of the two partial derivatives of the gradient vector $(\mathrm{D}u)_i$ at any pixel by a Bivariate Generalized Gaussian Distribution (BGGD) [12]. Namely, for all $i = 1, \ldots, n$ we assume that

$$\mathrm{P}((\mathrm{D}u)_i; p_i, \Sigma_i) = \frac{1}{2\pi|\Sigma_i|^{1/2}} \, \frac{p_i}{\Gamma(2/p_i)\, 2^{2/p_i}} \exp\left(-\frac{1}{2}((\mathrm{D}u)_i^T \Sigma_i^{-1} (\mathrm{D}u)_i)^{p_i/2} \right), \tag{7.5}$$

where Γ stands for the Gamma function, the *covariance matrices*

$$\Sigma_i = \begin{pmatrix} (\sigma_1)_i & (\sigma_3)_i \\ (\sigma_3)_i & (\sigma_2)_i \end{pmatrix} \in \mathbb{R}^{2\times 2}, \quad i = 1, \ldots, n,$$

are symmetric positive definite with determinant $|\Sigma_i|$ and $p_i/2$ is often referred to as *shape parameter*. Note that when in (7.5) $p_i = 2$ for every $i = 1, \ldots, n$, then the BGGD reduces to a standard bivariate Gaussian distribution with pixel-wise covariance matrices Σ_i. As suggested in [144, 143, 129], it is possible to decouple the spread and the directionality of the BGGD by introducing a further *scale parameter* $m > 0$, so that (7.5) takes the following form:

$$\mathrm{P}(x; p, \Sigma, m) = \frac{1}{\pi\Gamma\big(\frac{2}{p}\big)2^{\frac{2}{p}}} \, \frac{p}{2m|\Sigma|^{\frac{1}{2}}} \exp\bigg(-\frac{1}{2m^{p/2}}(x^T\Sigma^{-1}x)^{p/2} \bigg),$$

where, for the sake of better readability, we continue using the symbol Σ_i to denote the scatter matrix of the local distribution.

Proceeding similarly as in the previous chapters, we can then deduce the expression of the corresponding prior under such assumption. It reads:

$$\mathrm{P}(u) = \frac{1}{Z} \exp\bigg(-\frac{1}{2}\sum_{i=1}^{n} \frac{1}{m_i^{p_i/2}} \left((\mathrm{D}u)_i^T \Sigma_i^{-1} (\mathrm{D}u)_i \right)^{p_i/2} \bigg) \tag{7.6}$$

The symmetric positive definite matrices Σ_i contain information on both the directionality and the relative elongation of the elliptic level curves of the BGGD at pixel i. To see that explicitly, we consider their eigenvalue decomposition:

$$\Sigma_i = \mathrm{V}_i^T \mathrm{E}_i \mathrm{V}_i, \quad \mathrm{E}_i = \begin{pmatrix} e^{(1)}{}_i & 0 \\ 0 & e_i^{(2)} \end{pmatrix}, \quad e^{(1)}{}_i \geq e_i^{(2)} > 0, \quad \mathrm{V}_i^T \mathrm{V}_i = \mathrm{V}_i \mathrm{V}_i^T = \mathrm{I},$$

where for every $i = 1, \ldots, n$, $e^{(1)}{}_i, e_i^{(2)}$ are the (positive) eigenvalues of Σ_i and V_i is the orthonormal (rotation) modal matrix. We then rewrite the terms in the sum appearing in (7.6) as

$$\left((\mathrm{D}u)_i^T \Sigma_i^{-1} (\mathrm{D}u)_i \right)^{\frac{p_i}{2}} = \left((\mathrm{D}u)_i^T \mathrm{V}_i^T \mathrm{E}_i^{-1} \mathrm{V}_i (\mathrm{D}u)_i \right)^{\frac{p_i}{2}} = \left\| \mathrm{E}_i^{-1/2} \mathrm{V}_i \, (\mathrm{D}u)_i \right\|_2^{p_i},$$

whence by setting

$$\Lambda_i := \mathrm{E}_i^{-1/2}, \quad \mathrm{R}_{\zeta_i} := \mathrm{V}_i, \tag{7.7}$$

and after recalling the definition of the $\mathrm{DTV}_p^{\mathrm{sv}}$ regularizer given in (7.4), we observe that the prior in (7.6) can indeed be expressed as:

$$\mathrm{P}(u \mid \theta_{\mathrm{pr}}) = \frac{1}{Z} \exp\left(-\frac{1}{2} \mathrm{DTV}_p^{\mathrm{sv}}(u) \right), \tag{7.8}$$

where Z is the normalization constant and $m_i^{-p_i/2}$ plays the role of the local regularization weight α_i in (7.4). The vector of unknown parameters defining the prior in (7.8) reads

$$\theta_{\mathrm{pr}} = \begin{pmatrix} \theta_{\mathrm{pr}_1} \\ \theta_{\mathrm{pr}_2} \\ \cdots \\ \theta_{\mathrm{pr}_n} \end{pmatrix} \in \mathbb{R}^{n\times 4}\,, \quad \text{with} \quad \theta_{\mathrm{pr}_i} = \left(m_i, p_i, \zeta_i, e_i^{(1)}\right)\,,$$

where observe that the angle ζ_i and the maximum eigenvalue $e_i^{(1)}$ uniquely determine the local scatter matrix Σ_i when, a typical normalization constraint on the trace of Σ is imposed [144, 143, 129], namely

$$\mathrm{tr}(\Sigma_i) = e_i^{(1)} + e_i^{(2)} = d = 2\,,$$

with d being the dimension of the ambient space. By plugging the expression of the Gaussian likelihood and the BGGD prior (7.8) in the MAP inference formula (3.20) and after dropping the constant terms, we finally obtain the $\mathrm{DTV}_p^{\mathrm{sv}}$-$\mathrm{L}_2$ image restoration model (7.3) for blur and AWGN removal by setting $\mu = 1/\sigma^2$.

Also in this case, for the solution of the minimization problem in (7.3) and for the combined estimation of the unknown parameters, we are resorting to an alternated scheme as the ones proposed in the previous chapters. The maps of the four space-variant parameters, namely m, p, ζ and $e^{(1)}$, are updated and integrated in the model, so as to provide a more flexible regularization in the u-update.

Algorithm 5: IAS with non-informative hypeprior for the $\mathrm{DTV}_p^{\mathrm{sv}}$-$\mathrm{L}_2$ model

```
input: observed image $b \in \mathbb{R}^n$
output: restored image $u^*$
1. initialize: set $u^{(0)} = b$
2. for $t = 1, 2, \ldots$ until convergence do:
3.     · update $\left(m^{(t)}, p^{(t)}, \zeta^{(t)}, \left(e^{(1)}\right)^{(t)}\right) \in \arg\min \left\{-\log \mathrm{P}(u^{(t-1)} \mid m, p, \zeta, e^{(1)})\right\}$ ,
4.     · update $u^{(t)} \in \arg\min_{u\in\mathbb{R}^n} \Big\{ -\log \mathrm{P}\left(\mathrm{u} \mid \mathrm{m}^{(\mathrm{t})}, \mathrm{p}^{(\mathrm{t})}, \zeta^{(\mathrm{t})}, \left(\mathrm{e}^{(1)}\right)^{(\mathrm{t})}\right)$
           $-\log \mathrm{P}\left(b \mid u^{(t)}, p^{(t)}, \zeta^{(t)}, \left(e^{(1)}\right)^{(t)}\right) \Big\}$
5. end for
6. return: $u^* = u^{(t)}$
```

7.3 Automatic estimation of the $\mathrm{DTV}_p^{\mathrm{sv}}$ parameters

In this section, start addressing the parameter estimation problem, discussing in particular the constraint set for the parameters that are going to be estimated by means of the introduction of non-informative hyperpriors. We also remark that, unlike the previous cases, here the unknown entries of θ_{pr} will not be directly estimated but we will introduce some auxiliary variables. More in detail, we start addressing the problem of estimating the local scale parameter m_i, the shape parameter p_i and the entries of the scatter matrix Σ_i. By dropping for a moment the subscription, we observe that the requirement for the scatter matrix Σ to be symmetric positive definite means:

$$\Sigma = \begin{bmatrix} \sigma_1 & \sigma_3 \\ \sigma_3 & \sigma_2 \end{bmatrix} \qquad \text{with} \qquad \begin{cases} \sigma_1 > 0 \\ |\Sigma| = \sigma_1\sigma_2 - \sigma_3^2 > 0 \end{cases} .$$

By imposing condition (7.2) on the trace of the scatter matrix, we easily get the following expression of the constraint set C_{pr} for the parameters $p, m, \sigma_1, \sigma_2, \sigma_3$ to be well defined:

$$C_{\mathrm{pr}} := \begin{cases} p > 0 \\ m > 0 \\ \mathrm{tr}\,(\Sigma) = \sigma_1 + \sigma_2 = 2 \\ \sigma_1\sigma_2 - \sigma_3^2 > 0 \end{cases} \qquad \longrightarrow \qquad C_{\mathrm{pr}} = \begin{cases} p > 0 \\ m > 0 \\ \sigma_1^2 + \sigma_3^2 - 2\sigma_1 < 0. \end{cases} \tag{7.9}$$

The set C_{pr} is an open (unbounded) semi-cylinder in $\mathbb{R}^4$. After a change of coordinates which shifts the center of the circle in the $\sigma_1 - \sigma_3$ plane to the origin, we obtain the following expression of Σ^{-1} expressed in the new coordinates,

$$\tilde{\sigma}_1 := 1 - \sigma_1 \qquad \longrightarrow \qquad \Sigma^{-1} = \frac{1}{1 - \tilde{\sigma}_1^2 - \sigma_3^2} \begin{bmatrix} 1 + \tilde{\sigma}_1 & -\sigma_3 \\ -\sigma_3 & 1 - \tilde{\sigma}_1 \end{bmatrix} . \tag{7.10}$$

To avoid heavy notation, we will still denote in the following by σ_1 the same variable after this change of coordinates.

7.3.1 Estimation of the BGGD parameters via non-informative hyperpriors

In order to estimate the local parameters m, p, σ_1, σ_3, or equivalently, θ_{pr_i}, we are adopting the same strategy introduced in the previous chapter. For every $i = 1, \ldots, n$, let $\mathcal{S}_i = \{x_{i,j}\}_{j=1}^N$ be a set of N samples drawn from a BGGD. The samples $x_{i,j} \in \mathbb{R}^2$ represent the gradients in pixels belonging to a neighborhood $\mathcal{C}_i^r$ of radius r of pixel i. In formula, $x_{i,j} = (\mathrm{D}u)_j$, with $u_j \in \mathcal{C}_i^r$. For each i, we are interested in solving a problem of the form

$$(p^{(t)}, \sigma_1^{(t)}, \sigma_3^{(t)}, m^{(t)}) \longleftarrow \arg \min_{\theta \,\in\, C_{\mathrm{pr}}} \; \{\mathcal{F}(\theta; x) := -\log\; \mathrm{P}(\mathcal{S} \mid \theta)\} ,$$

where we dropped the subscription for the sake of better readability. The local likelihood function, under the hypothesis of independence of the samples in $\mathcal{S}$ reads

$$\begin{aligned} \mathrm{P}(\mathcal{S} \mid \theta) &= \prod_{x_j \in \mathcal{C}^r} \mathrm{P}(x_j \mid \theta) = \prod_{j=1}^N \mathrm{P}(x_j \mid \theta) \\ &= \left[\frac{1}{|\Sigma|^{1/2}} \frac{p}{2\pi\Gamma(\frac{2}{p})2^{2/p}m} \right]^N \exp\Big(- \frac{1}{2m^{p/2}} \sum_{j=1}^N (x_j^T \Sigma^{-1} x_j)^{p/2} \Big) . \end{aligned}$$

By recalling the fundamental property of Gamma function $\Gamma(z+1) = z\Gamma(z)$ for every $z \in \mathbb{R}$, we deduce:

$$\begin{aligned}\mathcal{F}(p,\sigma_1,\sigma_3,m;x) &= -\left[N\log\left(\frac{1}{|\Sigma|^{1/2}}\frac{1}{\pi\Gamma(\frac{2}{p}+1)2^{2/p}m}\right) - \frac{1}{2m^{p/2}}\sum_{j=1}^{N}(\,x_j^T\,\Sigma^{-1}\,x_j\,)^{p/2}\right]\\ &= N\log\left(|\Sigma|^{1/2}\pi\Gamma\left(\frac{2}{p}+1\right)2^{2/p}\right) + N\log m + \frac{1}{2m^{p/2}}\sum_{j=1}^{N}(\,x_j^T\,\Sigma^{-1}\,x_j\,)^{p/2}.\end{aligned}$$

Note that $\mathcal{F}$ is differentiable on C_{pr}. Therefore, by simply imposing the first order optimality condition for m, we get the following closed formula:

$$\frac{\partial\mathcal{F}}{\partial m} = \frac{N}{m} - \frac{p}{4m^{\frac{p}{2}+1}}\sum_{j=1}^{N}(x_j^T\Sigma^{-1}x_j)^{p/2} \quad\longrightarrow\quad m = \left(\frac{p}{4N}\sum_{j=1}^{N}(x_j^T\Sigma^{-1}x_j)^{p/2}\right)^{\frac{2}{p}}. \tag{7.11}$$

We now substitute this formula in the expression of $\mathcal{F}$, thus getting:

$$\mathcal{F}(p,\sigma_1,\sigma_3;x) = N\log\left(|\Sigma|^{1/2}\pi\Gamma\left(\frac{2}{p}+1\right)2^{2/p}\right) + \frac{2N}{p}\log\left(\frac{p}{4N}\sum_{j=1}^{N}(x_j^T\Sigma^{-1}x_j)^{p/2}\right) + \frac{2N}{p}. \tag{7.12}$$

By making explicit the dependence of $\mathcal{F}$ on the entries of Σ, and recalling that σ_2 can be expressed in terms of σ_1, we have that (7.12) turns into:

$$\begin{aligned}\mathcal{F}(p,\sigma_1,\sigma_3;x) = N\log\left(\frac{1}{|\Sigma|^{1/2}}\pi\Gamma\left(\frac{2}{p}+1\right)2^{2/p}\right) + \frac{2N}{p} + \frac{2N}{p}\log\frac{p}{4N}\\ + \frac{2N}{p}\log\left(\sum_{j=1}^{N}(\sigma_2 x_{j,1}^2 + \sigma_1 x_{j,2}^2 - 2\sigma_3 x_{j,1}x_{j,2})^{p/2}\right).\end{aligned} \tag{7.13}$$

We now study the behavior of $\mathcal{F}$ expressed as above as (p,σ_1,σ_3) approach the boundary of the set C_{pr} defined in (7.9). Thanks to the formula for m derived in (7.11), we start noticing that C_{pr} can be expressed in fact as a subset in $\mathbb{R}^3$ defined by the variables p, σ_1 and σ_3 only. By further switching to polar coordinates in the $\sigma_1 - \sigma_3$ plane, we get:

$$(\sigma_1,\sigma_3) = \varrho(\cos\phi,\sin\phi), \quad 0 \le \varrho < 1, \quad \phi \in [0,2\pi),$$

so that the matrices in (7.10) take the following form:

$$\Sigma = \begin{bmatrix}1-\varrho\cos\phi & \varrho\sin\phi\\ \varrho\sin\phi & 1+\varrho\cos\phi\end{bmatrix}, \qquad \Sigma^{-1} = \frac{1}{1-\varrho^2}\begin{bmatrix}1+\varrho\cos\phi & -\varrho\sin\phi\\ -\varrho\sin\phi & 1-\varrho\cos\phi\end{bmatrix}, \tag{7.14}$$

and the functional $\mathcal{F}$ in (7.13) becomes:

$$\begin{aligned}\mathcal{F}(p,\phi,\varrho;x) = N\log\left(\Gamma\left(\frac{2}{p}+1\right)\frac{\pi}{\sqrt{1-\varrho^2}}\left(\frac{p}{2N}\right)^{2/p}\right) + \frac{2N}{p} + \frac{2N}{p}\log\frac{p}{4N}\\ + \frac{2N}{p}\log\left[\sum_{j=1}^{N}((1+\varrho\cos\phi)x_{j,1}^2 + (1-\varrho\cos\phi)x_{j,2}^2 - 2\varrho\sin\phi\; x_{j,1}x_{j,2})^{p/2}\right]\end{aligned} \tag{7.15}$$

As a conclusion, the local estimation problem we are interested in takes the form of the following constrained optimization problem

$$(p^{(t)},\phi^{(t)},\varrho^{(t)}) \in \arg\min_{\substack{p\in(0,\infty),\\ \phi\in[0,2\pi),\\ \varrho\in[0,1)}} \mathcal{F}(p,\phi,\varrho). \tag{7.16}$$

Note that, since the problem (7.16) is formulated over a non-compact set of $\mathbb{R}^3$, the existence of a solution is in general not guaranteed.

7.3.2 Reformulation on a compact set

One possible way to overcome the problem of non-compactness consists in characterizing explicitly the configurations of the samples $\mathcal{C}^r$ for which the functional $\mathcal{F}$ in (7.15) does not attain its minimum inside C_{pr}. To do so, let us first rename the last term in (7.15) as:

$$A(\phi,\varrho) := \frac{2N}{p}\log\left[\sum_{i=j}^{N}((1+\varrho\cos\phi)x_{j,1}^2 + (1-\varrho\cos\phi)x_{j,2}^2 - 2\varrho\sin\phi\ x_{j,1}x_{j,2})^{p/2}\right].$$

For any $p\in(0,+\infty)$, if $A(\phi,\varrho)$ is bounded as $\varrho\to 1^-$, then the functional $\mathcal{F}$ in (7.15) tends to $+\infty$ and the minimum is necessarily attained in the interior of C_{pr}. However, if $A(\phi,\varrho)$ is unbounded as $\varrho\to 1^-$, nothing can be said about the behavior of $\mathcal{F}$ at the boundary and, as a consequence, nothing can be said about its minima. In particular, in this situation there may exist one or multiple configurations of the samples $x_1,\dots,x_N\in\mathcal{C}^r$ for which $\mathcal{F}$ tends to $-\infty$ at the boundary. In order to characterize such configurations, note that as $\varrho\to 1^-$ we have that by continuity:

$$A(\phi,\varrho)\to\frac{2N}{p}\log\left[\sum_{j=1}^{N}(\sqrt{1+\cos\phi}\ x_{j,1}-\sqrt{1-\cos\phi}\ x_{j,2})^p\right],$$

which tends to $-\infty$ if and only if the argument of the logarithm tends to zero, i.e. when

$$x_{j,2}=\sqrt{\frac{\sqrt{1+\cos\phi}}{\sqrt{1-\cos\phi}}}\,x_{j,1},\qquad \forall j=1,...,N. \tag{7.17}$$

This situation corresponds to the very particular case when the samples x_j lie all on the line passing through the origin with slope $\sqrt{\frac{\sqrt{1+\cos\phi}}{\sqrt{1-\cos\phi}}}$.

A possible way to guarantee the existence of solutions of the problem (7.16) is to re-formulate the problem over a compact subset of $\mathbb{R}^3$. Although this may sound a little bit artificial, note that for imaging applications such assumption makes perfect sense for different reasons. Firstly, as far as the range for the parameter ϱ is concerned, the degenerate configurations (7.17) happening as ϱ approaches 1^- are easily detectable in a pre-processing step and, in practice, very unlikely for natural images since they would correspond to situations where gradient components are linearly correlated for any sample $j=1,\dots,N$. Therefore, provided we can perform such preliminary check, the case $\varrho=1$ becomes admissible since no other possible configurations are allowed under this choice.

Secondly, regarding the admissible values for p, we notice that the more we enforce sparsity (i.e. the closer p is to zero), the more the BGGD will tend to a Dirac delta distribution, making the estimation of local anisotropy in a neighborhood of the point considered impossible (see Section 7.3 for more details). Hence, in analogy with Chapter 6, the exponent p is thought as confined in the closed interval $[\epsilon,2]$, with $\epsilon>0$.

After this observation, we can then reformulate the problem (7.16) as follows

$$(p^*,\phi^*,\varrho^*)\in\min_{p,\phi,\varrho}\mathcal{F}(p,\phi,\varrho;x) \tag{7.18}$$

$$\text{s.t.}\quad p\in[\varepsilon,2],\ \ 0\le\varrho\le 1,\ \ 0\le\phi\le 2\pi,$$

where now the constraint set is compact, which, combined with the continuity of $\mathcal{F}$, guarantees that the minimization problem admits a minimum.

Summarizing, for each pixel $i = 1, \ldots, n$, the triplet of estimated parameters $(p_i^*, \phi_i^*, \varrho_i^*)$ is involved in the computation of the matrices $\Lambda_i, \mathrm{R}_{\zeta_i}$ defining the regularizer in (7.1). Relying on (7.14), the eigenvalues $e_i^{(1)}, e_i^{(2)}$ can be easily computed. Observe that, due to the normalization condition on the trace introduced in (7.9), the minimum eigenvalue $e_i^{(2)}$ can be directly derived by the maximum eigenvalue $e_i^{(1)}$:

$$e_i^{(1)} = 1 + \varrho_i, \quad e_i^{(2)} = 2 - e_i^{(1)} = 1 - \varrho_i.$$

Therefore, recalling (7.7), the matrix Λ_i is obtained as follows:

$$\Lambda_i := \begin{pmatrix} \lambda_i^{(1)} & 0 \\ 0 & \lambda_i^{(2)} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{e_i^{(1)}}} & 0 \\ 0 & \frac{1}{\sqrt{e_i^{(2)}}} \end{pmatrix}. \tag{7.19}$$

Once $e_i^{(1)}$ is available, its corresponding eigenvector $(v_1)_i$, satisfying $\Sigma_i (v_1)_i = e^{(1)}{}_i (v_1)_i$, can be further calculated using the formula

$$(v_1)_i = \sqrt{\frac{1 + \cos\phi_i}{2}} \begin{bmatrix} \frac{\sin\phi_i}{\sqrt{1+\cos\phi_i}} \\ 1 \end{bmatrix}.$$

As a consequence, the local angle ζ_i describing the local orientation is computed by

$$\zeta_i = \arctan \frac{\sqrt{1 + \cos\phi_i}}{\sin\phi_i},$$

and the rotation matrix R_{ζ_i} is given as in (7.2).

In order to get a better insight on the space-variant and local flexibility of the proposed regularizer, it is helpful to represent the estimated BGGD to visualize its shape in the plane $((\mathrm{D}_h u)_i, (\mathrm{D}_v u)_i)$. To draw the corresponding level curves, we only need the maximum eigenvalue $e_i^{(1)}$ and the rotation angle ζ_i. Such curves are the ellipses having semi-axes a_i, b_i, and eccentricity ϵ_i given by:

$$a_i := \sqrt{e^{(1)}{}_i}, \quad b_i := \sqrt{e_i^{(2)}}, \quad \epsilon_i := \frac{\sqrt{{a_i}^2 - {b_i}^2}}{a_i} = \frac{\sqrt{e^{(1)}{}_i - e_i^{(2)}}}{\sqrt{e^{(1)}{}_i}}.$$

An illustrative drawing of the anisotropy ellipses described above is reported in Figure 7.1.

7.4 Existence and uniqueness of solutions

By applying again the lemma proved in [50, Lemma 2.7.1], whose statement has been given in Section 6.4, we will prove that existence of global minimizers for model (7.3) is guaranteed.

Proposition 6. *The* $\mathrm{DTV}_p^{\mathrm{sv}}$-$\mathrm{L}_2$ *functional* $\mathcal{J} : \mathbb{R}^n \to \mathbb{R}$ *defined in* (7.3) *is continuous, bounded from below by zero and coercive, hence it admits global minimizers.*

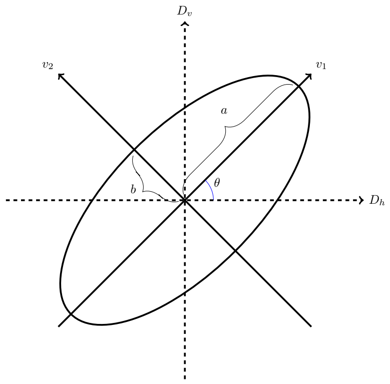


Figure 7.1: Representation of anisotropy ellipses describing BGGD level lines in the plane $\mathrm{D}_h - \mathrm{D}_v$ in terms of the eigenvalues and eigenvectors of the estimated matrix Σ.

Proof. Let $\mathrm{A}_1 \in \mathbb{R}^{2n\times n}$ be the matrix defined by

$$\mathrm{A}_1 = \mathrm{L}\mathrm{D}, \quad \mathrm{L} = \mathrm{diag}\left(\mathrm{L}_1, \mathrm{L}_2, \ldots, \mathrm{L}_n\right), \quad \mathrm{L}_i = \Lambda_i \mathrm{R}_{\zeta_i} \in \mathbb{R}^2, \;\; i = 1, \ldots, n, \tag{7.20}$$

with $\Lambda_i, \mathrm{R}_{\zeta_i} \in \mathbb{R}^2$ the full rank matrices in (7.2) and $\mathrm{D} \in \mathbb{R}^{2n\times n}$ a finite difference operator discretizing the image gradient, let $\mathrm{A}_2 = \mathrm{K}$, and let $f_1 : \mathbb{R}^{2n} \to \mathbb{R}$, $f_2 : \mathbb{R}^n \to \mathbb{R}$ be the functions defined by

$$\begin{aligned} f_1(y) &:= \textstyle\sum_{i=1}^{n} \left\|(y_{2i-1}, y_{2i})\right\|_2^{p_i}, & y \in \mathbb{R}^{2n}, \\ f_2(z) &:= \frac{\mu}{2}\left\|z - g\right\|_2^2, & z \in \mathbb{R}^n. \end{aligned} \tag{7.21}$$

Then, the $\mathrm{DTV}_p^{\mathrm{sv}}$-$\mathrm{L}_2$ energy functional in (7.3) can be written as

$$\mathcal{J}(u) = f_1(\mathrm{A}_1 u) + f_2(\mathrm{A}_2 u). \tag{7.22}$$

As the block diagonal matrix L in (7.20) has full rank (all matrices L_i have full rank), the linear operator A_1 has the same null space as the discrete gradient operator D. It follows that

$$\left(\mathrm{null}(\mathrm{A}_1) = \mathrm{null}(\mathrm{D}) = \mathrm{span}(\mathrm{I}_n)\right) \,\cap\, \left(\mathrm{null}(\mathrm{A}_2) = \mathrm{null}(\mathrm{K})\right) \;=\; \{\mathbf{0}_n\}\,,$$

in fact constant images do not belong to the null space of the linear blur operator K. Furthermore, functions f_1 and f_2 in (7.21) are clearly continuous, bounded from below by zero and coercive. It thus follows from Lemma 1 that the $\mathrm{DTV}_p^{\mathrm{sv}}$-$\mathrm{L}_2$ functional $\mathcal{J}$ in (7.22) is continuous, bounded from below by zero and coercive, hence it admits at least one global minimizer. □

Uniqueness of solutions is in general not guaranteed. However, under the assumption of analytically non-singular blur matrix K, if the functional is strictly convex, this trivially holds.

Corollary 2. *Let $\mathcal{J} : \mathbb{R}^n \to \mathbb{R}$ be the* $\mathrm{DTV}_p^{\mathrm{sv}}$-$\mathrm{L}_2$ *functional defined in* (7.3)*. If $p_i \geq 1$ for every $i = 1, \ldots, n$, then $\mathcal{J}$ is strongly convex. Hence it admits a unique global minimizer.*

Note, however, here we are more interested in the non-convex case, e.g. when there exists at least one $i \in \{1, \ldots, n\}$ such that $p_i < 1$, since in this better regularization properties are enforced in $\mathrm{DTV}_p^{\mathrm{sv}}$-$\mathrm{L}_2$. Therefore, in our applications uniqueness in general will not be guaranteed and we will be generally dealing with the case of local minima.

7.5 ADMM optimization

Consider the u-update step in the outer scheme in Algorithm 5,

$$u^{(t)} \in \arg\min_{u\in\mathbb{R}^n} \left\{ \mathcal{J}(u) := \sum_i^n \|\Lambda_i \mathrm{R}_{\zeta_i}(\mathrm{D}u)_i\|_2^{p_i} + \frac{\mu^{(t)}}{2}\|\mathrm{K}u - b\|_2^2 \right\},$$

where the parameters defining the local matrices $\Lambda_i, \mathrm{R}_{\zeta_i}$ and the local shape parameters p_i are considered fixed at iteration t, e.g. their iteration has already been performed.. Also here, we resort to ADMM algorithm. First, we introduce two auxiliary variables $w \in \mathbb{R}^n$ and $v \in \mathbb{R}^{2n}$ and rewrite model (7.3) in the following equivalent constrained form:

$$\begin{aligned} \{u^*, w^*, v^*\} \leftarrow \arg\min_{u,w,v} \left\{ \sum_{i=1}^n \|\Lambda_i \mathrm{R}_{\zeta_i} v_i\|_2^{p_i} + \frac{\mu}{2}\|w\|_2^2 \right\}, \\ \text{subject to} : w = \mathrm{K}u - b\,, \quad v = \mathrm{D}u\,, \end{aligned} \tag{7.23}$$

where $\mathrm{D} := (\mathrm{D}_h^T, \mathrm{D}_v^T)^T \in \mathbb{R}^{2n\times n}$ and $v_i := \left((\mathrm{D}_h u)_i\,,\,(\mathrm{D}_v u)_i\right)^T \in \mathbb{R}^2$.

The augmented Lagrangian functional is defined as follows:

$$\begin{aligned} \mathcal{L}(u, w, v; \rho_w, \rho_v) : =& \sum_{i=1}^n \|\Lambda_i \mathrm{R}_{\zeta_i} v_i\|_2^{p_i} + \frac{\mu}{2}\|w\|_2^2 - \langle \rho_v, v - Du \rangle + \frac{\gamma_v}{2}\|v - Du\|_2^2 \\ & - \langle \rho_w, w - (\mathrm{K}u - b) \rangle + \frac{\gamma_w}{2}\| w - (\mathrm{K}u - b)\|_2^2\,, \end{aligned} \tag{7.24}$$

where $\gamma_w, \gamma_v > 0$ are the scalar penalty parameters, while $\rho_w \in \mathbb{R}^n$, $\rho_v \in \mathbb{R}^{2n}$ are the vectors of Lagrange multipliers associated with the linear constraints $w = \mathrm{K}u - b$ and $v = \mathrm{D}u$ in (7.23), respectively. The scheme alternating the maximization with respect to the primal variables and the minimization with respect to the dual variables reads as

$$v^{(t)} \in \arg\min_{v\in\mathbb{R}^{2n}} \mathcal{L}(u^{(t-1)}, w^{(t-1)}, v; \rho_w^{(t-1)}, \rho_v^{(t-1)})\,, \tag{7.25}$$

$$w^{(t)} \in \arg\min_{w\in\mathbb{R}^{n}} \mathcal{L}(u^{(t-1)}, w, v^{(t)}; \rho_w^{(t-1)}, \rho_v^{(t-1)})\,, \tag{7.26}$$

$$u^{(t)} \in \arg\min_{u\in\mathbb{R}^{n}} \mathcal{L}(u, w^{(t)}, v^{(t)}; \rho_w^{(t-1)}, \rho_v^{(t-1)})\,, \tag{7.27}$$

$$\rho_w^{(t)} \in \rho_w^{(t-1)} - \gamma_w\left(w^{(t)} - (\mathrm{K}u^{(t)} - b)\right),$$

$$\rho_v^{(t)} \in \rho_v^{(t-1)} - \gamma_v\left(v^{(t)} - \mathrm{D}u^{(t)}\right).$$

We notice that sub-problems (7.27) and (7.26) for the primal variables u and w admit solutions based on formulas given in Section 5.5 for identical sub-problems. Moreover, for the update of the global regularization parameter μ, we resort again to the discrepancy principle, as already detailed in Chapter 5.

We thus focus on the minimization sub-problem for v in (7.25).

Sub-problem for v After simple algebraic manipulations, (7.25) can be re-written as follows:

$$v^{(t)} \in \arg\min_{v\in\mathbb{R}^{2n}} \sum_{i=1}^n \left\{ \|\Lambda_i \mathrm{R}_{\zeta_i} v_i\|_2^{p_i} + \frac{\gamma_v}{2}\left\| v_i - \left(\left(\mathrm{D}u^{(t-1)}\right)_i + \frac{1}{\gamma_v}\left(\rho_v^{(t-1)}\right)_i\right)\right\|_2^2 \right\}.$$

Solving the $2n$-dimensional minimization problem above is thus equivalent to solve the n following independent 2-dimensional problems:

$$v_i^{(t)} \in \arg\min_{v_i \in \mathbb{R}^2} \left\{ \|\Lambda_i \mathrm{R}_{\zeta_i} v_i\|_2^{p_i} + \frac{\gamma_v}{2} \left\| v_i - q_i^{(t-1)} \right\|_2^2 \right\}, \quad i = 1, \ldots, n\,, \tag{7.28}$$

where the vectors $q_i^{(t-1)} \in \mathbb{R}^2$ are defined explicitly at any iteration by

$$q_i^{(t-1)} \;:=\; \left(\mathrm{D}u^{(t-1)}\right)_i + \frac{1}{\gamma_v}\left(\rho_v^{(t-1)}\right)_i \;, \quad i = 1, \ldots, n\,.$$

We observe that the solutions of the n bivariate optimization problems in (7.28) requires the computation of a special proximal mapping-type operator, the only difference being the possible non convexity considered. We dedicate the following Section 7.5.1 to carefully discuss the solution of this optimization problem and show that it can be eventually re-written as a one-dimensional optimization problem and thus solved efficiently.

7.5.1 A non-convex proximal mapping solving

In this section, we describe a novel result in multi-variate non-convex proximal calculus which is crucial to solve efficiently the problem (7.28). Such problem can be interpreted as the calculation of a non-convex proximal mapping, see [85]. We then start recalling its definition.

Definition 3 (proximal map for non-convex functions)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a proper, lower semi-continuous and possibly non-convex function and let $\gamma > 0$. The proximal map of f with parameter γ is the set-valued function* $\mathrm{prox}_{\gamma f} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ *defined for any $q \in \mathbb{R}^n$ by:*

$$\mathrm{prox}_{\gamma f}(q) := \arg\min_{v \in \mathbb{R}^n} \left\{ f(v) + \frac{\gamma}{2} \|v - q\|_2^2 \right\}. \tag{7.29}$$

Note that under such definition the set $\mathrm{prox}_{\gamma f}(q)$ is in general not a singleton. Furthermore, for some particular choices of $\gamma > 0$ it may also be empty.

We present in the following the results concerned with the computation of the proximal map $\mathrm{prox}_{\gamma f}$ in (7.29), in the case when $f : \mathbb{R}^2 \to \mathbb{R}$ is the function

$$f(v) := \left(v^T \mathrm{A}\, v\right)^{p/2}, \quad v \in \mathbb{R}^2, \quad \mathrm{A} \in \mathbb{R}^{2\times 2} \text{ symmetric positive definite}, \quad p > 0\,. \tag{7.30}$$

The ADMM sub-step (7.28) will then be a special instance of (7.29) under the choice of f as above, $\gamma = \gamma_v$, $v = v_i$, $\mathrm{A} = \mathrm{R}_{\zeta_i}^T \Lambda_i^2 \mathrm{R}_{\zeta_i}$, $p = p_i$, and $q = q_i^{(t-1)}$, for $i = 1, \ldots, n$ and $t > 0$.

We now ensure that under the choice (7.30) above the minimization problem (7.29) admits solutions. Then, assuming that A has condition number $\kappa > 1$ we show how the calculation of the proximal map can be reduced to the solution of a one-dimensional problem, whose form depends on the input q and the matrix A. Note that the case $\kappa = 1$ boils down to consider a scalar and diagonal matrix A, which simplifies the problem and for which the results discussed in [111] can be used.

Proposition 7. *Under the choice (7.30), the optimization problem (7.29) admits at least one solution.*

Proof. Under the choice (7.30), both the terms in the objective function in (7.29) are continuous, bounded from below by zero and coercive over the entire domain $\mathbb{R}^2$. It clearly follows that the total objective function is continuous, bounded from below by zero and coercive, hence it admits at least one global minimizer. □

In the following, for $v, w \in \mathbb{R}^n$ we denote by $v \circ w$, $|v|$ and $\text{sign}(v)$ the component-wise (or Hadamard) product between v and w and the component-wise absolute value and sign of v, respectively.

Proposition 8. *Let $p,\ \gamma > 0$, $q \in \mathbb{R}^2$ and let $\mathrm{A} \in \mathbb{R}^{2\times 2}$ be a symmetric positive definite matrix with condition number $\kappa > 1$ and eigenvalue decomposition*

$$\mathrm{A} = \mathrm{V}^T \Lambda \mathrm{V}, \quad \mathrm{V}^T\mathrm{V} = \mathrm{V}\mathrm{V}^T = \mathrm{I}_2, \quad \Lambda = \text{diag}(\lambda_1, \lambda_2), \;\; \lambda_1 > \lambda_2 > 0. \tag{7.31}$$

Let us further define

$$\tilde{q} := \mathrm{V}q, \quad s := \text{sign}(\tilde{q}), \quad \bar{q} := |\tilde{q}|, \quad \bar{\gamma} := \frac{\gamma}{\lambda_2^{p/2}}, \quad \overline{\Lambda} := \text{diag}(\kappa, 1), \quad \kappa = \frac{\lambda_1}{\lambda_2}\,. \tag{7.32}$$

Then, any solution $v^* \in \mathbb{R}^2$ of the problem

$$v^* \in \arg\min_{v \in \mathbb{R}^2} \left\{ F(t) := (v^T \mathrm{A} v)^{p/2} + \frac{\gamma}{2}\, \|v - q\|_2^2 \right\}. \tag{7.33}$$

can be expressed as

$$v^* \,=\, V^T \left(s \circ z^*\right), \quad z^* \in \arg\min_{z \in \mathcal{H}_1} \, H(z)\,, \tag{7.34}$$

where the objective function $H : \mathbb{R}^2 \to \mathbb{R}$ and the feasible set $\mathcal{H}_1 \subset \mathbb{R}^2$ are defined by

$$H(z) := \left(z^T \,\overline{\Lambda}\, z\right)^{p/2} + \frac{\bar{\gamma}}{2} \,\|\, z - \bar{q} \,\|_2^2\,, \qquad \mathcal{H}_1 := \mathcal{H} \cap \left([0, \bar{q}_1] \times [0, \bar{q}_2]\right), \tag{7.35}$$

with $\mathcal{H}$ being the rectangular hyperbola defined by

$$\mathcal{H} := \left\{ (z_1, z_2) \in \mathbb{R}^2 : (z_1 - c_1)\,(z_2 - c_2) = c_1 c_2, \quad c_1 = -\frac{\bar{q}_1}{\kappa - 1}, \; c_2 = \frac{\kappa\, \bar{q}_2}{\kappa - 1} \right\}. \tag{7.36}$$

Proof. We start noticing that the matrix Λ in (7.31) can be factorized as $\Lambda = \lambda_2 \overline{\Lambda}$, where $\overline{\Lambda}$ is defined in (7.32). By substituting such factorization into (7.31), we can reformulate problem (7.33) as:

$$v^* \in \arg\min_{v \in \mathbb{R}^2} \left\{ \lambda_2^{p/2} \left(v^T \mathrm{V}^T \,\overline{\Lambda}\, V v\right)^{p/2} + \frac{\gamma}{2}\, \|v - q\|_2^2 \right\}. \tag{7.37}$$

After introducing the bijective linear change of variable

$$y := \mathrm{V}v \iff v = \mathrm{V}^T y,$$

we have that problem (7.37) can be equivalently expressed as

$$\begin{aligned} v^* &= \mathrm{V}^T y^*\,, \\ y^* &\in \arg\min_{y \in \mathbb{R}^2} \left\{ G(y) := \left(y^T \,\overline{\Lambda}\, y\right)^{p/2} + \frac{\bar{\gamma}}{2}\, \|y - \tilde{q}\|_2^2 \right\}, \end{aligned} \tag{7.38}$$

where $\overline{\gamma}$ and $\tilde{q}$ are defined in (7.32).

If $\tilde{q}_1 = \tilde{q}_2 = 0$ then one can trivially show that clearly $y^* = (0,0) \implies t^* = (0,0)$. We can then assume that $\tilde{q} \in \mathbb{R}^2 \setminus \{0\}$ and exploit symmetries of the function G in (7.38) to restrict the optimization problem to the case where $\tilde{q}$ lies in the first quadrant only. First, we notice that, for any given $a \in \mathbb{R}$ and $b \in \mathbb{R} \setminus \{0\}$, we have

$$
\begin{aligned}
a^2 =&\ (\mathrm{sign}(b))^2\, a^2 = (\mathrm{sign}(b)\, a)^2 \,, && (7.39)\\
(a-b)^2 =&\ (a - \mathrm{sign}(b)|b|)^2 = \left(\mathrm{sign}(b)\left(\frac{a}{\mathrm{sign}(b)} - |b|\right)\right)^2 \\
=&\ (\mathrm{sign}(b))^2\,(\mathrm{sign}(b)a - |b|)^2 = (\mathrm{sign}(b)a - |b|)^2 \,. && (7.40)
\end{aligned}
$$

By now recalling definitions of function G in (7.38) and of matrix $\overline{\Lambda}$ in (7.32), and then using (7.39)-(7.40), we can write

$$
\begin{aligned}
G(y) =&\ \left(\kappa y_1^2 + y_2^2\right)^{p/2} + \frac{\overline{\gamma}}{2}\left((y_1 - \tilde{q}_1)^2 + (y_2 - \tilde{q}_2)^2\right) \\
=&\ \left(\kappa\,(\mathrm{sign}(\tilde{q}_1)y_1)^2 + (\mathrm{sign}(\tilde{q}_2)y_2)^2\right)^{p/2} + \frac{\overline{\gamma}}{2}\left((\mathrm{sign}(\tilde{q}_1)y_1 - |\tilde{q}_1|)^2 + \left(\mathrm{sign}(\tilde{q}_2)y_2 - |\tilde{q}_2|^2\right)\right).
\end{aligned}
$$

By setting $\mathrm{S} := \mathrm{diag}\left(\mathrm{sign}\left(\tilde{q}_1\right), \mathrm{sign}\left(\tilde{q}_2\right)\right)$ we can now set

$$
z := \mathrm{S}y \quad \Longleftrightarrow \quad y = \mathrm{S}^{-1}z,
$$

which is a linear bijective change of variable since $\tilde{q}_1, \tilde{q}_2 \in \mathbb{R} \setminus \{0\} \implies \mathrm{sign}\left(\tilde{q}_1\right), \mathrm{sign}\left(\tilde{q}_2\right) \in \{-1, 1\}$. Recalling the definition of s and $\overline{q}$ in (7.32), we thus get that the optimization problem (7.38) is equivalent to

$$
\begin{aligned}
y^* =&\ s \circ z^* \,, \\
z^* \in&\ \arg\min_{z \in \mathbb{R}^2} \left\{ H(z) := \left(z^T\, \overline{\Lambda}\, z\right)^{p/2} + \frac{\overline{\gamma}}{2}\, \|\, z - \overline{q}\,\|_2^2 \right\} \,, && (7.41)
\end{aligned}
$$

where the vector $\overline{q} = (|\tilde{q}_1|, |\tilde{q}_1|)$ now lies in the first (open) quadrant $(0, +\infty)^2$.

We now prove that the solutions z^* in (7.41) belong to the arc of hyperbola $\mathcal{H}_1$ defined in (7.35). To this aim, we consider the following one-parameter family of ellipses depending on a parameter $R > 0$:

$$
\begin{aligned}
\mathcal{E}_R :=&\ \left\{(z_1, z_2) \in \mathbb{R}^2 \colon\ z^T\overline{\Lambda}z = R^2\right\} = \left\{(z_1, z_2) \in \mathbb{R}^2 \colon\ \kappa\, z_1^2 + z_2^2 = R^2\right\} \\
=&\ \left\{(z_1, z_2) \in \mathbb{R}^2 \colon\ z_1 = z_1(\zeta; R) = \frac{R}{\sqrt{\kappa}}\cos\zeta,\ z_2 = z_2(\zeta; R) = R\sin\zeta,\ \zeta \in [0, 2\pi[\right\} && (7.42)
\end{aligned}
$$

and, as a start, we show that the minimizers of the restriction of the function H in (7.41) to any ellipse $\mathcal{E}_R$ in (7.42) lie on the hyperbola $\mathcal{H}$ in (7.36). In Figure 7.2 we show the hyperbola $\mathcal{H}$ (magenta solid line) with its two orthogonal asymptotes, the arc $\mathcal{H}_1$ defined in (7.35) (red solid thick line) and one ellipse $\mathcal{E}_R$ (blue dashed line) as in (7.42).

Let us observe first that when restricted to an ellipse $\mathcal{E}_R$ of the form in (7.42), the objective function H depends only on ζ (R can be regarded as a fixed parameter). The restriction $H_R : \mathbb{R} \to \mathbb{R}$ takes then the following form

$$
H_R(\zeta; R) = R^p + \frac{\overline{\gamma}}{2}\left(\left(\frac{R}{\sqrt{\kappa}}\cos\zeta - \overline{q}_1\right)^2 + \left(R\sin\zeta - \overline{q}_2\right)^2\right).
$$

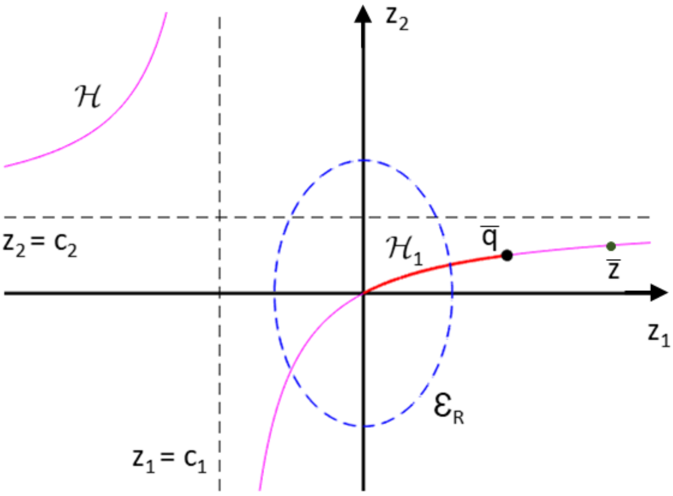


Figure 7.2: Graphical representation for the bivariate minimization problem (7.41).

For any $R > 0$, the function H_R above is clearly periodic with period 2π, bounded (from below and above) and infinitely many times differentiable in ζ, hence the minimizers of H_R can be sought for among its stationary points in the interval $[0, 2\pi)$. The first-order derivative of H_R is as follows:

$$\begin{aligned} H_R'(\zeta; R) &= \bar{\gamma} \left(-\frac{R}{\sqrt{\kappa}} \sin\zeta \left(\frac{R}{\sqrt{\kappa}} \cos\zeta - \bar{q}_1 \right) + R\cos\zeta \left(R\sin\zeta - \bar{q}_2 \right) \right) \\ &= \bar{\gamma}\, \frac{\kappa - 1}{\sqrt{\kappa}} \;\; \big(\left(z_1(\zeta;R) - c_1\right) \left(z_2(\zeta;R) - c_2\right) - c_1 c_2 \,\big) \end{aligned} \tag{7.43}$$

where (7.43) follows after some simple algebraic manipulations from the parametrization in (7.42), with c_1, c_2 constants defined in (7.36). Since $\gamma > 0$, $\kappa > 1$ by assumption, the scalar quantity $\bar{\gamma}\,(\kappa - 1)/\sqrt{\kappa}$ in (7.43) is positive, hence we have

$$H_R'(\zeta; R) = 0\,(>0, <0) \iff \left(z_1(\zeta;R) - c_1\right)\left(z_2(\zeta;R) - c_2\right) - c_1 c_2 = 0\,(>0, <0)\,. \tag{7.44}$$

It thus follows that, for any fixed $R > 0$ (that is, for any ellipse $\mathcal{E}_R$ in (7.42)), any stationary point $z(\zeta_R^* : R)$ of H_R satisfies

$$\left(z_1\left(\zeta_R^*; R\right),\, z_2\left(\zeta_R^*; R\right) \right) \in \mathcal{E}_R \cap \mathcal{H}\,,$$

i.e. it belongs to the set of intersection points between the ellipse $\mathcal{E}_R$ and the hyperbola $\mathcal{H}$ (see the two intersection points in Figure 7.2). It also follows from (7.44) that the intersection point in the first quadrant is the global minimizer for H_R, whereas the one in the third quadrant is the global maximizer. Since previous considerations hold true for any ellipse $\mathcal{E}_R$, then any global minimizer z^* of the unrestricted objective function H in (7.41) must belong to the restriction of the hyperbola $\mathcal{H}$ in (7.36) to the first quadrant.

Finally, it is easy to further shrink the locus of potential global minimizers z^* to the arc $\mathcal{H}_1$ defined in (7.35). Let us argue by contradiction and suppose there exists a global minimizer $\bar{z}$ belonging to the restriction of the hyperbola $\mathcal{H}$ to the first quadrant but not to $\mathcal{H}_1$ - see Figure

7.2. We have:

$$H(\bar{z}) - H(\bar{q}) = \underbrace{\left(\bar{z}^T \,\overline{\Lambda}\, \bar{z}\right)^{p/2} - \left(\bar{q}^T \,\overline{\Lambda}\, \bar{q}\right)^{p/2}}_{>0} + \underbrace{\frac{\overline{\gamma}}{2}\left(\| \,\bar{z} - \bar{q}\,\|_2^2 - \| \,\bar{q} - \bar{q}\,\|_2^2\right)}_{>0} > 0,$$

whence $\bar{z}$ can not be a global minimizer for the function H. □

In the following corollary we exploit and complete the results stated in previous Proposition 8 by showing how the bivariate minimization problem in (7.34) can be reduced to an equivalent univariate problem.

Corollary 3. *The minimizers $z^* \in \mathbb{R}^2$ in (7.34) can be obtained as follows:*

$$z^* = \left(z_1^*, c_2 \left(\frac{z_1^*}{z_1^* - c_1} \right) \right) ,$$

with c_1, $c_2 \in \mathbb{R}$ defined in (7.36) and $z_1^* \in \mathbb{R}$ the solution(s) of the following 1-dimensional constrained minimization problem:

$$z_1^* \;\in\; \arg \min_{\xi \,\in\, [0,\bar{q}_1]} \left\{ h(\xi) := \; (h_1(\xi))^{p/2} \,+\, \frac{\overline{\gamma}}{2}\, h_1(\xi) \,-\, \frac{\overline{\gamma}}{2}\, h_2(\xi) \right\} ,$$

$$h_1(\xi) = \xi^2 \left(\kappa + \frac{c_2^2}{(\xi - c_1)^2} \right) , \quad h_2(\xi) = \xi \, (\kappa - 1) \left(\xi - 2c_1 + 2 \, \frac{c_2^2}{\kappa(\xi - c_1)} \right) .$$

Proof. The proof is immediate by deriving the expression of z_2 as a function of z_1 from the definition of the hyperbola $\mathcal{H}$ in (7.36), then substituting this expression in the objective function H in (7.35) and, finally, carrying out some algebraic manipulations. □

To summarize, we report in Algorithm 6 the pseudocode of the proposed ADMM iterative scheme used to solve the saddle-point problem (7.24).

7.6 Parameters estimation results

In this section, an extensive evaluation on the accuracy of the ML estimation procedure described in Section 7.3 is carried out.

In order to assess the quality of the estimation, we introduce the following statistical notions.

Definition 4. *Let $\omega > 0$ be an unknown parameter of a fixed probability distribution p_ω and for $\ell > 0$ let ω_j, $j = 1, \ldots, \ell$ be estimates of ω obtained by a given estimation procedure. The* sample estimator *$\hat{\omega}$ of ω is defined as the average:*

$$\hat{\omega} := \frac{\sum_{j=1}^{\ell} \omega_j}{\ell}.$$

We can then define the relative bias *$\mathcal{B}_{\hat{\omega}}$, the* empirical variance *$\mathcal{V}_{\hat{\omega}}$ and the* relative root mean square error *$rmse_{\hat{\omega}}$ of the estimator $\hat{\omega}$ as:*

$$\mathcal{B}_{\hat{\omega}} := \frac{\mathbb{E}(\hat{\omega} - \omega)}{\omega}, \quad \mathcal{V}_{\hat{\omega}} := \frac{1}{\ell - 1} \sum_{j=1}^{\ell} (\, \omega_j - \hat{\omega} \,)^2, \quad rmse_{\hat{\omega}} := \frac{\sqrt{\mathcal{V}_{\hat{\omega}} \,+\, \mathcal{B}_{\hat{\omega}}^2}}{\omega}.$$

Algorithm 6: ADMM-based algorithm for the $\mathrm{DTV}_p^{\mathrm{sv}}$-$\mathrm{L}_2$ model

```
input:  $b \in \mathbb{R}^n$, $r > 0$, $\tau \approx 1$, $\gamma_v > 0$,$\gamma_w > 0$
output: restored image $u^*$
initialize:  set $u^{(0)} = b$, $\rho_w^{(0)} = \mathbf{0}_n$, $\rho_v^{(0)} = \mathbf{0}_{2n}$
for  $t$ = 1, 2, ... until convergence  do:
    update parameters
        · $p_i^{(t)}$, $\Sigma_i^{(t)}$ by solving (7.16) in terms of $\left(p_i^{(t)}, \phi_i^{(t)}, \varrho_i^{(t)}\right)$ for every $i = 1, \ldots, n$
        · $m_i^{(t)}$ by (7.11) for every $i = 1, \ldots, n$
        · $\mu^{(t)}$ by (5.35)
    update primal variables
        · $v^{(t)}$ as in Section 7.5.1
        · $w^{(t)}$ by (5.32)
        · $u^{(t)}$ by solving (5.33)
    update dual variables
        · $\rho_w^{(t)}$ by (5.25)
        · $\rho_v^{(t)}$ by (5.26)
end for
return:  $u^* = u^{(t)}$
```

In the following, the accuracy and the precision of the estimator is evaluated by analyzing its performance on the estimation of the parameters $(p, e^{(1)}, \zeta)$. As discussed in Section 7.3.2, parameters $e^{(1)}, \zeta$ can be derived from ϱ, ϕ and from (7.19), we recall that $e^{(1)} = \left(\frac{1}{\lambda^{(1)}}\right)^2$. In addition, we also consider how the quality of the estimation of $(p, e^{(1)}, \zeta)$ affects the estimation of the scale parameter m, which is computed directly via the formula (7.11) as a non-linear function of (p, ϕ, ϱ) or, equivalently, of $(p, e^{(1)}, \zeta)$, as well as of the samples. The non-linearity may affect the accuracy of its estimation.

7.6.1 Parameter estimation: accuracy and precision

We now perform some tests assessing the accuracy and the precision of the ML estimation procedure proposed in Section 7.3 in terms of the quantities defined above. As a first test we compare the results obtained by applying the ML procedure to estimate a BGGD of parameters $(\bar{p}, \bar{e}^{(1)}, \bar{\zeta}, \bar{m}) = (1, 1.4, 45°, 0.3)$. We run our tests for an increasing number $N \in \{10, 10^2, 10^3, 10^4, 10^5, 10^6\}$ of samples drawn from the distribution. For each value of N, the estimation procedure is run $\ell = 200$ times. For any $j = 1, \dots, \ell$ we estimate the parameter triple $(p^*, \phi^*, \varrho^*)_j$ and consider the corresponding estimators of the true parameters as defined in Definition 4. The results are shown in Figures 7.3 - 7.5.

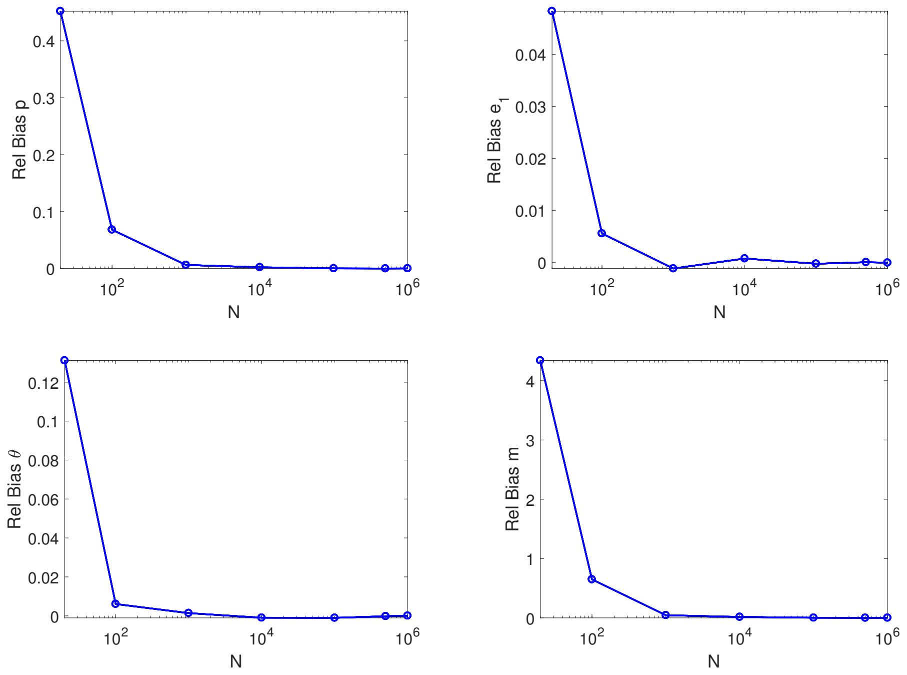


Figure 7.3: Plots of relative bias for estimated $(p^*, {e^{(1)}}^*, \zeta^*, m^*)$ in semi-logarithmic scale on x-axis.

For all parameters (including the scale parameter m), the behavior of relative bias, variance and relative root mean square error as the number of samples increases reveals good precision and accuracy. In particular, low values of such error quantities are already obtained when $N \approx 10^2$.

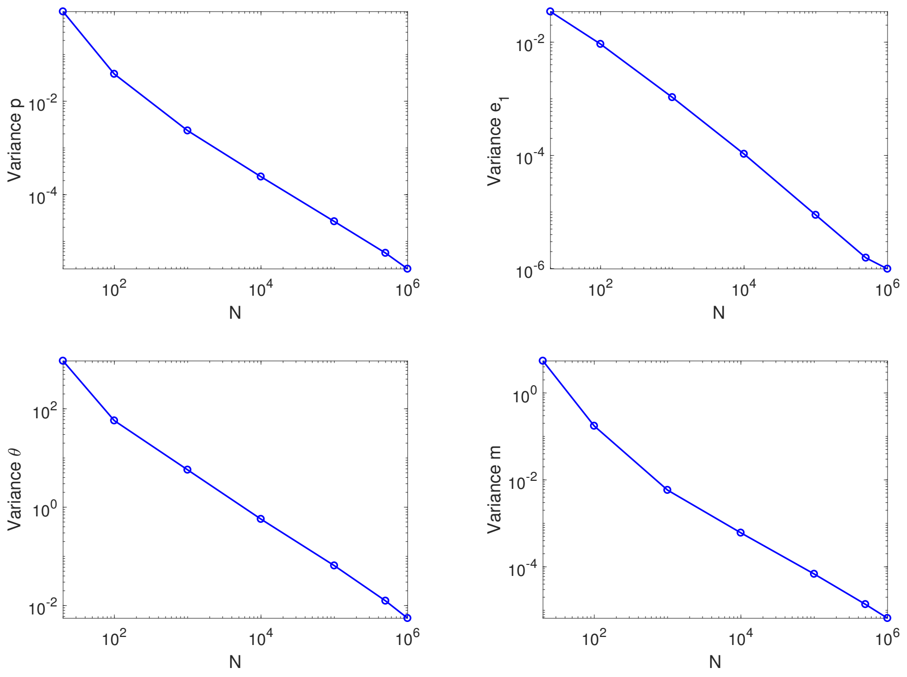


Figure 7.4: Plot of the empirical variance for estimated $(p^*, {e^{(1)}}^*, \zeta^*, m^*)$ in semi-logarithmic scale on x-axis.

7.6.2 Parameter estimation on synthetic neighbourhoods

We now test the ML estimation procedure on a simple synthetic image reported in Figure 7.6a. Here, the goal is to evaluate the effectiveness of the estimation when discriminating between different image regions such as edges, corners and circular profiles in terms of the functional shape of the estimated BGGD. In the following test, we estimate the parameters of the unknown BGGD in three different situations where a pixel surrounded by a 11×11 neighborhood is chosen to lie on a vertical edge (Fig. 7.6), a corner (Fig. 7.7) and on a circular profile (Fig. 7.8). In order to avoid degenerate configurations of the gradients, such as the ones described in (7.17), we preliminary corrupt the image by a small Additive White Gaussian noise (AWGN) with $\sigma = 0.03$.

Edge points In Fig. 7.6b, we report the scatter plot of the gradients of the edge points in the red-bordered region countered in Figure 7.6a, which, as expected, shows its distribution along the x-axis. The parameter estimation procedure of the BGGD at one of such edge points is run by taking 121 samples of gradients in the 11×11 neighborhood. The estimation procedure results in the following parameters $(p^*,\ {e^{(1)}}^*,\ \zeta^*,\ m^*) = (0.07,\ 1.60, -177.82°,\ 2 * 10^{-5})$. Note that the low value of the parameter p leads to a very fat tail distribution, as shown in Fig. 7.6c. The orientation and the eccentricity of the level curves are in line with the clear directionality of the samples as it can be seen in Figure 7.6d.

Corner points For the corner example in Figure 7.7, the scatter plot of the gradients is reported in Figure 7.7b. The ML procedure results in this case in the estimation $(p^*,\ {e^{(1)}}^*,\ \zeta^*,\ m^*) = (0.07,\ 1.08,\ 72.49°,\ 3 * 10^{-7})$. The estimated PDF is reported in Fig. 7.7c. Similarly as

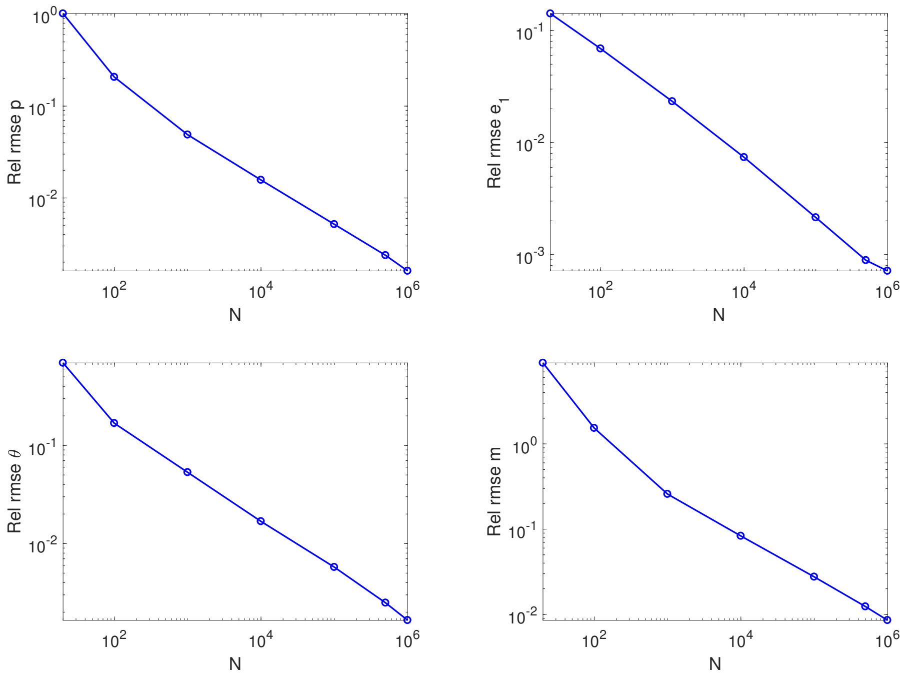
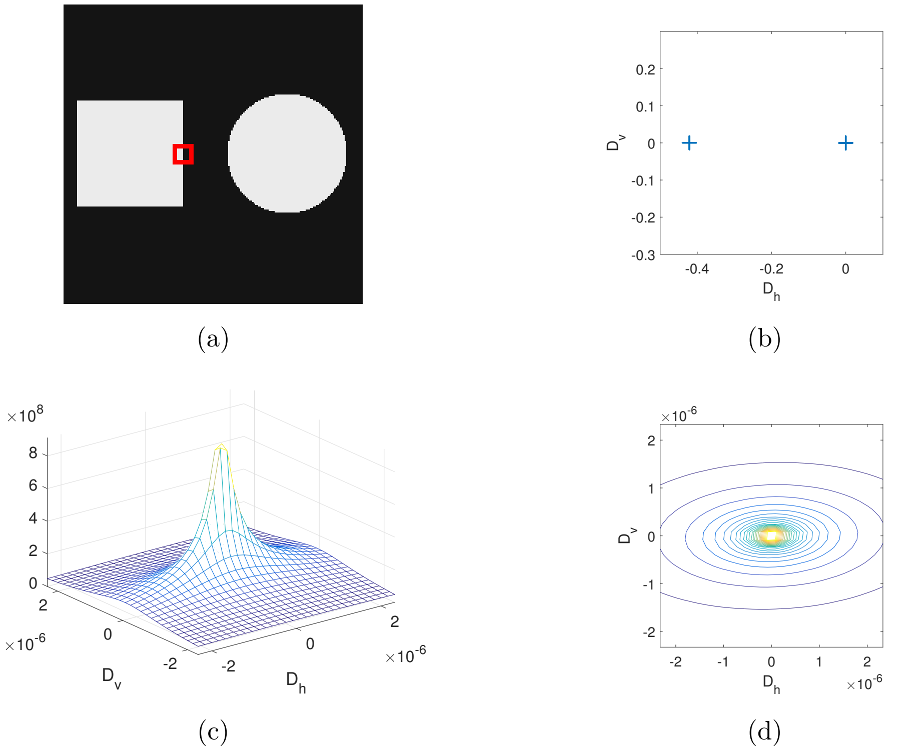


Figure 7.5: Plot of relative root mean square error for estimated $(p^*, e^{(1)^*}, \zeta^*, m^*)$ in semi-logarithmic scale on x-axis.



(a)

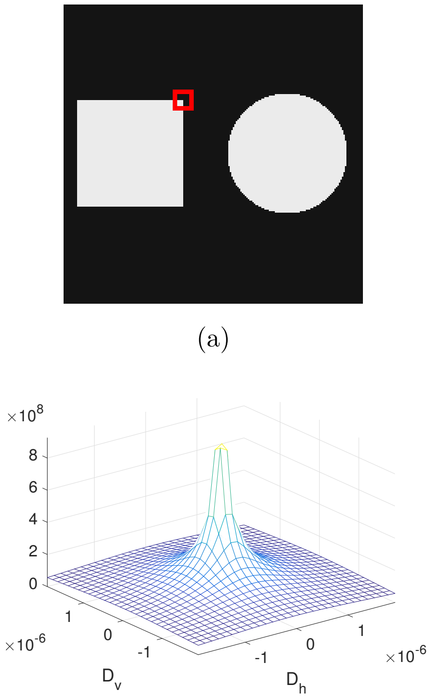
(b)

(c)

(d)

Figure 7.6: 7.6a: BGGD Parameter estimation for a synthetic geometrical image. Test for edge image pixel. 7.6b: Scatter plot of the gradients in the read-bordered region. 7.6c: PDF with estimated parameters $(p^*,\ e^{(1)^*},\ \zeta^*,\ m^*) = (0.07,\ 1.60,\ -177.82°,\ 2*10^{-5})$. 7.6d: Level curves of the estimated PDF.

before, note that a very fat-tail distribution is estimated. On the other hand, since ${e^{(1)}}^* \approx 1$, we also have ${e^{(2)}}^* \approx 1$ and the eccentricity of the ellipse $\epsilon \approx 0$. We can conclude that, in this case, the distribution is almost isotropic and the angle ζ has a negligible influence on the orientation of the level curves as it can be seen in Figure 7.6d.



(a)

(b)

(c)

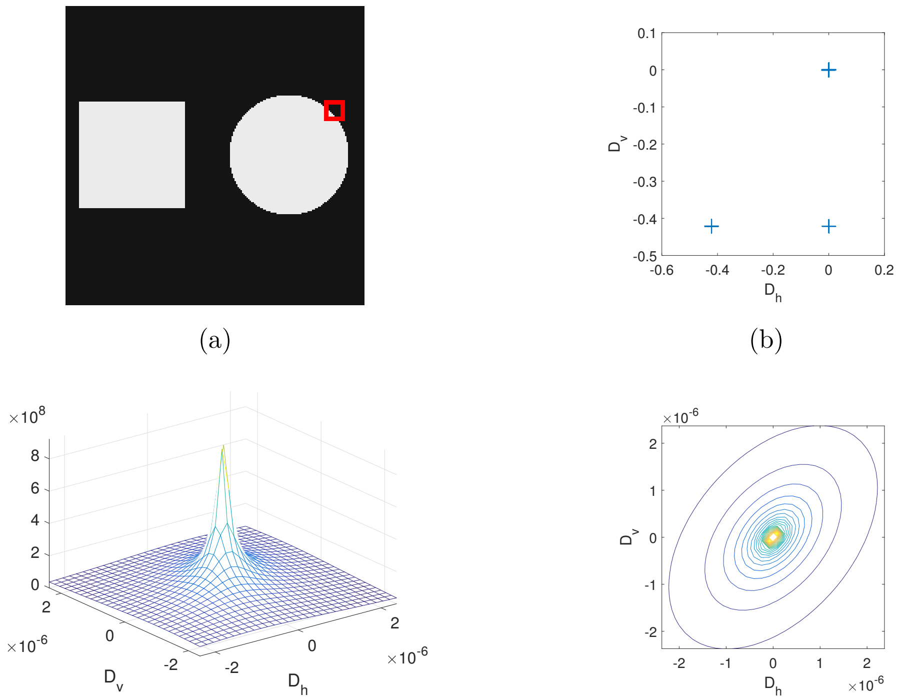
(d)

Figure 7.7: 7.7a: BGGD Parameter estimation for a synthetic geometrical image. Test for corner image pixel. 7.7b: Scatter plot of the gradients in the read-boarded region. 7.7c: PDF with estimated parameters $(p^*,\ {e^{(1)}}^*,\ \zeta^*,\ m^*) = (0.07,\ 1.08,\ 72.49°,\ 3 * 10^{-7})$. 7.7d: Level curves of the estimated PDF.

Circle points Finally, we consider the ML parameter estimation procedure in correspondence with a pixel lying on a circular profile, see Figure 7.8. In this case, the estimated parameters are $(p^*, {e^{(1)}}^*, \zeta^*, m^*) = (0.08, 1.44, 49.28°, 2 * 10^{-6})$. The values obtained for ${e^{(1)}}^*$ and ζ^* reflect the spatial distribution of the gradients in Figure 7.8b.

7.6.3 Parameter estimation on synthetic images

Motivated by the good results above, we report in this section the numerical experiments concerned with the estimation of the four parameters $(p^*, {e^{(1)}}^*, \zeta^*, m^*)$ at any image pixel. For the following estimations, we fix a neighborhood of 3×3 pixels. It is worth remarking here that the tests in section 7.6.1 have been computed on samples directly drawn from a BGGD. For such example, we remarked on how a large number of samples reflects on a reliable estimation of the BGGD parameters. When dealing with real images, however, our goal rather consists in estimating the parameters of the BGGD of the local gradient from the surrounding ones, since, clearly, one single sample is not sufficient to get a reliable estimate. However, the samples



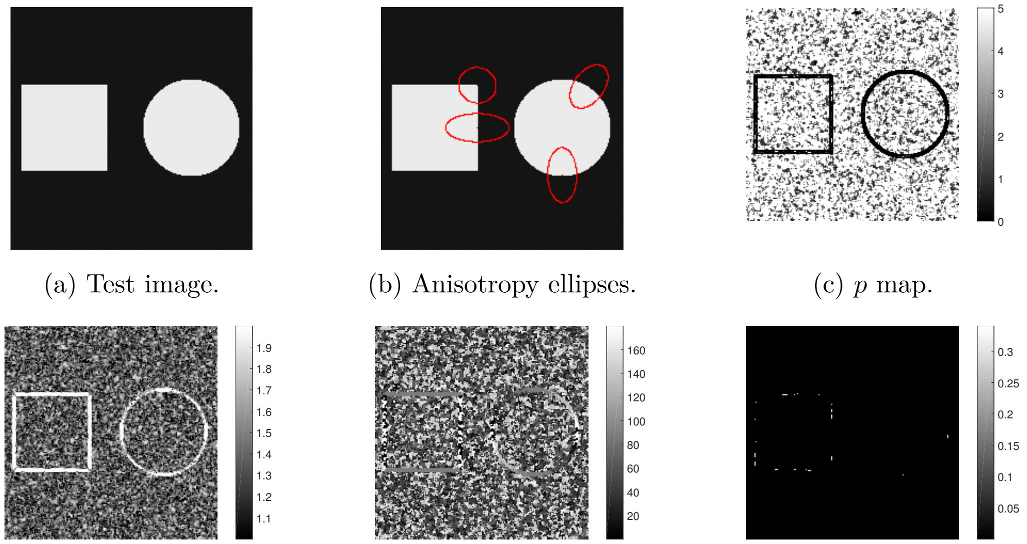
(a) (b)

(c) (d)

Figure 7.8: 7.8a: BGGD Parameter estimation for a synthetic geometrical image. Test on image pixel lying on circular profile. 7.8b: Scatter plot of the gradients in the read-boarded region. 7.8c: PDF with estimated parameters $(p^*, {e^{(1)}}^*, \zeta^*, m^*) = (0.08, 1.44, 49.28°, 2 * 10^{-6})$. 7.8d: Level curves of the estimated PDF.

involved in the estimation procedure are in general not drawn from the same BGGD as their parameters may be different. Thus, their number has to be limited in order to reduce modeling errors as much as possible. In conclusion, the size of the neighborhood is a trade off between the local properties of the image and the robustness of the estimate procedure, the former requiring small neighborhoods, the latter requiring larger ones. In order to avoid degenerate configurations, we corrupt the images by AWGN with $\sigma = 0.03$. Moreover, the search interval for the shape parameter p is set equal to $[0.1, 5]$. We start considering the synthetic test image used already in the experiment above, i.e. Figure 7.9a. Here we perform the estimation of the parameters at any pixel and report the local parameter maps in Figure 7.9c, 7.9d, 7.9e and 7.9f. Furthermore, we report in Figure 7.9b the anisotropy ellipses representing the level curves of the estimated PDF, drawn as described in Section 7.3.2, whose orientation, given by the ζ-map in 7.9e, is in line with what we expected and with the test proposed in the previous sub section (see Fig. 7.6 - 7.7). One can also observe that the higher values in the $e^{(1)}$-map are estimated to be along the edges, describing the strong anisotropy of the level curves there, while the higher values in the p-map are in the piece-wise constant regions. This can be explained by saying that in these regions the estimation procedure detects a plain Bivariate Gaussian Distribution characterized by a shape parameter $p = 2$. This is of course due to the presence of AWGN.

The same experiments are proposed for `geometric` test image in Figure 7.10a. Even though such image presents edges displaced along different orientations and details on different scales, the results showed in Figure 7.10b-7.10f confirm the robustness of estimator in distinguishing between different image regions.



(a) Test image.

(b) Anisotropy ellipses.

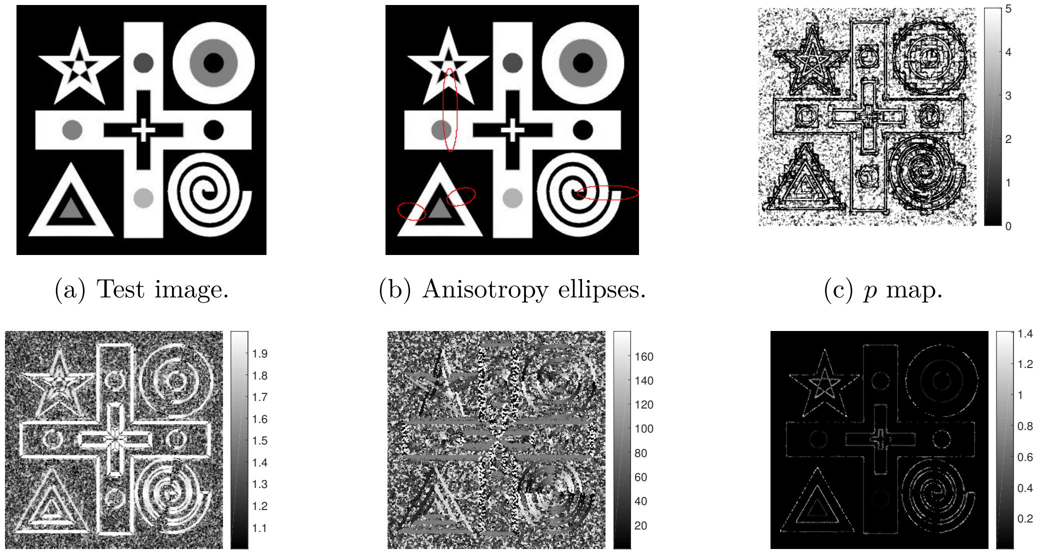
(c) p map.

(d) $e^{(1)}$ map.

(e) ζ map.

(f) m map.

Figure 7.9: Test on `synthetic` image.



(a) Test image.

(b) Anisotropy ellipses.

(c) p map.

(d) $e^{(1)}$ map.

(e) ζ map.

(f) m map.

Figure 7.10: Test on `geometric` image.

Remark 3. *In order to generate the samples used in the parameter map estimation above, one has to choose a suitable discretization of the image gradient. Here, we considered central differences schemes. Compared to standard forward/backward difference schemes, this choice avoids the undesired correlation between the horizontal and the vertical components. As preliminary numerical tests showed, such correlation may result indeed into a deviation between the estimated ζ^* from the one estimated above.*

σ	TV-L$_2$	TV$_p$-L$_2$	HWTV-L$_2$	TV$^{\rm sv}_{p,\alpha}$-L$_2$	DTV$^{\rm sv}_p$-L$_2$
0.02	2.46	3.14	3.20	3.23	**3.61**
0.03	1.74	1.99	2.12	2.14	**2.79**
0.06	1.59	2.02	2.11	2.13	**2.90**

Table 7.1: ISNR values for the `barbara` test image for decreasing $\sigma = 0.02, 0.03, 0.06$.

7.7 Computed examples

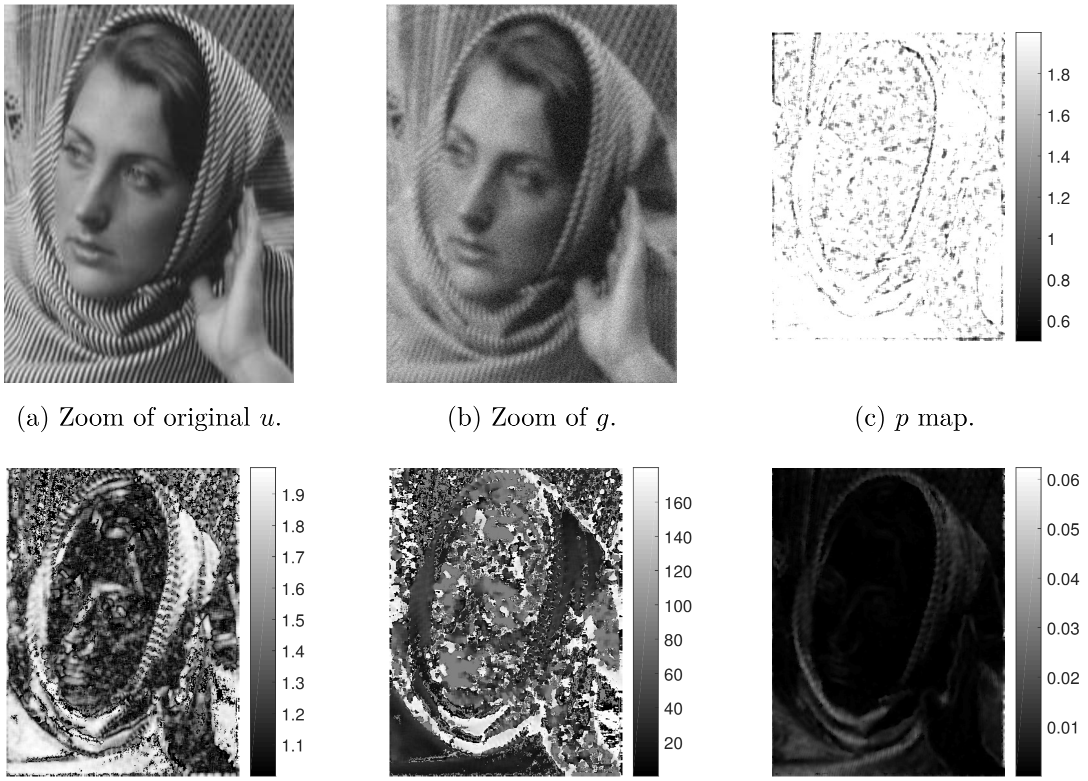
In this section, we evaluate the performance of the DTV$^{\rm sv}_p$-L$_2$ image reconstruction model (7.3) applied to the restoration of grey-scale images corrupted by (known) blur and AWGN.

The DTV$^{\rm sv}_p$-L$_2$ model will be compared with the following ones:

- the TV-L$_2$ model;
- the HWTV-L$_2$ model in Chapter 5;
- the TV$_p$-L$_2$ model, with constant $p \in (0, 2]$;
- the TV$^{\rm sv}_{p,\alpha}$-L$_2$ in Chapter 6.

As far as the update of the local parameters is concerned, we remark that a closed formula for the scale parameters m_i is available, while the local shape parameters p_i, the local orientations ζ_i and the local maximum eigenvalue $e_i^{(1)}$ are the solution of the minimization problem (7.18) written in terms of auxiliary variables. Due to the large number of unknowns, here the maps of the parameters will be computed only at the beginning of the iterations.
For the numerical solution of the DTV$^{\rm sv}_p$-L$_2$ model we use the ADMM-based algorithm 6 where for all tests we manually set the penalty parameters γ_v and γ_w.

Barbara image We start testing the reconstruction algorithm on a zoom of a high resolution (1024 × 1024) `barbara` test image with size 471 × 361, characterized by the joint presence of texture and cartoon regions. The image here has been corrupted by Gaussian blur of `band`= 9 and `width`= 2 and AWGN with different standard deviation $\sigma = 0.02, 0.03, 0.06$. The original image and the observed image, as well as the four parameter maps, computed considering a neighborhood of size 7 × 7, are shown in Figure 7.11. In order to avoid inaccurate estimations of the parameters due to the presence of possibly large noise, the parameter p^* in the TV$_p$-L$_2$ model as well as the local maps of the parameters in the DTV$^{\rm sv}_p$-L$_2$ have been computed after few iterations (usually 5) of the TV-L$_2$ model. Furthermore, as discussed in Section 7.3.2, the p parameter has been computed by restricting the admissible range to $[0.1, 2]$. In Tables 7.1 and 7.2 the ISNR and SSIM values achieved by the TV-L$_2$, the TV$_p$-L$_2$ (with estimated global $p = 0.92$), the HWTV-L$_2$, the TV$^{\rm sv}_{p,\alpha}$-L$_2$ and the DTV$^{\rm sv}_p$-L$_2$ models for different values of σ are reported. We note that the proposed model outperforms the competing ones. As shown in Figure 7.12, the flexibility of the DTV$^{\rm sv}_p$ regularizer strongly improves the reconstruction quality mainly in terms of better texture preservation.



(a) Zoom of original u.

(b) Zoom of g.

(c) p map.

(d) $e^{(1)}$ map.

(e) ζ map.

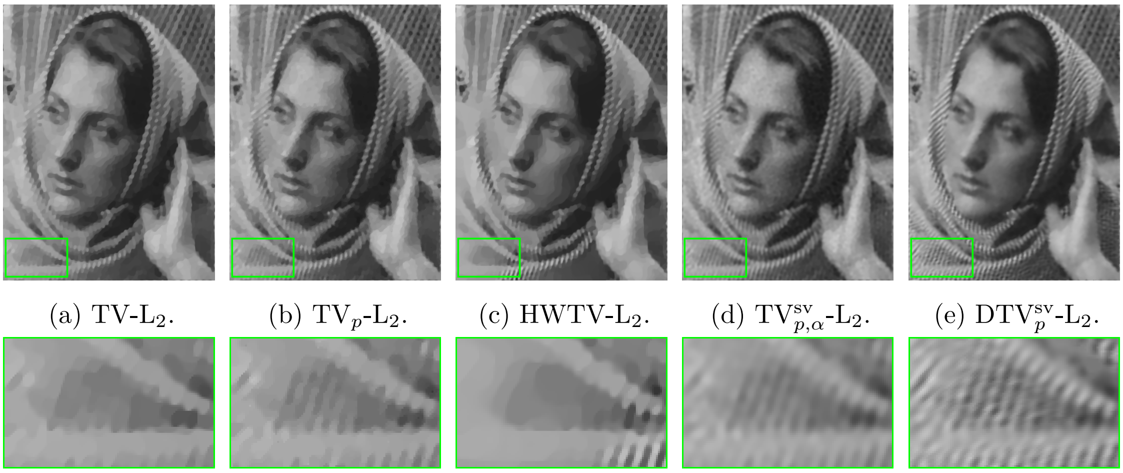
(f) m map.

Figure 7.11: Parameter maps for a zoom of the `barbara` test image. Image is corrupted by Gaussian blur and AWGN with $\sigma = 0.06$.

Natural image As a second test, we compared the performance of DTV_p^{SV}-L_2 restoration model on a 500×500 portion of a high resolution (1024×1024) natural test image characterized by fine-scale textures of different types. As in the previous example, we similarly corrupt the image by AWGN and Gaussian blur of `band`$= 9$ and `width`$= 2$ with $\sigma = 0.02, 0.03, 0.06$. The original and the observed images, as well as the four parameter maps computed considering neighborhoods of size 3×3 are shown in Figure 7.13. Similarly as for the numerical test above few preliminary iterations of TV-L_2 are performed before computing the parameter maps. The research interval for the p parameter has been set equal to $[0.1, 2]$. It is worth remarking that the very small neighborhood size used for the parameter estimation is the one yielding the best restoration results for this test. We believe that this is motivated by the very fine scale of details in the test image. In Tables 7.3 and 7.4, the ISNR and SSIM values achieved by the compared

σ	TV-L_2	TV_p-L_2	HWTV-L_2	$\text{TV}_{p,\alpha}^{\text{SV}}$-$\text{L}_2$	DTV_p^{SV}-L_2
0.02	0.80	0.83	0.83	0.83	**0.85**
0.03	0.74	0.75	0.78	0.77	**0.80**
0.06	0.65	0.68	0.70	0.69	**0.74**

Table 7.2: SSIM values for the `barbara` test image for increasing $\sigma = 0.02, 0.03, 0.06$.



(a) TV-L$_2$. (b) TV$_p$-L$_2$. (c) HWTV-L$_2$. (d) TV$^{\text{sv}}_{p,\alpha}$-L$_2$. (e) DTV$^{\text{sv}}_p$-L$_2$.

(f) Zoom of (a) (g) Zoom of (b) (h) Zoom of (c) (i) Zoom of (d) (j) Zoom of (e)

Figure 7.12: Detail of reconstruction of `barbara` image 7.11. Texture components are much better preserved by encoding directional information.

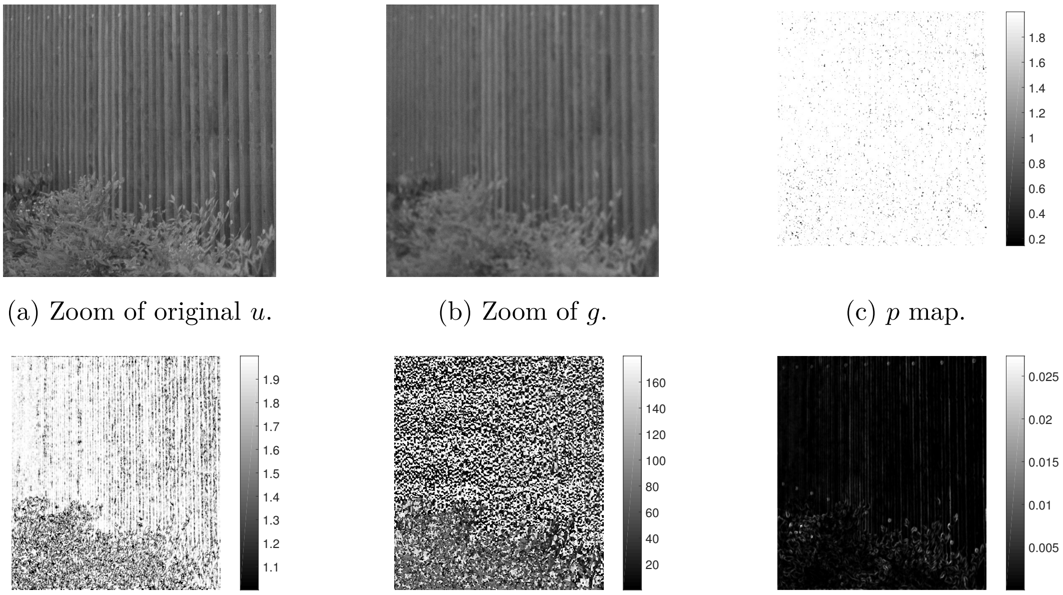
σ	TV-L$_2$	TV$_p$-L$_2$	HWTV-L$_2$	TV$^{\text{sv}}_p$-L$_2$	DTV$^{\text{sv}}_p$-L$_2$
0.02	2.07	2.43	2.40	2.53	**2.78**
0.03	1.83	2.06	2.01	2.26	**2.56**
0.06	0.94	1.55	1.67	1.86	**2.45**

Table 7.3: ISNR values for the test image in 7.13 for $\sigma = 0.02, 0.03, 0.06$.

models for different values of σ are reported. Also in this case, the proposed model outperforms the competing ones. Note that the improvement is actually more significant in correspondence of higher noise levels. In Figure 7.14, a visual comparison between the reconstructions obtained by the different models for $\sigma = 0.06$ is proposed.

σ	TV-L$_2$	TV$_p$-L$_2$	HWTV-L$_2$	TV$^{\text{sv}}_p$-L$_2$	DTV$^{\text{sv}}_p$-L$_2$
0.02	0.78	0.79	0.79	0.80	**0.81**
0.03	0.76	0.77	0.78	0.78	**0.79**
0.06	0.70	0.72	0.74	0.74	**0.76**

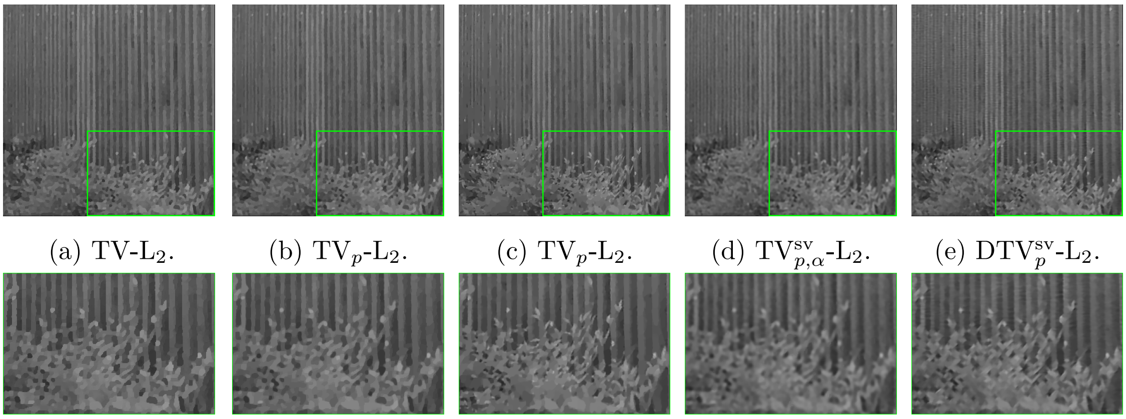
Table 7.4: SSIM values for the test image in 7.13 for $\sigma = 0.02, 0.03, 0.06$.



(a) Zoom of original u.
(b) Zoom of g.
(c) p map.

(d) $e^{(1)}$ map.
(e) ζ map.
(f) m map.

Figure 7.13: Parameter maps for a zoom of a natural test image. Image is corrupted by AWGN and blur for a $\sigma = 0.06$.



(a) TV-L$_2$.
(b) TV$_p$-L$_2$.
(c) TV$_p$-L$_2$.
(d) TV$^{\text{sv}}_{p,\alpha}$-L$_2$.
(e) DTV$^{\text{sv}}_{p}$-L$_2$.

(f) Zoom of 7.14a.
(g) Zoom of 7.14b.
(h) TV$_p$-L$_2$.
(i) Zoom of 7.14d.
(j) Zoom of 7.14e.

Figure 7.14: Detail of reconstruction of natural test image 7.13. Texture components are much better preserved by encoding directional information.

Part III

Informative hyperpriors

Chapter 8

Sparse reconstructions with generalized gamma hyperpriors

Part III of the thesis is characterized by the adoption of informative hyperpriors in the hierarchical model for solving the signal restoration linear inverse problem. Here, we limit our analysis to the case of signals admitting a sparse representation in a given basis. This assumption helps in getting an easier intuition on the expected behavior of the unknown parameters. The framework outlined in the following also applies to the case of sub-sampled data, where K is a wide matrix, and of additive Gaussian noise, not necessarily white.

Sparse recovery has been a very active field of research in the last decades, especially due to surge of interest in compressed sensing. Besides the classical references mentioned in Section 4.2, in [32] the authors formulate the sparse recovery problem as an inverse problem in a Bayesian framework, by modeling the components in the sparse signal as independent zero-mean Gaussian random variables whose variances follow a gamma distribution. In this chapter, we extend the analysis proposed in [32] to the case of generalized gamma hyperprior. In particular, we will focus on the convexity properties of the resulting energy functionals.

8.1 Hierarchical Bayesian Models

Consider the linear observation model with additive Gaussian noise,

$$b = \mathrm{K}x + e\,, \quad e \sim \mathcal{N}(0, \Sigma), \tag{8.1}$$

where $\mathrm{K} \in \mathbb{R}^{m \times n}$ is the blur matrix - typically, we consider the case $m \leq n$ - $\Sigma \in \mathbb{R}^{m \times m}$ is the noise symmetric positive definite covariance matrix that we assume to be known, and $x \in \mathbb{R}^n$ is the unknown that we are interested in recovering. We remark that x can be either a one-dimensional signal or a vectorized image, and this motivates the change of notation with respect to Part II. Notice that here the noise is not necessarily scaled white Gaussian, since Σ is not assumed to be a scaled identity; in any case, without loss of generality, we may assume that $\Sigma = \mathrm{I}$, since given a symmetric factorization $\Sigma^{-1} = \mathrm{S}^{\mathrm{T}}\mathrm{S}$, noise can be *whitened* through a linear transformation of $\mathrm{K} \to \mathrm{SK}$ and $b \to \mathrm{S}b$.

After whitening, the likelihood distribution is of the form

$$\mathrm{P}(b \mid x) \propto \exp\Big(-\frac{1}{2}\|\mathrm{K}x - b\|^2 \Big).$$

We assume that, possibly after a change of variables, the unknown is represented in a basis where the generative vector x is sparse; this also means that from now on a synthesis approach will be adopted. The a priori belief that x is sparse is encoded by modeling its components as independent random variables following a zero mean *conditionally Gaussian* distribution, i.e.,

$$x_j \sim \mathcal{N}(0, \theta_{\mathrm{pr}_j}), \quad \theta_{\mathrm{pr}_j} > 0, \quad 1 \leq j \leq n,$$

with unknown prior variances θ_{pr_j}. In the following, we are omitting the subscription for the j-th variance and the vector of variances will be just denoted by θ. According to the Bayesian paradigm, the unknown variances are also modeled as random variables, hence the expression for the conditional Gaussian prior must take into account the portion of the normalizing factor that depends on the variances,

$$\mathrm{P}(x \,|\, \theta) \propto \frac{1}{\prod_{j=1}^{n} \sqrt{\theta_j}} \exp\Big(-\frac{1}{2}\|\mathrm{D}_\theta^{-1/2} x\|^2 \Big), \quad \mathrm{D}_\theta = \mathrm{diag}(\theta_1, \ldots, \theta_n).$$

In this manner, the a priori believed sparsity of x can be formulated as a property of the variances of the components, with smaller variances promoting values closer to zero and encoded in the *hyperprior* $\mathrm{P}_{\mathrm{hyper}}(\theta)$. Recalling the framework outlined in Section 3.2.4, in the Bayesian setting, where all unknowns are modeled as random variables, solving (8.1) is tantamount to estimating x and θ, or more generally, to exploring their joint posterior distribution conditioned on b. The joint prior distribution $\mathrm{P}_{\mathrm{prior}}(x, \theta)$ is the product of the conditional prior $\mathrm{P}(x \mid \theta)$ and the hyperprior. It follows from Bayes' formula that the posterior distribution $\mathrm{P}(x, \theta \mid b)$ is

$$\mathrm{P}(x, \theta \mid b) \propto \mathrm{P}_{\mathrm{prior}}(x, \theta)\, \mathrm{P}(b \mid x) = \mathrm{P}(x \mid \theta)\, \mathrm{P}_{\mathrm{hyper}}(\theta)\, \mathrm{P}(b \mid x).$$

The a priori believed sparsity of the signal at hand, to gather with computational consideration, guides the choice of a prior for the hyperparameters θ_j. More specifically, we promote sparsity of the signal by selecting the hyperprior from the parametric family of generalized gamma distributions:

$$\mathrm{P}_{\mathrm{hyper}}(\theta) = \mathrm{P}_{\mathrm{hyper}}(\theta \mid r, \beta, \vartheta) = \frac{|r|^n}{\gamma(\beta)^n} \prod_{j=1}^{n} \frac{1}{\vartheta_j} \left(\frac{\theta_j}{\vartheta_j} \right)^{r\beta - 1} \exp\Bigg(-\left(\frac{\theta_j}{\vartheta_j} \right)^{r} \Bigg),$$

where $r \in \mathbb{R} \setminus \{0\}$, $\beta > 0$, $\vartheta_j > 0$; further restrictions on the parameters of the generalized gamma may be necessary to guarantee finite mean and variance. Observe that the hyperprior could be generalized further by letting each component θ_j have its own hyperparameter r and β. This generalization is not considered here.

In this context, the Maximum A Posteriori (MAP) estimate of x in (8.1) is used to represent the distribution. We are thus interested in solving the following problem:

$$(x^*, \theta^*) = \arg \min_{x \in \mathbb{R}^n, \theta \in \mathbb{R}^n_+} \{ -\log \mathrm{P}(x, \theta \mid b) \} = \arg \min_{x \in \mathbb{R}^n, \theta \in \mathbb{R}^n_+} \{ \mathcal{F}(x, \theta) \}, \tag{8.2}$$

where, with our choices of prior, hyperprior, and likelihood,

$$
\begin{aligned}
&\mathcal{F}(x,\theta) = \mathcal{F}(x,\theta \mid r,\beta,\vartheta)\\
=&\frac{1}{2}\|\mathrm{K}x-b\|^2 + \frac{1}{2}\|\mathrm{D}_\theta^{-1/2}x\|^2 - \left(r\beta - \frac{3}{2}\right)\sum_{j=1}^n \log\frac{\theta_j}{\vartheta_j} + \sum_{j=1}^n \left(\frac{\theta_j}{\vartheta_j}\right)^r \\
=&\frac{1}{2}\|\mathrm{K}x-b\|^2 + \mathcal{P}(x,\theta \mid r,\beta,\vartheta).
\end{aligned}
\tag{8.3}
$$

In the following, we will refer to $\mathcal{P}(x,\theta \mid r,\beta,\vartheta)$ as the penalty term in the MAP objective function. Our aim in this section is to analyze hierarchical Bayesian models with generalized gamma hyperpriors for different choices of the hyperparameters. In particular, we are interested in shedding some light on

1. how the sparsity of the MAP estimate depends on the hyperparameters;
2. how, for some choices of the hyperparameters, the MAP penalty term approaches classical penalty terms;
3. the dependency of the convexity - or lack thereof - of the MAP objective function on the hyperparameters;
4. the performance of the Iterative Alternating Sequential (IAS) minimization algorithm for the computation of the MAP estimate, reviewed in the next subsection, with various generalized gamma hyperpriors for the reconstruction of sparse signals from underdetermined noisy data.

8.1.1 IAS Algorithm

Our algorithm of choice for the solution of the minimization problem (8.2) is an alternating sequential scheme whose properties and performance for some choices of hyperparameters have been analyzed in [23, 32, 25]. Given an initial $\theta^{(0)}$ and setting $t \geq 1$, the IAS algorithm proceeds through a sequence of simple alternating updates of the form

$$x^{(t)} = \arg\min_{x\in\mathbb{R}^n}\{\mathcal{F}(x,\theta^{(t-1)})\}, \quad \theta^{(t)} = \arg\min_{\theta\in\mathbb{R}^n_+}\{\mathcal{F}(x^{(t)},\theta)\},$$

until a convergence criterion is met. As far as stopping criteria for IAS are concerned, two natural convergence criteria can be utilized: either the relative change of the objective function value is below a given threshold, or the relative change in the variable updates is below a threshold. In the computed examples, both criteria are used.

Among the appealing features of the IAS scheme applied in this framework, we mention the fact that both updating steps are particularly simple to implement, and that the algorithm has been shown to converge [23], with a convergence rate at least linear [32] for some classes of problems. We review the updating steps below.

Step 1: Updating x

Due to the structure of the objective function (8.3), the updating of x given θ reduces to solving a quadratic optimization problem

$$x^{(t)} = \arg\min_{x\in\mathbb{R}^n}\{\|\mathrm{K}x-b\|_2^2 + \|\mathrm{D}_\theta^{-1/2}x\|_2^2\}, \quad \theta = \theta^{(t-1)}\,,$$

or, equivalently, to finding the solution of the linear system

$$\begin{bmatrix} \mathrm{K} \\ \mathrm{D}_\theta^{-1/2} \end{bmatrix} x = \begin{bmatrix} b \\ 0 \end{bmatrix} \tag{8.4}$$

in the least squares sense. The latter is a well posed problem because $\theta \in \mathbb{R}^n_+$; if x is of large dimensions or K is not explicitly available, an iterative least squares solver may be the method of choice to solve (8.4). Due to the well-posedness of the problem, the iteration will continue until a sufficient reduction of the residual norm has been achieved. A computationally efficient way to determine an approximation of the MAP solution that is particularly appealing when the data vector is much lower dimensional than x, has been proposed in the cited articles on IAS and further analyzed in [26]. After the change of variable,

$$\mathrm{D}_\theta^{-1/2} x = w \sim \mathcal{N}(0, \mathrm{I}_n),$$

which is equivalent to whitening x via a Mahalanobis transformation, the linear system (8.4) becomes

$$\begin{bmatrix} \widetilde{\mathrm{K}}_\theta \\ \mathrm{I}_n \end{bmatrix} w = \begin{bmatrix} b \\ 0 \end{bmatrix}, \quad \widetilde{\mathrm{K}}_\theta = \mathrm{K}\mathrm{D}_\theta^{1/2}. \tag{8.5}$$

It has been pointed out repeatedly in the literature that the Tikhonov regularized solution is close to the approximate solution obtained by solving the linear system

$$\widetilde{\mathrm{K}}_\theta w = b \tag{8.6}$$

with an iterative linear solvers, equipped with a suitable early termination rule discussed below. When the iterative solver employed for the solution of (8.6) is the Conjugate Gradient for Least Squares (CGLS) method [90], the kth iterate satisfies

$$w_k = \mathrm{argmin}\{\|b - \widetilde{\mathrm{K}}_\theta w\| \mid w \in \mathscr{K}_k(\widetilde{\mathrm{K}}_\theta^{\mathrm{T}} b, \widetilde{\mathrm{K}}_\theta^{\mathrm{T}} \widetilde{\mathrm{K}}_\theta)\}, \quad x_k = \mathrm{D}_\theta^{1/2} w_k, \tag{8.7}$$

where

$$\mathscr{K}_k(\widetilde{\mathrm{K}}_\theta^{\mathrm{T}} b, \widetilde{\mathrm{K}}_\theta^{\mathrm{T}} \widetilde{\mathrm{K}}_\theta) = \mathrm{span}\{(\widetilde{\mathrm{K}}_\theta^{\mathrm{T}} \widetilde{\mathrm{K}}_\theta)^\ell \widetilde{\mathrm{K}}_\theta^{\mathrm{T}} b \mid 0 \le \ell \le k-1\},$$

is the kth Krylov subspace associated with the vector $\widetilde{\mathrm{K}}_\theta^{\mathrm{T}} b$ and the matrix $\widetilde{\mathrm{K}}_\theta^{\mathrm{T}} \widetilde{\mathrm{K}}_\theta$. The quantity $b - \widetilde{\mathrm{K}}_\theta w_k$ whose norm is minimized is the discrepancy vector corresponding to w_k; in the traditional inverse problems literature, the Morozov discrepancy principle states that the iterations should be stopped right before the norm of the discrepancy falls below the noise level. Recalling that the standard deviation of the m-variate white noise is $\sqrt{m}$, the Morozov stopping criterion can be written as

$$\|b - \widetilde{\mathrm{K}}_\theta w_k\| \le \sqrt{m}.$$

On the other hand, letting

$$G(w) = \left\| \begin{bmatrix} \mathrm{K}\mathrm{D}_\theta^{1/2} \\ \mathrm{I}_n \end{bmatrix} w - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|^2 = \|b - \mathrm{K}\mathrm{D}_\theta^{1/2} w\|^2 + \|w\|^2, \tag{8.8}$$

denote the norm of the discrepancy of the original linear system (8.5), it follows that the least squares solution of the original problem minimizes the functional G. It is therefore natural to monitor the behavior of $G(x_k)$ as the iterations proceed, and continue iterating only as long as $G(w_k)$ keeps decreasing. While it is known that the norm of the discrepancy of (8.5) decreases and $\|w_k\|$ increases with the number of CGLS iterations, we do not know a priori how the increase/decrease rates are related to each other, so without further analysis, it is not clear if a minimum is reached before the maximum allowed number of iterations.

In light of these observations, we propose the following approximation to the least squares solution of (8.5) via the approximate solution of (8.6).

Definition 5. *The* reduced Krylov subspace *(RKS) solution for the problem* (8.5) *is w_{k_0} defined by* (8.7)*, with k_0 chosen to be the first index k satisfying the criterion*

$$(\mathscr{C}): \quad \|b - \widetilde{\mathrm{K}}_\theta w_{k+1}\| \leq \sqrt{m}, \qquad \text{or} \qquad G(w_{k+1}) > \tau G(w_k),$$

where $\tau - 1 = \epsilon > 0$ is a small safeguard parameter.

In the following, we refer to the IAS algorithm as *approximate IAS* when the minimization of (8.8) is replaced by the RKS solution, as opposed to the original *exact IAS*.

Step 2: Updating θ

The update of the prior variance θ is based on the first order optimality condition. Since the parameters θ_j are mutually independent, the update can be carried out separately for each component. It follows from the form of the MAP objective function that the updated jth component of θ must satisfy the algebraic equation

$$\frac{\partial \mathcal{F}}{\partial \theta_j} = -\frac{1}{2}\frac{x_j^2}{\theta_j^2} - \left(r\beta - \frac{3}{2}\right)\frac{1}{\theta_j} + r\frac{\theta_j^{r-1}}{\vartheta_j^r} = 0, \quad x_j = x_j^{t+1}. \tag{8.9}$$

There are combinations of the hyperparameter values for which the solution is available in closed form. We derive a computationally efficient form for the general case in the ensuing discussion.

The IAS algorithm has a similarity to a class of a reweighted least squares methods [79, 56], or fixed point iterative methods with lagged diffusivity [57], providing iterative algorithms to compute ℓ_1-regularized solutions to inverse problems. For similar alternating iterative methods in the Bayesian framework, we refer to [4, 8] applied to compressed sensing and imaging.

8.2 Scaling

The analyses of the IAS algorithm published in [23, 32], were limited to some specific hyperpriors from the generalized gamma family. Before extending the analysis to the family of generalized gamma hyperpriors, we reformulate the problem in non-dimensional form. To that end we introduce non-dimensional parameters z_j and ξ_j such that

$$x_j = \vartheta_j^{1/2} z_j, \quad \theta_j = \vartheta_j \xi_j,$$

and express the objective function in terms of these variables as

$$\Phi(z,\xi) = \frac{1}{2}\|\widehat{\mathrm{K}}z - b\|^2 + \frac{1}{2}\sum_{j=1}^n \frac{z_j^2}{\xi_j} - \left(r\beta - \frac{3}{2}\right)\sum_{j=1}^n \log \xi_j + \sum_{j=1}^n \xi_j^r, \tag{8.10}$$

where $\widehat{\mathrm{K}}$ is a column-scaled version of K, that is,

$$\widehat{\mathrm{K}} = \left[\ \vartheta_1^{1/2} a^{(1)}, \cdots, \vartheta_n^{1/2} a^{(n)}\ \right] = \mathrm{K}\,\mathrm{diag}(\vartheta_1^{1/2}, \cdots, \vartheta_{\mathrm{n}}^{1/2}).$$

Column scaling the forward map is a common practice in some geophysics and biomedical imaging applications, where it has been motivated by sensitivity considerations. Define the sensitivity of the forward map $x \mapsto \mathrm{K}x$ with respect to the jth component x_j as

$$s_j = \left\| \frac{\partial(\mathrm{K}x)}{\partial x_j} \right\| = \|\mathrm{K}e_j\| = \|a^{(j)}\|,$$

where e_j is the jth canonical unit vector in $\mathbb{R}^n$ and $a^{(j)}$ is the jth column of the matrix K. Then, weighting the component x_j by the corresponding sensitivity scalar can be seen as a way to avoid favoring solutions with support concentrated near the receiver locations. While this observation can be used as a guidance for selecting the value of the hyperparameters ϑ_j, in the Bayesian framework this reasoning was considered problematic because the selection of the prior should not depend on the observation model. A Bayesian justification for such choice of ϑ in the case where $r = 1$ has been given recently in [32, 24]. The following theorem generalizes the result to the case of general gamma hyperpriors. The result is formulated in terms of the signal-to-noise ratio (SNR) of the inverse problem (8.1),

$$\mathrm{SNR} = \frac{\mathsf{E}\{\|b\|^2\}}{\mathsf{E}\{\|\varepsilon\|^2\}}.$$

Theorem 1. *(a) Assuming that a support set $S \subset \{1, 2, \ldots, n\}$ is given, the SNR conditional to the unknown x being supported on S, denoted by SNR_S, is given by*

$$\mathrm{SNR}_S = \frac{\sum_{j\in S} \nu(r,\beta)\vartheta_j}{\mathrm{trace}(\Sigma)} + 1, \quad \nu(r,\beta) = \frac{\gamma(\beta + 1/r)}{\gamma(\beta)},$$

provided that $\beta > -1/r$.

(b) Let $p_k = \mathbb{P}\{\|x\|_0 = k\}$ denote the probability that the support of the signal has cardinality k, for $k = 1, 2, \ldots, n$. Then the exchangeability condition ($\mathscr{E}$),

$$(\mathscr{E}): \quad \mathrm{SNR}_S = \mathrm{SNR}_{S'} \textit{ whenever } S \textit{ and } S' \textit{ are of the same cardinality,}$$

is satisfied if and only if ϑ_j is chosen as

$$\vartheta_j = \frac{C}{\|a^{(j)}\|^2}, \quad C = \frac{(\overline{\mathrm{SNR}} - 1)\mathrm{trace}(\Sigma)}{\nu(r,\beta)} \sum_{k=1}^{n} \frac{p_k}{k}. \tag{8.11}$$

Proof. The proof is a slight modification of that for the gamma hyperprior ($r = 1, \beta > 3/2$) in [32] to account for the fact that if θ_j follows the generalized gamma distribution, then

$$\mathsf{E}\{\theta_j\} = \nu(r,\beta)\vartheta_j.$$

□

An important corollary of the above theorem is that, under the stated assumptions, scaling the columns $a^{(j)}$ by $\vartheta_j^{1/2}$ is tantamount to making them all of the same norm $\sqrt{C}$. From the point of view of linear inverse problems, this scaling renders the data equally sensitive to each component of the unknown x.

Furthermore, as already pointed out in [32], the theorem provides a Bayesian argument to choose the value of Tikhonov regularization parameter in linear inverse problems from an estimated SNR and a priori belief about the cardinality of the support. The effect of the sensitivity scaling, which has been first demonstrated in [32] by computed examples, will be again highlighted in Chapter 10.

8.3 Variance updating: a closer look

In this section, we analyze in detail the process of updating the variance vector given an updated estimate of the signal. After scaling the variables as described in the previous section to arrive at a non-dimensional formulation, the algebraic relation (8.9) for the non-dimensional variance ξ_j given the non-dimensional signal z_j becomes

$$-\frac{1}{2}z^2 - \eta\xi + r\xi^{r+1} = 0, \quad \eta = r\beta - \frac{3}{2}, \tag{8.12}$$

where we omit the subscript j to simplify the notation. Since the expression depends only on the square of z, we restrict our discussion to the case where z assumes non-negative values, the negative values being covered by symmetry.

The following result characterizes the variance as the solution of an initial value problem.

Lemma 2. *If $r < 0$ and $\eta < -3/2$, or $r > 0$ and $\eta > 0$, formula (8.12) defines an implicit function*

$$\varphi(z) = \xi, \quad \varphi : \mathbb{R}_+ \to \mathbb{R}_+,$$

which is smooth and strictly increasing. Moreover, ξ is the solution of the initial value problem

$$\varphi'(z) = \frac{2z\varphi(z)}{2r^2\varphi(z)^{r+1} + z^2}, \quad \varphi(0) = \left(\frac{\eta}{r}\right)^{1/r}. \tag{8.13}$$

Proof. Starting from (8.12), we define the function

$$g(\xi) = \xi(r\xi^r - \eta), \quad 0 < \xi_0 = \left(\frac{\eta}{r}\right)^{1/r} \leq \xi < \infty,$$

which is differentiable, with $g'(\xi) = -\eta + r(r+1)\xi^r$. For $r \leq -1$ and $\eta < -3/2$, the derivative is always positive. For $-1 < r < 0$ and $\eta < -3/2$, the condition $\xi > \xi_0$ implies that $\xi^r < \eta/r$, and consequently, $g'(\xi) > r\eta > 0$, and for $r > 0$ and $\eta > 0$, we have $\xi^r < \eta/r$, and therefore $g'(\xi) > \eta r > 0$. Hence, the function $g(\xi)$ is strictly increasing for $\xi > \xi_0$. Furthermore,

$$g(\xi_0) = 0, \quad \lim_{\xi\to\infty} g(\xi) = \infty.$$

Therefore the equation

$$g(\xi) = \xi(r\xi^r - \eta) = \frac{1}{2}z^2$$

has a unique solution $\xi = \xi(z) \in [\xi_0, \infty)$ for every $z \geq 0$, hence the mapping from z to the solution (8.12) defines a strictly increasing function $\xi = \varphi(z)$. We have $\varphi(z) = g^{-1}(z^2/2)$ and by the implicit function theorem, the function g^{-1}, as an inverse of a differentiable function is differentiable, the function φ is differentiable. Substituting $\xi = \varphi(z)$ in (8.12) ,

$$-\frac{1}{2}z^2 - \eta\varphi(z) + r\varphi(z)^{r+1} = 0, \tag{8.14}$$

and differentiating with respect to z, we get

$$((r+1)r\varphi(z)^r - \eta)\varphi'(z) = z,$$

or, equivalently,

$$\left(r^2\varphi(z)^r + \frac{1}{\varphi(z)}(r\varphi(z)^{r+1} - \eta\varphi(z))\right)\varphi'(z) = \left(r^2\varphi(z)^r + \frac{z^2}{2\varphi(z)}\right)\varphi'(z) = z,$$

yielding the differential equation (8.13). □

The outline of the IAS scheme with informative hyperprior is given in Algorithm 7.

Algorithm 7: IAS with informative hypeprior

```
input: observed signal $b \in \mathbb{R}^m$
output: restored signal $x^*$
initialize: set $\theta^{(0)} = \vartheta$
for $t = 1, 2, \ldots$ until convergence do:
    · update $x^{(t)}$ by solving (8.5) in terms of $w$ and rescaling, $x^{(t)} = \mathrm{D}_\theta^{1/2} w$
    for $j = 1, \ldots n$ do:
        · update $\theta_j^{(t)}$ by solving (8.13) in terms of $\xi_j$ and rescaling, $\theta_j = \vartheta_j \xi_j$
    end for
end for
return: $x^* = x^{(t)}$, $\theta^* = \theta^{(t)}$
```

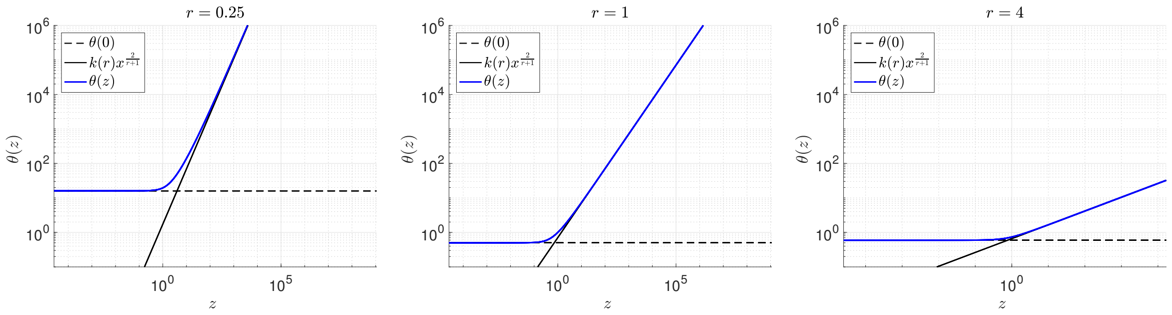
The characterization of ξ in terms of the differential equation (8.13) can be used to compute effectively the values of the updates of the variances in the IAS algorithm, as we will show in the computed examples. Moreover, Lemma 2 makes it possible to analyze the asymptotic behavior of the variance parameter when the corresponding value of z is either close to zero or very large.

Lemma 3. *The asymptotic behavior of φ when z is close to zero is*

$$\varphi(z) = \left(\frac{\eta}{r}\right)^{1/r} + \frac{1}{2\eta r}z^2 + \mathcal{O}(z^4). \tag{8.15}$$

whereas the asymptotics for $z > 0$ large is

$$\varphi(z) = \kappa z^{2/(r+1)}\left(1 + o(1)\right), \quad \kappa = \left(\frac{1}{2r}\right)^{1/(r+1)} \tag{8.16}$$



(a) (b) (c)

Figure 8.1: Logarithmic plots of the updating functions with different values of the parameters r, with $\eta = 0.5$ in each case. The asymptotics given by Lemma 3 as well as the initial values $\varphi(0) = (\eta/r)^{1/r}$ are indicated in this figure.

when $r > 0$, *and*

$$\varphi(z) = \kappa z^2 \left(1 + o(1)\right), \quad \kappa = \frac{1}{2|\eta|}$$

when $r < 0$.

Proof. The asymptotic behavior of φ for z near zero can be obtained from its Taylor expansion at $z = 0$. It follows from (8.13) that $\varphi(0) = \left(\frac{\eta}{r}\right)^{1/r}$, $\varphi'(0) = 0$, and differentiating (8.13) with respect to z yields

$$\varphi''(0) = \frac{1}{r^2 \varphi(0)^r} = \frac{1}{r\eta}.$$

The asymptotic estimate follows from the observation that the third derivative of φ vanishes at $z = 0$.

To obtain the asymptotics of φ for large z and $r > 0$, observe that (8.14) implies

$$\lim_{z \to \infty} \varphi(z) \to \infty,$$

therefore, since $(1 + o(1))^{-1} = 1 + o(1)$, for large z,

$$\frac{1}{2} z^2 = \varphi(z)^{r+1} \left(r - \frac{\eta}{\varphi(z)^r} \right) = r\varphi(z)^{r+1} \left(1 + o(1)\right),$$

implying (8.16). Similarly, if $r < 0$, we write

$$\frac{1}{2} z^2 = \xi \left(|\eta| - \frac{|r|}{\xi^{|r|}} \right) = |\eta| \xi \left(1 + o(1)\right),$$

completing the proof. □

Figure 8.1 shows the updating functions in a logarithmic scale with selected values of the parameter r.

The asymptotic behavior of the updating function is helpful for understanding the role of the model parameters r and β. For this interpretation, we need the following theorem establishing the equivalence of the IAS optimization of the objective function with respect to the pair (z, ξ) in $\mathbb{R}^{2n}$ and the optimization of the objective function along the manifold $\xi = \varphi(z)$, with the last equality to be understood as componentwise.

Lemma 4. *Let (z^*, ξ^*) be a local minimizer of the objective function $\Phi(x, \xi)$ given by (8.10). Then, the point z^* is a local minimizer of $\Psi(z) = \Phi(z, \varphi(z))$. Conversely, if z^* is a local minimizer of $\Psi(z)$, then $(z^*, \varphi(x^*))$ is a local minimizer of $\Phi(x, \xi)$.*

Proof. If (z^*, ξ^*) is a local minimizer of Φ, then it must satisfy

$$\frac{\partial \Phi}{\partial \xi_j}(z^*, \xi^*) = 0,$$

implying that $\xi^* = \varphi(z^*)$. Let $U = B_1 \times B_2 \in \mathbb{R}^{2n}$ be a neighborhood of (z^*, ξ^*) such that for any $(z, \xi) \in U$, $\Phi(z^*, \xi^*) \leq \Phi(z, \xi)$. Since φ is continuous, for each z in some neighborhood $B_1' \subset B_1$ of z^*, $\varphi(z) \in B_2$, therefore $\Phi(z^*, \varphi(z^*)) \leq \Phi(z, \varphi(z))$, that is, z^* is a local minimizer of Ψ.

Conversely, let z^* be a local minimizer of Ψ. Then, there is a neighborhood B of z^* such that for any $z \in B$, $\Phi(z^*, \varphi(z^*)) \leq \Phi(z, \varphi(z))$. However, for each z, $\theta = \varphi(z)$ is the unique minimizer of $\theta \mapsto \Phi(z, \theta)$, therefore

$$\Phi(z^*, \varphi(z^*)) \leq \Phi(z, \varphi(z)) \leq \Phi(z, \theta), \quad (z, \theta) \in B \times \mathbb{R}^+,$$

implying that $(z^*, \varphi(z^*))$ indeed is a local minimizer of Φ. □

It follows from the lemma that in order to study the sparsity promoting properties of the various hyperpriors, one can consider the objective function $\Psi(z) = \Phi(z, \varphi(z))$, and in particular, the scaled penalty term

$$\Pi(z) = \frac{1}{2} \sum_{j=1}^{n} \frac{z_j^2}{\varphi(z_j)} - \eta \sum_{j=1}^{n} \log \varphi(z_j) + \sum_{j=1}^{n} \varphi(z_j)^r.$$

We will use this observation together with the asymptotic forms of the updating function to elucidate how the regularization properties of the penalty functions change with the hyperparameter values. Before addressing the general case, we consider some special choices of the parameter values.

8.3.1 Special generalized gamma hyperpriors

There are hyperparameter combinations for which the updating function for the variances is available in closed form. Some of these special cases have been used in numerical computations in earlier works [30, 21].

Gamma distribution and ℓ_1 prior

The most thoroughly analyzed hyperprior in the context of the IAS algorithm is the gamma distribution, which is a generalized gamma with $r = 1$ and $\eta > 0$. With that choice of parameters, equation (8.9) simplifies to

$$-\frac{1}{2}z^2 - \eta\xi + \xi^2 = 0,$$

and can be readily solved for ξ, yielding

$$\xi = \varphi(z) = \frac{1}{2}\left(\eta + \sqrt{\eta^2 + 2z^2}\right).$$

As pointed out in [25, 32], substituting $\xi_j = \varphi(z_j)$ in the MAP penalty function and letting η go to zero yields

$$\Pi(z) = \sum_{j=1}^n \left\{ \frac{1}{2}\frac{z_j^2}{\xi_j} \quad - \quad \eta \log \xi_j + \xi_j \right\} = \sum_{j=1}^n \left\{ \frac{z_j^2}{\eta + \sqrt{\eta^2 + 2z_j^2}} \right.$$
$$\left. -\eta \log \frac{1}{2}(\eta + \sqrt{\eta^2 + 2z_j^2}) + \frac{1}{2}(\eta + \sqrt{\eta^2 + 2z_j^2}) \right\}$$
$$\to \sqrt{2}\sum_{j=1}^n |z_j|, \text{ as } \eta \to 0+,$$

that is, in the limit, the penalty function approaches the ℓ_1-penalty. In [32], it was further shown that the unique solution of the IAS algorithm converges to the solution with the ℓ_1-penalty, thus recovering a compressible solution, if the data came from a sparse generative model. For further results, we refer to [32].

Inverse gamma distribution and Student prior

The second special case is that of the inverse gamma hyperprior, corresponding to setting $r = -1$. In this case, equation (8.9) becomes

$$\frac{1}{2}z^2 - \eta\xi - 1 = 0,$$

and the update formula is

$$\xi = \varphi(z) = \frac{1}{2k}(z^2 + 2), \quad k = \beta + \frac{3}{2}.$$

As for the gamma hyperprior, substituting $\xi_j = \varphi(x_j)$ in the MAP penalty functional yields

$$\begin{aligned} \Pi(z) &= \sum_{j=1}^n \left\{ \frac{1}{2}\frac{z_j^2}{\xi_j} + k \log \xi_j + \frac{1}{\xi_j} \right\} = \sum_{j=1}^n \left\{ \frac{z_j^2 + 2}{2\xi_j} + k \log \xi_j \right\} \\ &= n(k - \log 2k) + \sum_{j=1}^n \log(z_j^2 + 2)^k, \end{aligned}$$

which corresponds to the prior model

$$\mathrm{P}_{\mathrm{prior}}(z) \propto \exp(-\Pi(z)) \propto \prod_{j=1}^n \frac{1}{(z_j^2 + 2)^k},$$

We observe that as $\beta \to 0+$, $k \to 3/2$, and the distribution of the individual components z_j approaches the Student distribution,

$$St(z \mid \nu) \propto \frac{1}{\left(1 + \frac{z^2}{\nu}\right)^{(\nu+1)/2}},$$

with parameter $\nu = 2$, a prominently fat tailed distribution that favors outliers, thus promoting sparsity.

Generalized gamma and ℓ_p prior

The third special case that we consider here is $r\beta = 3/2$, in which the update formula becomes

$$-\frac{1}{2}z^2 + r\xi^{r+1} = 0,$$

or

$$\xi = \varphi(z) = \frac{|z|^{2/(r+1)}}{(2r)^{1/(r+1)}}.$$

Substituting $\xi_j = \varphi(z_j)$ in the MAP penalty functional we get

$$\begin{aligned}\Pi(z) &= \sum_{j=1}^n \left\{ \frac{1}{2}\frac{z_j^2}{\xi_j} + \xi_j^r \right\} = \sum_{j=1}^n \left\{ \frac{(2r)^{1/(r+1)}}{2}|z_j|^{2-2/(r+1)} + \frac{1}{(2r)^{1/(r+1)}}|z_j|^{2r/(r+1)} \right\} \\ &= C_r \sum_{j=1}^n |z_j|^{2r/(r+1)}, \quad C_r = \frac{r+1}{(2r)^{r/(r+1)}}.\end{aligned}$$

and letting

$$p = \frac{2r}{r+1}, \quad 0 < p < 2,$$

yields

$$\Pi(z) = C_r \sum_{j=1}^n |z_j|^p, \quad 0 < p < 2$$

The ℓ_p-penalties for $0 < p < 1$ are known for their sparsity promoting properties, and have been analyzed extensively in the literature. However, since are non-convex, they pose challenges when it comes to computing the corresponding regularized solution.

8.3.2 General case: asymptotics

Consider now the penalty functional $\Pi(z) = \sum_{j=1} \Pi_j(z_j)$ in the general case with $r > 0$. From Lemma 3 we see that if $|z_j|$ is large, the penalty function of the jth component can be written as

$$\begin{aligned}\Pi_j(z_j) &= \frac{1}{2}\frac{z_j^2}{\varphi(z_j)} + \varphi(z_j)^r - \eta \log \varphi(z_j) \\ &= \frac{1}{2\kappa}|z_j|^{2-2/(r+1)}(1+o(1)) + \kappa^r |z_j|^{2r/(r+1)}(1+o(1)) \\ &\qquad - \frac{2\eta}{r+1}\log(|z_j|(1+o(1))) \\ &\propto |z_j|^p(1+o(1)), \quad p = \frac{2r}{r+1}.\end{aligned}$$

Similarly, for small values of $|z_j|$, (8.15) yields the asymptotic estimate

$$\begin{aligned}\Pi_j(z_j) &= \frac{1}{2}\frac{z_j^2}{a + bz_j^2 + \mathcal{O}(z_j^4)} + (a + bz_j^2 + \mathcal{O}(z_j^4))^r - \eta \log(a + bz_j^2 + \mathcal{O}(z_j^4)) \\ &= C_1 + C_2 z_j^2 + \mathcal{O}(z_j^4))\end{aligned}$$

with $a = (\eta/r)^{1/r}$, $b = 1/(2\eta r)$, and C_1 and C_2 are some scalars. Therefore, for large $|z_j|$, the penalty behaves like an ℓ_p-penalty with $p = 2r/(r+1) \in (0, 2)$, while for small $|z_j|$, the penalty is essentially Gaussian.

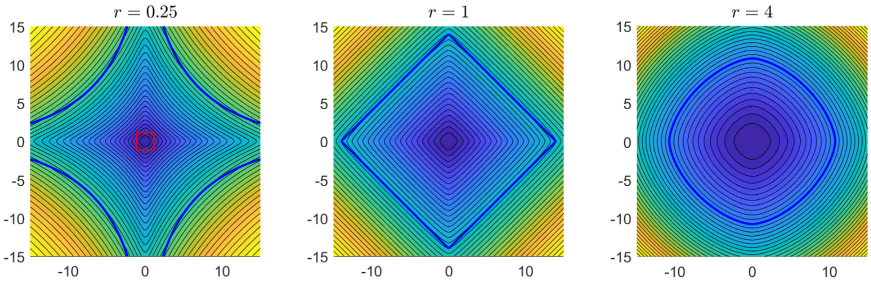


Figure 8.2: Level set plots of the reduced penalty function $\Pi(z)$ in two dimensions with different values of r, corresponding asymptotically to ℓ_p-penalties with $p = 2/5$ (left), $p = 1$ (center) and $p = 8/5$ (right). The corresponding ℓ_p-sphere is superimposed with dark blue. The boundary of the convexity region (see Section 8.4) for $r = 1/4$ is marked by the red square.

Figure 8.2 shows the level curves of the function $\Pi(z)$ in two dimensions for some parameter choices. From these plots it is clear that for large values of $\|z\|$, the level sets look like the ℓ_p spheres, while for small values, the level curves become increasingly circular as predicted by the asymptotic formulas.

8.4 Convexity

Our first goal in this section is to find out for which choices of the parameters (r, β) the objective function Φ given by (8.10) is globally convex for all $(z, \xi) \in \mathbb{R}^n \times \mathbb{R}^n_+$, or, alternatively, convex in a specified subset. The following theorem summarizes the results [27].

Theorem 2. *Let $\beta > 0$ and $r \neq 0$, and let $\Phi(z, \xi) = \Phi(z, \xi \mid r, \beta)$ be the objective function (8.10) for the dimensionless formulation of the problem.*

(a) If $r \geq 1$ and $\eta = r\beta - 3/2 > 0$, the function $\Phi(z, \xi)$ is convex everywhere.

(b) If $0 < r < 1$ and $\eta = r\beta - 3/2 > 0$, or, if $r < 0$ and $\beta > 0$, the function $\Phi(z, \xi)$ is convex provided that

$$\xi_j < \overline{\xi} = \left(\frac{\eta}{r|r-1|}\right)^{1/r}.$$

Proof. Recall that the positive definiteness of the Hessian is a sufficient condition for the convexity of the underlying functional. Consider the block partitioning of the Hessian of Φ,

$$\mathrm{H} = \mathrm{H}(z, \xi) = \begin{bmatrix} \nabla_z \nabla_z \Phi(z, \xi) & \nabla_z \nabla_\xi \Phi(z, \xi) \\ \nabla_\xi \nabla_z \Phi(z, \xi) & \nabla_\xi \nabla_\xi \Phi(z, \xi) \end{bmatrix},$$

where,

$$\begin{aligned} \nabla_z \nabla_z \Phi(z, \xi) &= \mathrm{D}_\xi^{-1} + \widehat{\mathrm{K}}^{\mathrm{T}} \widehat{\mathrm{K}}, \\ \nabla_z \nabla_\xi \Phi(z, \xi) &= \nabla_\xi \nabla_z \Phi(x, \theta) = \mathrm{diag}\Big(-\frac{z_j}{\xi_j^2}\Big), \\ \nabla_\xi \nabla_\xi \Phi(z, \xi) &= \mathrm{diag}\Big(\frac{z_j^2}{\xi_j^3} + r(r-1)\xi_j^{r-2} + \eta \frac{1}{\xi_j^2}\Big), \end{aligned}$$

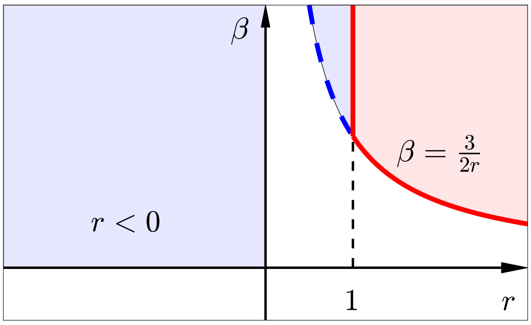


Figure 8.3: Convexity regions in the (r, β) plane. The red shadowing denotes parameter choices for which the MAP objective function is convex everywhere, and the blue shadowing parameter choices for which the MAP objective function is only locally convex. The curve $\beta = 3/(2r)$ marks the parameter pairs for which the hierarchical model is an ℓ_p-penalty priors, with $p = 2r/(r+1)$, which are convex if $p \geq 1$ or, equivalently, $r \geq 1$. The vertical line $r = 1$ corresponds to the family of gamma hyperpriors.

For any vector $q = \begin{bmatrix} u \\ v \end{bmatrix} \in \mathbb{R}^{2n}$, we have

$$\begin{aligned} q^T \mathrm{H} q =& \|\widehat{\mathrm{K}}u\|^2 + \sum_{j=1}^n \frac{u_j^2}{\xi_j} + \sum_{j=1}^n \left(\frac{z_j^2}{\xi_j^3} v_j^2 + r(r-1)\xi_j^{r-2} v_j^2 + \eta \frac{v_j^2}{\xi_j^2} \right) - 2 \sum_{j=1}^n \frac{z_j}{\xi_j^2} u_j v_j \\ =& \|\widehat{\mathrm{K}}u\|^2 + \sum_{j=1}^n \frac{1}{\xi_j} \left(u_j - \frac{z_j}{\xi_j} v_j \right)^2 + \sum_{j=1}^n \phi_j(\xi_j \mid r, \beta) v_j^2, \end{aligned} \tag{8.17}$$

where

$$\phi_j(\xi_j \mid r, \beta) = r(r-1)\xi_j^{r-2} + \eta \frac{1}{\xi_j^2}.$$

Note that the first two terms in (8.17) are always non-negative, so the positivity of the quadratic form defined by the Hessian it is guaranteed if $\phi_j(\xi_j \mid r, \beta) > 0$ for all j, $1 \leq j \leq n$. The proof for the different cases follows by enforcing this condition. □

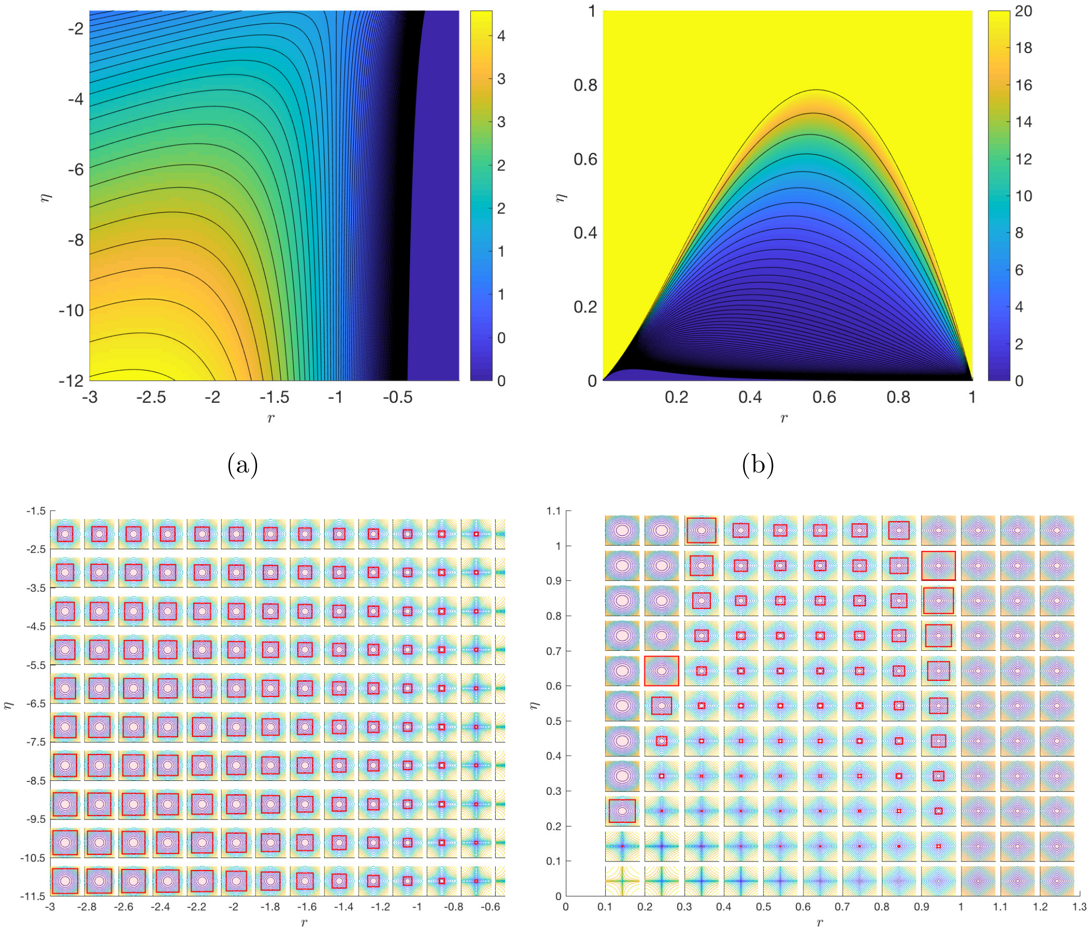
Figure 8.3 shows the regions in the r, β plane corresponding to hyperparameter choices leading to convex or conditional convex MAP objective functions. Observe that the ℓ_p-penalty corresponds to the boundary $\beta = 3/(2r)$, with $p = 2r/(r+1)$. In particular, for $p \leq 1$, the generalized gamma family provides nearby penalty functionals that yield at leas a locally convex objective function.

We define the *convexity radius* $\rho = \rho(r, \beta) \geq 0$, by

$$\rho = \varphi^{-1}(\overline{\xi}),$$

that is, for $\|z\|_\infty < \rho$, we have $\|\xi\|_\infty < \overline{\xi}$ guaranteeing the convexity. If the objective function is globally convex as in the case (a) of Theorem 2, we set $\rho = \infty$

Figure 8.4 shows graphically the convexity radius as a function of the parameters r and η, as well as the evolution of the level curves in two dimensions of the reduced objective function together with the convexity spheres.



(a)

(b)

(c)

(d)

Figure 8.4: The radius of the convexity region as a function of r and η for generalized gamma hyperpriors with$-3 \leq r < 0$ (a) and $0 < r \leq 1$ (b). The two panels on the bottom row show, for different choices of (r, η) in the generalized gamma family, the level curves of the corresponding functionals. In each tile, a red curve, if present, marks the boundary of the region inside which the functional is convex. The absence of a red curve indicates that for that choice of (r, η) the functional is always convex.

8.5 Stable convexity

Consider the IAS algorithm for computing the MAP estimate, and denote the current iterate by $(z^{(t-1)}, \xi^{(t-1)})$. The update of z requires the solution of the minimization problem

$$z^{(t)} = \arg\min_{z \in \mathbb{R}^n} \left\{ \frac{1}{2} \|\widehat{\mathrm{K}} z - b\|_2^2 + \frac{1}{2} \sum_{j=1}^{n} \frac{z_j^2}{\xi_j^{(t-1)}} \right\}.$$

We say that $\Phi(z, \xi)$ is *stably convex* if there is a $T > 0$ such that, for $t > T$,

$$\|z^{(t-1)}\|_\infty < \rho \Rightarrow \|z^{(t)}\|_\infty < \rho = \varphi^{-1}(\bar{\xi}),$$

where ρ is the convexity radius. Stable convexity is tantamount to guaranteeing that once the IAS iterates $(z^{(t-1)}, \xi^{(t-1)})$ enter the convexity basin they do not leave it, thus keeping the optimization problem convex.

To find a sufficient condition for stable convexity, we need an estimate of the ℓ^∞-norm of the least squares solution of the system

$$\begin{bmatrix} \widehat{\mathrm{K}} \\ \mathrm{D}_\xi^{-1/2} \end{bmatrix} z = \begin{bmatrix} b \\ 0 \end{bmatrix}, \quad \|\xi\|_\infty < \overline{\xi}.$$

In the following, we assume that the columns $\widehat{a}^{(j)}$ of $\widehat{\mathrm{K}}$ have been scaled according to the sensitivity and satisfy $\|\widehat{a}^{(j)}\|_2 = C^{1/2}$. The next lemma provides an estimate for the size of the components of the updated z [27].

Lemma 5. *Assume that $\xi_j < \overline{\xi}$. Then the entries of the solution z of the normal equations,*

$$(\widehat{\mathrm{K}}^{\mathrm{T}}\widehat{\mathrm{K}} + \mathrm{D}_\xi^{-1})z = \widehat{\mathrm{K}}^{\mathrm{T}}b$$

satisfy

$$|z_j| \leq \frac{C\xi_j}{1 + C\xi_j}(hC\overline{\xi} + 1)\|\widehat{\mathrm{K}}^{\mathrm{T}}b\|_2, \quad h = \max_j \{ \sum_{i \neq j} |\cos\measuredangle(\widehat{a}^{(i)}, \widehat{a}^{(j)})| \}.$$

Proof: In terms of the quadratic forms associated with the symmetric positive definite matrices, we have that

$$\widehat{\mathrm{K}}^{\mathrm{T}}\widehat{\mathrm{K}} + \mathrm{D}_\xi^{-1} \geq \mathrm{D}_\xi^{-1},$$

from which it follows that

$$(\widehat{\mathrm{K}}^{\mathrm{T}}\widehat{\mathrm{K}} + \mathrm{D}_\xi^{-1})^{-1} \leq \mathrm{D}_\xi,$$

establishing the following inequality for the induced ℓ_2-norms,

$$\|(\widehat{\mathrm{K}}^{\mathrm{T}}\widehat{\mathrm{K}} + \mathrm{D}_\xi^{-1})^{-1}\|_2 \leq \|\mathrm{D}_\xi\|_2 \leq \overline{\xi},$$

and further, the estimate

$$\|z\|_\infty \leq \|z\|_2 \leq \overline{\xi}\|\widehat{\mathrm{K}}^{\mathrm{T}}b\|_2.$$

Next we express the matrix $\widehat{\mathrm{K}}^{\mathrm{T}}\widehat{\mathrm{K}}$ as the sum of the two matrices $C\mathrm{I}$ and R containing, respectively, its diagonal and off-diagonal entries,

$$\widehat{\mathrm{K}}^{\mathrm{T}}\widehat{\mathrm{K}} = C\mathrm{I} + \mathrm{R}, \quad \mathrm{R}_{ij} = \begin{cases} C\cos\measuredangle(\widehat{a}^{(i)}, \widehat{a}^{(j)}), & i \neq j, \\ 0, & i = j. \end{cases}$$

Substituting this expression in the normal equations gives

$$\mathrm{diag}(C + 1/\xi_j)z + \mathrm{R}z = \widehat{\mathrm{K}}^{\mathrm{T}}b,$$

yielding the following upper bounds for the components of the solution,

$$|z_j| \leq \frac{\xi_j}{C\xi_j + 1}\left(|(\mathrm{R}z)_j| + |(\widehat{\mathrm{K}}^{\mathrm{T}}b)_j|\right).$$

Furthermore, since

$$|(\mathrm{R}z)_j| \leq \|\mathrm{R}\|_\infty\|z\|_\infty = \max_k|\sum_{i \neq k}\mathrm{R}_{ik}|\|z\|_\infty = Ch\|z\|_\infty,$$

replacing $\|z\|_\infty$ with its upper bound and observing that $\|(\widehat{\mathrm{K}}^{\mathrm{T}}b)_j| \leq \|\widehat{\mathrm{K}}^{\mathrm{T}}b\|_2$, we have

$$|z_j| \leq \frac{\xi_j}{C\xi_j + 1}\Big(Ch\|z\|_\infty + \|(\widehat{\mathrm{K}}^{\mathrm{T}}b)\|_2\Big) \leq \frac{C\xi_j}{1 + C\xi_j}(hC\overline{\xi} + 1)\|\widehat{\mathrm{K}}^{\mathrm{T}}b\|_2$$

thus completing the proof. □

While not definitive, the previous lemma points to some of the factors that contribute to the stable convexity. First, we observe that if $|z_j^t| \ll 1$, choosing the shape parameter $\eta > 0$ small implies that $\xi_j^t = \varphi(|z_j^t|) \ll 1$. The above lemma suggests that in the IAS iterations, small entries remain small, and therefore one can hope that they remain below the convexity bound. On the other hand, if the columns of the matrix $\widehat{\mathrm{K}}$ are almost orthogonal, $h \ll 1$ and we have an upper bound close to the norm $\|\widehat{\mathrm{K}}^{\mathrm{T}}b\|_2$ for the entries $|z_j|$. In that case, choosing the parameters (r, β) so that $\rho = \tau\|\widehat{\mathrm{K}}^{\mathrm{T}}b\|_2$ for some safeguard factor $\tau > 1$ guarantees stable convexity of the objective function. The quantity h is closely related to the mutual coherence of the matrix [61], and to the Welch bounds for frames, widely studied in frame theory and signal processing literature [162].

Remark 4. *In general, one may not have an a priori guarantee that the components of the unknown are bounded by a constant smaller than the convexity radius. However, if we know a priori that $|x_j| < M$ for some $M > 0$, we may choose the parameters (r, η) so that $\rho \geq M$, guaranteeing global convexity and thus the existence of a unique minimizer. A natural question that arises then is, how the IAS algorithm should be modified for a case in which a box constraint is part of the prior. This question is addressed in the next chapter.*

8.6 IAS with bound constraints

Consider the constrained optimization problem:

$$\text{minimize } \Phi(z, \xi) \text{ subject to the constraints } 0 \leq z \leq H,$$

for some $H > 0$. The minimizer corresponds to the MAP estimate under the belief that the components of the solution are nonnegative and not larger than H. More general box contraints can be treated in a similar way.

We begin by introducing the penalty function

$$G(z) = \begin{cases} 0, & \text{when } 0 < z \leq H, \\ \infty & \text{otherwise,} \end{cases}$$

and write the posterior density with the bound constraints as

$$\mathrm{P}(z, \xi \mid b) \propto \exp\left(-\Phi(z, \xi) - G(z)\right) = \exp\left(-\Phi_G(z, \xi)\right).$$

Following the ideas in [131, 62], consider the Moreau-Yoshida envelope of the objective function,

$$\Phi_G^\lambda(z, \xi) = \Phi(z, \xi) + G^\lambda(z),$$

where

$$G^\lambda(z) = \min_{u \in \mathbb{R}^n}\left\{G(u) + \frac{1}{2\lambda}\|z - u\|^2\right\},$$

with $\lambda > 0$ an auxiliary parameter. Its can be shown [116] that the Moreau-Yoshida envelope is differentiable with respect to z, and its gradient is of the form

$$\nabla_z \Phi_G^\lambda(z,\xi) = \nabla_z \Phi(z,\xi) + \frac{1}{\lambda}(z - \mathrm{prox}_G^\lambda(z)),$$

where the proximal operator is defined as

$$\mathrm{prox}_G^\lambda(z) = \arg\min_{v\in\mathbb{R}^n}\left\{G(v) + \frac{1}{2\lambda}\|z-v\|^2\right\} = \left\{\begin{array}{ll} z, & \text{if } G(z)=0 \\ \mathrm{Q}z, & \text{if } G(z)=\infty \end{array}\right. ,$$

and Q is the orthogonal projector onto the feasible set $[0,H]^n$. Since the derivatives of the objective function with respect to the parameters ξ_j are unaffected by the inclusion of the bounds, a natural extension of the IAS algorithm for bound constrained problems can be obtained by modifying the solution of the least squares minimization problem as follows:

Given the current ξ^t:

(a) Find $z = z^*$ by solving $\nabla_z \Phi(z, \xi^{(t-1)}) = 0$ in the least squares sense,

(b) Define $z^{(t)} = \mathrm{prox}_G^\lambda(z^*)$ as the projection of z^* onto the feasible set.

Observe that, for the computation of the MAP estimate it is not necessary to specify the auxiliary parameter λ, as the proximal operator is a projection regardless of the value of λ. It was proved in [62] that, as $\lambda \to 0+$, the posterior distribution defined in terms of the Moreau-Yoshida envelope converges in the sense of total variation towards the posterior distribution augmented by the positivity constraint.

8.7 Computed examples

In this section, we present computed examples that illustrate how the choice of the hyperprior from the generalized gamma family affects sparsity of the computed MAP solution.

Example 1 The first computed example is a one-dimensional deconvolution problem with an Airy convolution kernel. The generative model is a piecewise constant signal $f : [0,1] \to \mathbb{R}$, $f(0) = 0$, and the data consist of discrete noisy observations,

$$b_j = \int_0^1 A(s_j - t) f(t) dt + \varepsilon_j, \quad 1 \le j \le m, \quad A(t) = \left(\frac{J_1(\kappa|t|)}{\kappa|t|}\right)^2,$$

where J_1 is the Bessel function of the first kind and κ is a scaling controlling the width of the kernel. We set $\kappa = 40$, yielding a kernel with FWHM $= 0.08$. We discretize the integral as

$$\int_0^1 A(s_j - t) f(t) dt \approx \sum_{j=1}^n w_k A(s_j - t_k) f(t_k), \quad 1 \le k \le n,$$

where $t_k = (k-1)/(n-1)$ and the w_k's are the trapezoidal quadrature weights. To generate the data, we use a dense discretization with $n = n_{\mathrm{dense}} = 1253$, while the forward model used for solving the inverse problem assumes $n = 500$. The observation points are given by $s_j = (4+j)/100$, $1 \le j \le m = 91$, and the noise added is assumed to be scaled white noise,

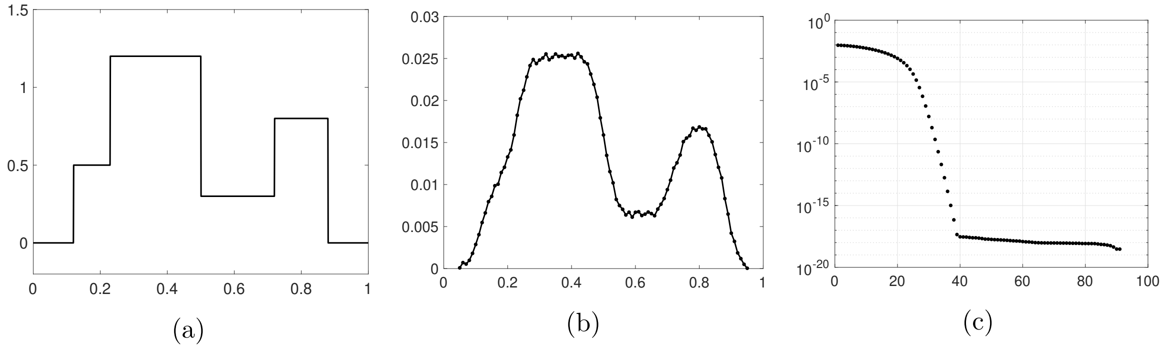


Figure 8.5: The generative model (a), the blurred and noisy data vector $b \in \mathbb{R}^{91}$ (b), singular values of the discrete blurring kernel $\mathrm{K} \in \mathbb{R}^{91\times 500}$ used for the solution of the inverse problem (c). Since only the first 30 singular values are significantly different from zero, the matrix is numerically singular.

with standard deviation σ set to 1% of the maximum of the noiseless generated signal. We denote $x_j = f(t_j)$. Figure 8.5 shows the generative signal and the data.

To compute the update of the hyperparameter ξ given the current vector z, we first sort the values of z so that $0 \leq |z_{j_1}| \leq \ldots \leq |z_{j_n}|$, and subsequently solve numerically the differential equation (8.13) at these values. Observe that this solution is fast since the propagation needs not to be restarted from zero, but rather we only need to propagate from $|z_{j_\ell}|$ to $|z_{j_{\ell+1}}|$ to get the next value. The integration was done using the RK45 solver of Matlab.

While the generative signal, a piecewise constant function, is not sparse, it admits a sparse representation in terms of its increments $z_j = x_j - x_{j-1}$ over the interval of definition. If we assume that $x_0 = 0$, then

$$z = \mathrm{B}\,x\,, \quad \mathrm{B} = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ -1 & 1 & \ldots & 0 \\ & & \ddots & \\ 0 & \ldots & -1 & 1 \end{bmatrix} \in \mathbb{R}^{n\times n}, \tag{8.18}$$

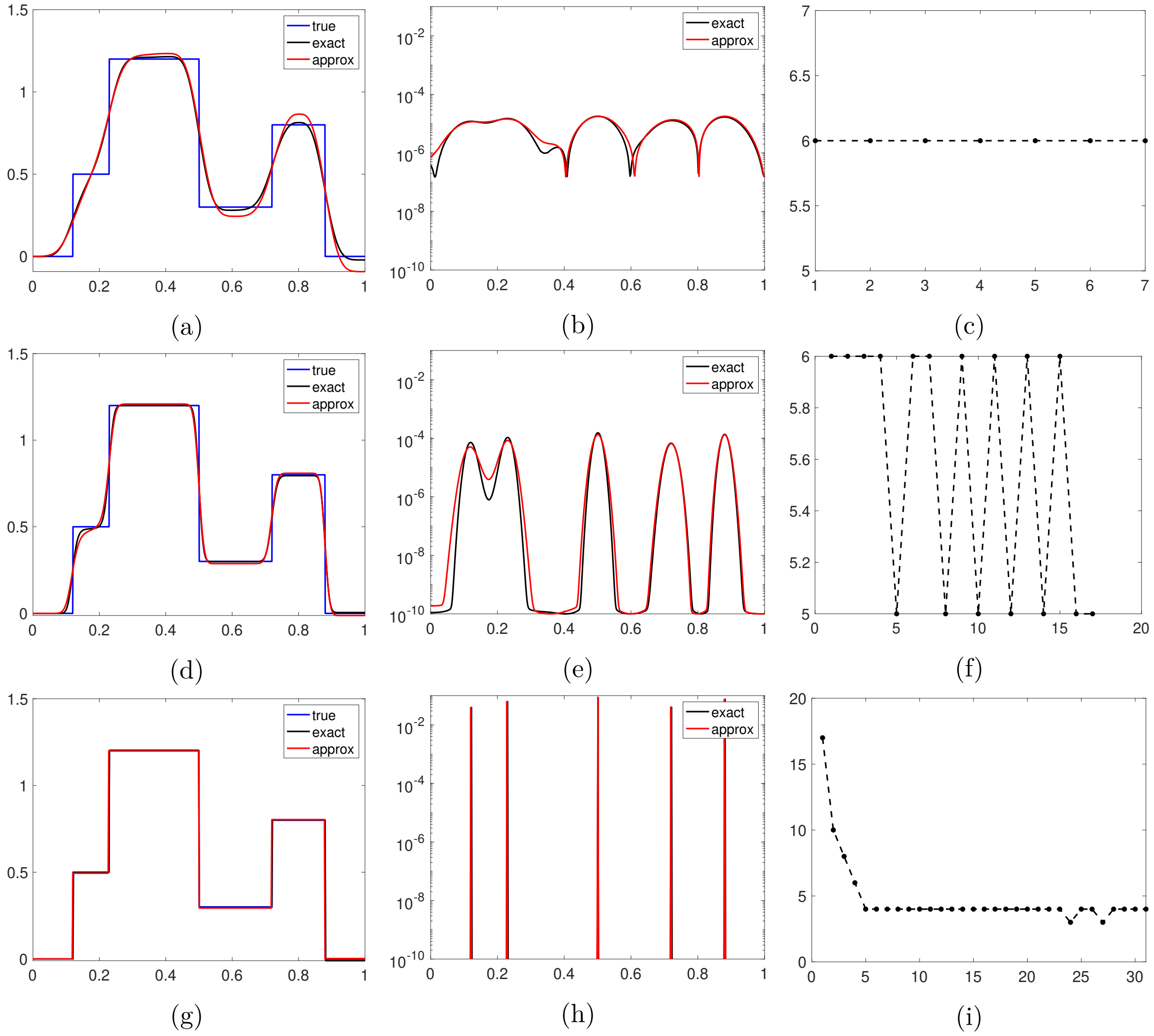
hence

$$x = \mathrm{C}\,z \quad \text{with} \quad \mathrm{C} = \mathrm{B}^{-1} = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 1 & 1 & \ldots & 0 \\ \vdots & & \ddots & \\ 1 & \ldots & 1 & 1 \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

Therefore our inverse problem is to estimate the vector z, assumed to be sparse, from the data vector b, given the forward model

$$b = \mathrm{K}\mathrm{C}z + e, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I}), \quad \mathrm{K}_{jk} = w_k A(s_j - t_k).$$

To illustrate how the sparsity of the MAP estimate determined by the IAS algorithm is affected by the choice of the hyperprior in the generalized gamma family, we show the results with the hyperpriors corresponding to $r = 3$, $r = 1$ and $r = 0.5$, see Figure 8.6. The results clearly demonstrate that with decreasing r, the sparsifying properties are strengthened. Observe that the dramatic decrease of the CGLS iterations, compared to the numerical rank of the matrix, makes the approximate IAS very attractive for large problems.

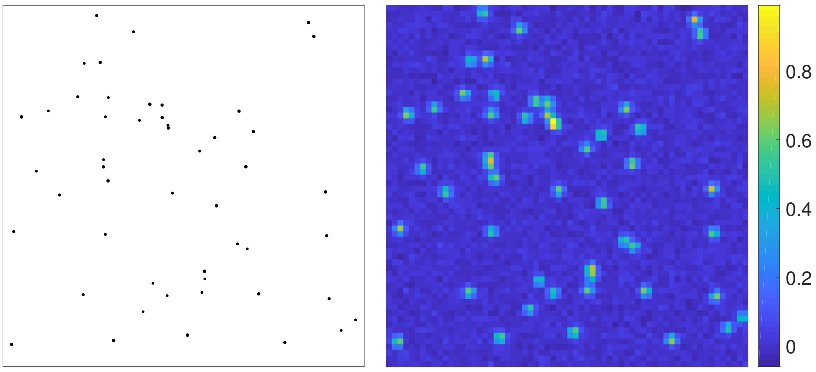


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 8.6: Reconstructions of the signal x (left), the hyperparameter θ (center) and the count of CGLS iterations per each IAS update when the approximate method is employed. In the top row, the parameter values are $r = 3$ and $\eta = 10^{-5}$, in the middle row, $r = 1$ and $\eta = 10^{-5}$, and in the bottom row $r = 1/2$ and $\eta = 10^{-5}$. The results with both the exact and approximate IAS are shown.



(a) (b)

Figure 8.7: The generative model (a), an impulse image of 50 point sources with variable amplitude. The 64×64 blurred and noisy observation, degraded by Gaussian blur and additive while Gaussian noise, scaled so as to achieve SNR ≈ 25, corresponding to a standard deviation of about 1.8% of the maximum noiseless signal (b).

Example 2 In the second example, we consider the problem of estimating a nearly black two-dimensional object. The generating model is an impulse image, defined as a distribution on $\Omega = [0,1] \times [0,1]$,

$$d\mu(p) = \sum_{k=1}^{J} a_k \delta(p - p_k) dp, \quad p_k \sim \text{Uniform}(\Omega), \quad a_k \sim \text{Uniform}([1.5, 2]),$$

and we assume that the distribution is observed with a Gaussian convolution kernel,

$$A(p, p') = \frac{1}{2\pi w^2} e^{-\|p-p'\|^2/2w^2}, \quad w = 0.01, \tag{8.19}$$

the discrete and noisy data being given at observation points $q_j \in \Omega$ by

$$b_j = \int_\Omega A(q_j, p') d\mu(p') + \varepsilon_j = \sum_{k=1}^{K} a_k A(q_j, p_k) + \varepsilon_j.$$

To solve the inverse problem, we divide the image Ω in $n = 128 \times 128 = 16\,384$ pixels, denoted by Ω_ℓ, and discretize the kernel, approximating

$$\int_\Omega A(q_j, p) d\mu(p) \approx \sum_{\ell=1}^{n} \underbrace{|\Omega_\ell| A(q_j, q'_\ell)}_{=\mathrm{K}_{j\ell}} x_\ell, \quad x_\ell = \frac{1}{|\Omega_\ell|} \int_{\Omega_\ell} d\mu(p),$$

where q'_ℓ denotes the center point of the pixel Ω_ℓ and $|\Omega_\ell|$ is its area. In this example, we assume that the number of observation points is $m = 64 \times 64 = 3\,844$, hence the forward model is defined by a matrix $\mathrm{K} \in \mathbb{R}^{m \times n}$. The noiseless signal is then corrupted by scaled white Gaussian noise with standard deviation approximately 1.8% of the maximum of the noiseless signal. In this case, since the signal itself is sparse, no change of variable is needed. Figure 8.7 shows the positions of the point masses in the true impulse image, as well as the noisy blurred image with kernel width $w = 0.01$.

We consider three hyperpriors from the generalized gamma family, corresponding to $r = 1$, $r = 0.5$, and $r = -1$. To promote sparsity of the first two cases se set $\eta = 10^{-5}$, while in the

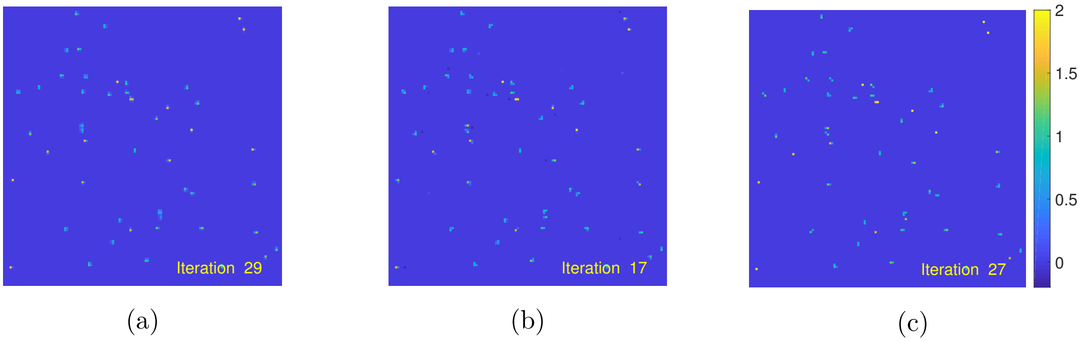


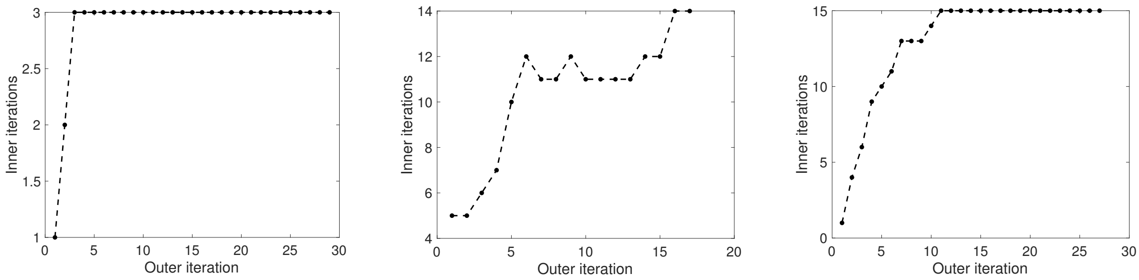
Figure 8.8: Reconstructions of the impulse image from blurred noisy observation. The reconstructed image is of size 128 × 128, and the hyperparameter values are, from left to right: $(r,\eta) = (1, 10^{-5})$, $(r,\eta) = (1/2, 10^{-5})$, and $(r,\beta) = (-1,3)$. The images are in the same scale.

third case, where $r = -1$ and η does not have the same role as for positive values of r, we set $\beta = 3$. We scale the hyperparameters by a constant value, setting $\vartheta_j = \vartheta_0 =$ constant, and to make the results comparable, we select the parameter ϑ_0 so that the lower bound for the scaling parameters θ_j are equal,

$$\vartheta_0\varphi(0) = \vartheta_0 \left(\frac{\eta}{r}\right)^{1/r} = 10^{-9}.$$

In this example, we consider only the approximate IAS algorithm.

The final reconstructions, shown in Figure 8.8 are almost identical, and the number of iterations are comparable. The number of the CGLS inner iterations per outer iteration in each case is low, no more than 15. To see a difference in the performance for the three parameter choices, we show how the reconstruction of the hyperparameter θ proceeds. Figure 8.10 shows logarithmic plots of the vector θ, rendered as a pixel image, after 2,4,8, and 16 iterations. The first observation is that even if the lower bound for the parameter vector θ was set equal, each choice of hyperprior has its charcteristic scale: The value $r = 0.5$ yields the lowest values, while in the case $r = -1$ the interval is shifted to considerably higher values. Interestingly, however, the ratio between the largest and smallest value is in the same range. This observation is important, as the ratio informs us about the relative weights of each column of K in the scaling $\mathsf{K} \to \mathsf{K}\mathsf{D}_\theta^{1/2}$. The column scaling performs an effective model reduction, identifying the relevant columns of K and suppressing irrelevant ones. Each hyperparameter selection in the end identifies the same relevant columns, however the plots in Figure 8.10 show that the choice $r = 1$ is the most conservative, while when $r = -1$ the suppression of irrelevant columns happens sooner. Therefore one can argue that the parameter choices that correspond to less convex case pursue more greedily the support, however, the lack of convexity also makes it possible that the support corresponds to a local, rather than global minimum.



(a) (b) (c)

Figure 8.9: The number of CGLS iterations in each outer iteration. The hyperparameter values correspond to those in Figure 8.8.

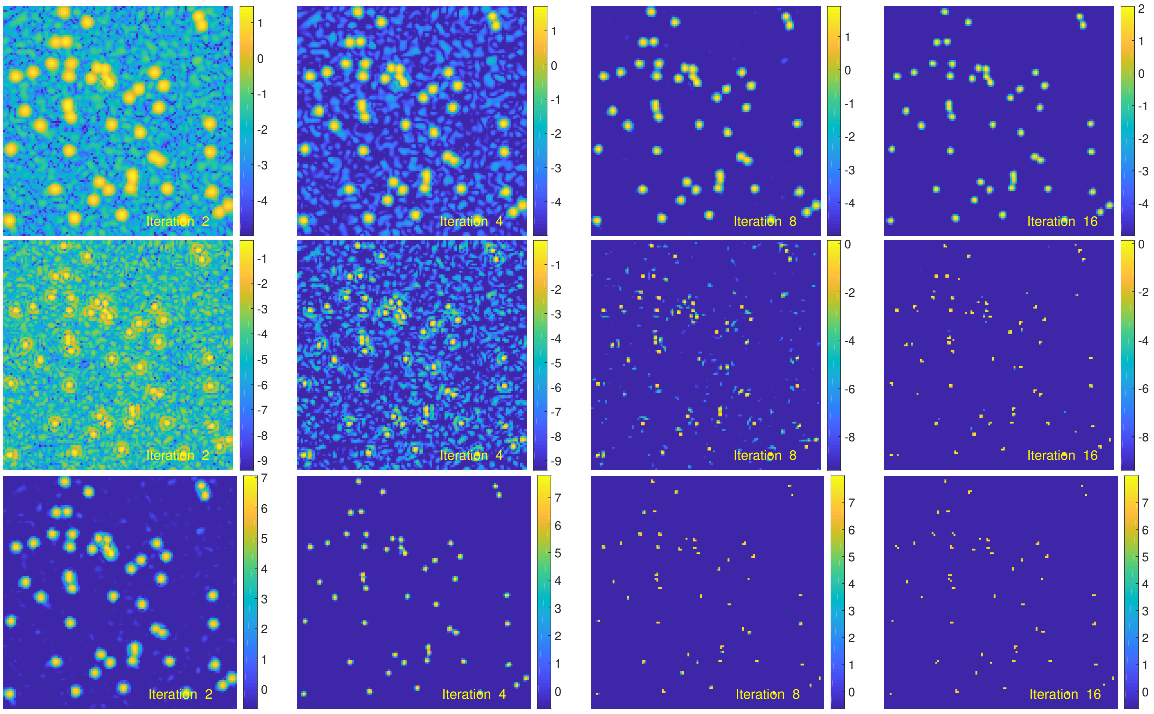


Figure 8.10: Logarithmic plot, top to bottom, of the estimate of θ for the three hyperpriors corresponding to Figure 8.8 at the end of the outer iteration 2, 4, 8 and 16 (left to right).

Chapter 9

Hybrid solvers for hierarchical Bayesian inverse problems

In Chapter 8, the analysis proposed in [32] on the convexity and sparsity promotion of the Gibbs energy functional with gamma hypeprior, has been extended to the case of generalized gamma hypepriors. In particular, we observed that the global convexity is in general lost when using a generalized gamma hyperprior with $r < 1$, leaving open the possibility of stopping at a local minimizer of the energy functional. Nevertheless, Theorem 2 in Section 8.4 states that it is always possible to detect a region in which the Gibbs energy functional is convex.

With the purpose of getting a better trade-off between the sparsity promotion of the model and the robustness of the optimization algorithm, several hybrid approaches have been investigated in literature, some of them relying on ℓ_1-ℓ_2 [165, 115] or ℓ_2-ℓ_0 [63, 66] minimization problems and mixed norms [105, 106]. In a Bayesian framework, the sum of ℓ_p-norms with different p, e.g. $p = 0, 1, 2$, has been set as a regularizer in order to exploit the sparsity promotion and convexity properties of the underlying norms [40]. There, the authors also consider a hierarchical framework for estimating the unknown parameters in the variational model, whose number is limited to 3, namely the three global regularization parameters weighting the contribution of each norm in the regularizer.

Here, in this perspective, we show how the hyperpriors analyzed in Chapter 8 can be mixed in order to couple the strong sparsity promotion of non-convex functionals with the convergence guarantees of convex functionals.

We start recalling the expression of the energy functional corresponding to the adoption of a conditionally Gaussian prior for x with a generalized gamma hyperprior:

$$\begin{aligned}\mathcal{F}(x,\theta \mid r,\beta,\vartheta) &= \frac{1}{2}\|\mathrm{K}x - b\|^2 + \frac{1}{2}\|\mathrm{D}_\theta^{-1/2}x\|^2 - \left(r\beta - \frac{3}{2}\right)\sum_{j=1}^{n}\log\frac{\theta_j}{\vartheta_j} + \sum_{j=1}^{n}\left(\frac{\theta_j}{\vartheta_j}\right)^r \\ &= \frac{1}{2}\|\mathrm{K}x - b\|^2 + \mathcal{P}(x,\theta;r,\beta,\vartheta)\,. \qquad (9.1)\end{aligned}$$

For the sake of completeness, we also quickly recall the statement of Theorem 2, Section 8.4, that has been proven for the non-dimensional energy functional $\Phi(z,\xi)$ and can be easily extended to $\mathcal{F}(x,\theta)$:

(a) If $r \geq 1$ and $\eta = r\beta - 3/2 > 0$, the functional $\mathcal{F}(x,\theta)$ is convex everywhere.

(b) If $0 < r < 1$ and $\eta = r\beta - 3/2 > 0$, or, if $r < 0$ and $\beta > 0$, the function $\mathcal{F}(x, \theta)$ is convex provided that

$$\theta_j < \bar{\theta} = \vartheta_j \left(\frac{\eta}{r|r-1|} \right)^{1/r} , \; j = 1, \dots, n \, .$$

Another remark which is useful to state in advance concerns the way in which the θ-update in Algorithm 7 is formulated. In fact, in non-dimensional settings, Lemma 2 draws a connection between the update of ξ_j and the solution of an ordinary differential equation, by stating that,

$$\varphi(z_j) = \xi_j \text{ with } \varphi \text{ solving (8.13)} \, . \tag{9.2}$$

The relation in (9.2) can be reformulated in terms of x_j and θ_j as

$$\theta_j = g(|x_j| \mid r, \beta, \vartheta_j) \, , \tag{9.3}$$

with $g : \mathbb{R}_+ \to \mathbb{R}_+$ being the counterpart of function φ in the physical space whose dependence on the hyperparameters has been highlighted.
Finally, from now on we will refer to the approximate IAS algorithm introduced in Chapter 8 as IAS.

9.1 Hybrid IAS algorithms

From the point of view of optimization, the benefits of choosing $r \geq 1$ guaranteeing the global convexity are obvious. This, however, should be contrasted with the convergence rate and how strongly each of the hypermodels promote sparsity. As a rule, moving away from global convexity tends to increase the greediness of the algorithm in the sense that it promotes more strongly the sparsity. In this section, we propose two different modifications to the IAS algorithm that aim at taking advantage of the global convexity of the objective function corresponding to the gamma hyperprior, $r = 1$, thus guaranteeing convergence, and the fast convergence of the hyperpriors with $r < 1$ with only a locally convex objective function. In both proposed algorithms, the gamma hyperprior model is used to initial convergence towards the unique global minimum, after which a switch to locally convergent hypermodel is done. We refer to the two proposed algorithms as *local* and *global* hybrid models: in the former one the hypermodel is switched only for components entering in the convexity region of the model $r < 1$ with strong sparsity promotion properties, while in the global model, the full objective function is switched. Both algorithms are extensively tested with computed examples.

9.1.1 Local hybrid IAS

To define the local hybrid algorithm, consider the objective function $\mathcal{F}(x \, , \theta \mid r, \vartheta, \beta)$ with given model parameters (r, β, ϑ) that the IAS algorithm seeks to minimize. Starting from (9.1), we write the objective function as

$$\mathcal{F}(x, \theta \mid r, \vartheta, \beta) = \frac{1}{2} \|\mathrm{K}x - b\|_2^2 + \sum_{j=1}^{n} \mathcal{P}_j(x_j, \theta_j \mid r, \vartheta_j, \beta),$$

where

$$\mathcal{P}_j(x_j, \theta_j \mid r, \vartheta_j, \beta) = \frac{1}{2}\frac{x_j^2}{\theta_j} - \left(r\beta - \frac{3}{2}\right)\log\frac{\theta_j}{\vartheta_j} + \left(\frac{\theta_j}{\vartheta_j}\right)^r.$$

In the IAS algorithm, we choose the parameters r, β and ϑ and keep them fixed during the full iteration process. In the proposed hybrid algorithm, the model parameters are updated dynamically for each component, the selection criterion being based on whether the component pair (x_j, θ_j) satisfies the convexity criterion given in Theorem 2, Section 8.4.

More precisely, consider two hypermodels, with parameters $(r^{(1)}, \beta^{(1)}, \vartheta^{(1)})$ and $(r^{(2)}, \beta^{(2)}, \vartheta^{(2)})$, where $r^{(2)} < 1 \leq r^{(1)}$, $r^{(2)} \neq 0$. We refer to these models as $\mathscr{M}_1$ and $\mathscr{M}_2$, respectively. We start the IAS algorithm by using the model $\mathscr{M}_1$.

After t alternating minimization steps, let $(x, \theta) = (x^{(t)}, \theta^{(t)})$ denote the current approximation of the IAS algorithm. For each coordinate x_j, we now compute the θ_j update - see (9.3) - corresponding to model $\mathscr{M}_2$,

$$\theta_j^{(2)} = g(|x_j| \mid r^{(2)}, \beta^{(j)}, \vartheta_j^{(2)}) = g^{(2)}(|x_j|).$$

If the value satisfies

$$\theta_j^{(2)} < \overline{\theta}_j = \vartheta_j^{(2)}\left(\frac{\eta^{(2)}}{r^{(2)}|r^{(2)} - 1|}\right)^{1/r^{(2)}},$$

we continue the updating of θ_j using the model $\mathscr{M}_2$. Observe that since the function $g^{(2)}$ is strictly increasing, we may write the above condition in terms of x_j,

$$|x_j| < \left[g^{(2)}\right]^{-1}(\overline{\theta}_j) = \overline{x}_j.$$

Let $I \subset \{1, 2, \ldots, n\}$ denote an index set such that

$$j \in I \text{ if and only if } x_j < \overline{x}_j,$$

and denote by I^c its complement. We define a hybrid objective function,

$$\begin{aligned}\mathcal{F}(x, \theta \mid I) \quad = \quad & \frac{1}{2}\|\mathrm{K}x - b\|_2^2 + \sum_{j \in I^c} \mathcal{P}_j(x_j, \theta_j \mid r^{(1)}, \vartheta_j^{(1)}, \beta^{(1)}) + \\ & \sum_{j \in I} \mathcal{P}_j(x_j, \theta_j \mid r^{(2)}, \vartheta_j^{(2)}, \beta^{(2)}).\end{aligned}$$

To guarantee the convexity of the objective function above, we impose a bound constraint

$$|x_j| < \overline{x}_j \text{ for } j \in I, \tag{9.4}$$

as detailed in Section 8.6.

Before discussing further the algorithm, consider the selection of the model parameters, and in particular ϑ. For model $\mathscr{M}_1$, the vector $\vartheta^{(1)}$ can be chosen based on the sensitivity analysis as suggested in Section 8.2. We choose the hyperparameter vector $\vartheta^{(2)}$ using the following design criterion: *If $x_j = 0$, the updating value for θ_j given by the IAS algorithm is the same regardless of the choice of the hypermodel.* We recall that the update of θ_j in the IAS algorithm is given by (8.12). In particular,

$$g(0 \mid r, \beta, \vartheta_j) = \vartheta_j\left(\frac{\eta}{r}\right)^{1/r}, \quad \eta = r\beta - 3/2,$$

which leads to the scaling

$$\vartheta_j^{(2)} = \left(\frac{\eta^{(1)}}{r^{(1)}}\right)^{1/r^{(1)}} \left(\frac{r^{(2)}}{\eta^{(2)}}\right)^{1/r^{(2)}} \vartheta_j^{(1)}.$$

We summarize the proposed local hybrid IAS in an Algorithm 8.

Algorithm 8: Local hybrid IAS

```
input:  observed signal $b \in \mathbb{R}^m$
output:  restored signal $x^*$
 1.  initialize:  set $t = 0$, $\theta^{(0)} = \vartheta^{(1)}$, $I = \emptyset$
 2.  for  $t = 1, 2, \ldots$ until convergence  do:
 3.      update $x^{(t)}$ by solving (8.5) in terms of $w$ and rescaling, $x^{(t)} = \mathrm{D}_\theta^{1/2} w$
 4.      project components $x_j^{(t)}$, $j \in I$, to $[-\overline{x}, \overline{x}]$
 5.      for  $j = 1, \ldots, n$
 6.          if  $\theta_j \geq \overline{\theta}$
 7.              update $\theta_j^{(t)} = g(|x_j^{(t)}| \mid r^{(1)}, \beta^{(1)}, \vartheta_j^{(1)})$
 8.          else
 9.              update $\theta_j^{(t)} = g(|x_j^{(t)}| \mid r^{(2)}, \beta^{(2)}, \vartheta_j^{(2)})$
10.              update $I = I \cup \{j\}$
11.          endif
12.  end for
13.  $x^* = x^{(t)}$, $\theta^* = \theta^{(t)}$
```

Before discussing a modification of the above algorithm, a comment on the projection on convexity interval (step 4) is of order. The projection step is included in the algorithm to ensure that the index set I of components being updated using the hypermodel $\mathscr{M}_2$ is monotonically increasing, which, in general, is not automatically guaranteed. In other words, once that $(x_j^{(t)}, \theta_j^{(t)})$ enters the convexity basin, it can not be ensured that it remains therein at the next IAS iteration, unless the hybrid objective function $\mathcal{F}(x, \theta \mid I)$ is stably convex. The definition of stable convexity has been given in Section 8.5, together with Lemma 5 that suggests a few conditions under which stable convexity is achieved. However, even if in the proposed numerical examples the conditions in Lemma 5 will not be checked, results show that the projection step is in practice not necessary, and the bound constraint $|x_j| < \overline{x}$ is not active.

9.1.2 Global hybrid IAS

We remark that, in order for the convexity of the energy functional to be guaranteed, the hybrid scheme summarized in Algorithm 8 may result to be a very cautious choice: it is expected to give a substantial contribution in terms of background sparsification, but it may fail in enhancing sudden discontinuities in the signal. Hence, one could take the risk of minimizing a non-convex objective function, rather focusing on finding an initial guess sufficiently close to the global minimum.

In the proposed Algorithm 9, we run the IAS algorithm first with model $\mathscr{M}_1$ corresponding to a conservative parameter choice with guaranteed convergence towards a global minimum, and after a fixed number $\bar{t}$ of iterations, we switch to hypermodel $\mathscr{M}_2$ lacking the global convexity, but with strong sparsity promotion. We refer to this scheme as *global* hybrid IAS, since the change of distribution involves all the variances θ_j unlike in the local version where only selected components followed the model $\mathscr{M}_2$.

Algorithm 9: Global hybrid IAS

```
input: observed signal $b \in \mathbb{R}^m$
output: restored signal $x^*$
 1. initialize: set $\theta^0 = \vartheta^{(1)}$
 2. for  $t = 1, 2, \ldots$ until convergence  do:
 3.     update $x^{(t)}$ by solving (8.5) in terms of $w$ and rescaling, $x^{(t)} = \mathrm{D}_\theta^{1/2} w$
 4.     for $j = 1, \ldots, n$
 5.         if $t < \bar{t}$
 6.             update $\theta_j^{(t)} = g(|x_j^{(t)}| \mid r^{(1)}, \beta^{(1)}, \vartheta_j^{(1)})$
 7.         else
 8.             update $\theta_j^{(t)} = g(|x_j^{(t)}| \mid r^{(2)}, \beta^{(2)}, \vartheta_j^{(2)})$
 9.         endif
10. end for
11. $x^* = x^{(t)}$, $\theta^* = \theta^{(t)}$
```

In the description of the algorithm above, the selection of the switch value $\bar{t}$ is defined as an input. A natural modification is to run the model $\mathscr{M}_1$ as long as the variances θ keep changing significantly. Since in general, we have little information of the nature of the minima of the objective function when $r < 1$, a definitive automatic switching rule is not easy to justify. In the following section, we consider the performance of the algorithm in the light of computed examples.

9.2 Computed examples

In this section, we evaluate the performance of the proposed local and global hybrid IAS. In particular, we will turn off the bound constraint in (9.4) and monitor the behavior of the variances along the iterations of Algorithm 8, tracking the ones entering the convexity region to check whether they remain therein or not. In addition, in order to assess the robustness of the global hybrid IAS, we will record the variances in the convexity region at the switching iteration

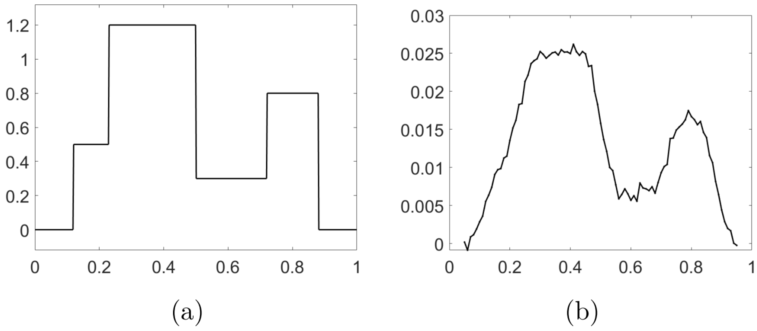


Figure 9.1: The generative model (a), the blurred and noisy data vector $b \in \mathbb{R}^{91}$ (b).

$\bar{t}$ and at the final iteration of Algorithm 9. This analysis is aimed at detecting the areas of the signal where the non-convexity may lead to undesirable effects.

In the following examples, the generative distributions for the hybrid schemes are the gamma ($r^{(1)} = 1$) and the inverse gamma hyperpriors ($r^{(2)} = -1$). The performance of local and global hybrid IAS is thus compared with the one of the plain gamma and plain inverse gamma hyperprior. In the global case, the iteration number at which we switch to the second non-convex hyperprior is $\bar{t} = 10$.

Example 1 As a first example, we consider a one-dimensional deconvolution problem. The generative model as well as the observed data, corrupted by an Airy convolution kernel, have been generated according to the procedure detailed in Example 1 of Section 8.7. Here, we use a dense discretization with $n = n_{\text{dense}} = 1253$ points, while the forward model used for solving the inverse problem assumes $n = 500$. The number of observation points is $m = 91$, and the noise added is assumed to be scaled white noise, with standard deviation σ set to 2% of the maximum of the noiseless generated signal. Figure 9.1 shows the generative signal and the data.

The generative signal is not sparse, but it admits a sparse representation in terms of its increments $z_j = x_j - x_{j-1}$ over the interval of definition

$$z = \mathrm{B}\, x\,,$$

with B defined as in (8.18). Therefore our inverse problem is to estimate the vector z, assumed to be sparse, from the data vector b, given the forward model

$$b = \mathrm{KB}z + e, \quad e \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_m).$$

The restoration results, together with the final variance vector θ and the CGLS steps per IAS iteration, are shown in Figure 9.2. One can notice that the first increment is not easy to detect due to the level of corruption in the signal. In fact, it is not sharply restored when a plain gamma hyperprior is adopted, while it is not detected at all in the case of a plain inverse gamma hyperprior. More specifically, the non-convexity of the MAP objective function for the inverse gamma hyperprior drives the IAS solution towards a local minimum; in fact, the first increment is sharply detected but at the wrong position. In the third row of Figure 9.2, the results obtained via the hybrid local hyperprior are shown. The components of θ plotted in blue follow the inverse gamma distribution at the last iteration of IAS, while the red ones distribute according to a gamma distribution. The action of the inverse gamma helps in sparsifying the background. Finally, the global hybrid hyperprior returns a sharp restoration and detects the five jumps at the original positions. It is also worth remarking that, in all the cases performed, the number of CGLS steps per each IAS iteration equals the number of increments detected. This reflects again the capability of IAS of quickly determining the cardinality of the support.

We now want to give a closer look to the behavior of θ for the hybrid IAS. In Figure 9.3a we show the variances following a gamma distribution (yellow) and the variances switching distribution (green). One can notice that when θ_j enters the convexity region, it remains therein. In other words, once that a distribution switch is performed, it is never reversed. This means that, the activation of the bound constraint (9.4) would have not given any contribution. Then, as far as the global strategy is concerned, we examine the θ vector at iteration $\bar{t}-1$, that is the last iteration of the global hybrid IAS in which a gamma hyperprior is adopted - see Figure 9.3b. The entries of θ below the convexity bound, plotted in blue in Figure 9.3b, are the ones such that the switch to the inverse gamma distribution can be considered *safe*. Then, the same analysis is proposed for all the iterations of global hybrid IAS, as shown in the right panel of Figure 9.3c, whence one can notice that after the switch the decrease of variances proceeds at a higher rate.

Example 2 In the second example, we consider an image restoration problem. We consider a Gaussian convolution kernel as in (8.19) of width $w = 0.015$. The image domain is discretized using a $n \times n$ grid, $n = 136$, whereas the number of observation points is $m = 68 \times 68$. The noiseless signal is corrupted by scaled white noise with standard deviation approximately 2% of the maximum of the noiseless signal. Also in this case, the original image is not sparse in the standard coordinate basis, but it is sparse in terms of its vertical and horizontal increments:

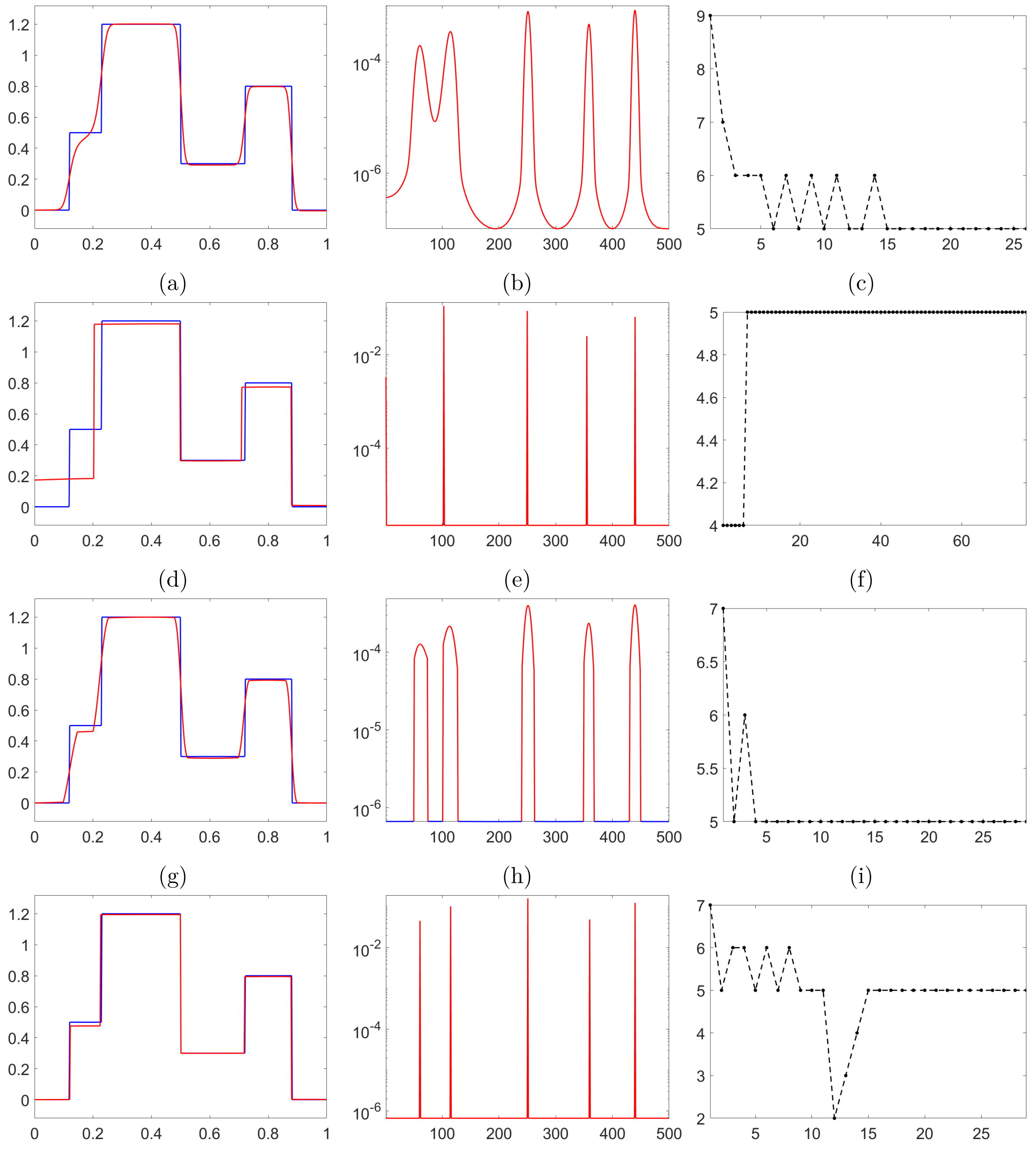
$$z = \mathrm{L}x\,, \quad \mathrm{L} = \begin{bmatrix} \mathrm{D}_v \\ \mathrm{D}_h \end{bmatrix} \in \mathbb{R}^{2n^2 \times n^2}$$

with $\mathrm{D}_v, \mathrm{D}_h$ given in (3.6).

The original image, the observed data and the vector of increments z computed on the original image are shown in Figure 9.4. Note that, due to the binary nature of the image, all the algorithms are performed by constraining x_j in $[0, 1]$, $1 \leq j \leq n^2$, , while the constraint (9.4) is still not active in the local hybrid IAS.

The restored images for the plain IAS with Gamma and Inverse Gamma hyperprior and for the Local and Global IAS are shown in the first column of Figure 9.5. As a further analysis, in Figure 9.6 we also report the logarithmic plot of variances θ_j and the profile of the restorations along the black dashed cut, compared to the corresponding profile of the original image, reported in red. As expected, the restoration via Gamma hyperprior presents rounded corners, while the adoption of an Inverse Gamma hyperprior returns artifacts along the edges. Both these effects are mitigated in the restorations obtained by means of a hybrid approach, especially in global settings.

In Figure 9.7, the number of CGLS steps per outer IAS iteration for the four models is shown.



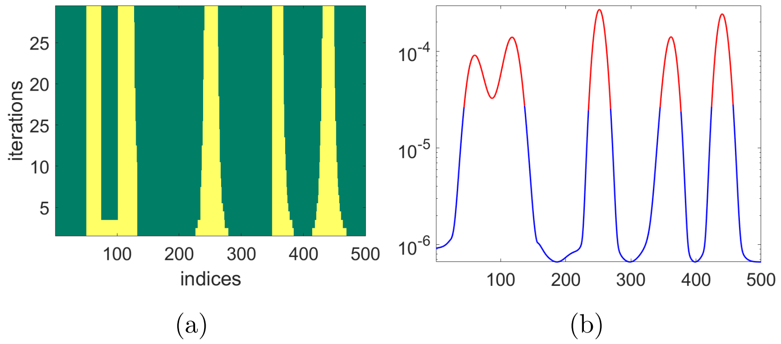
(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

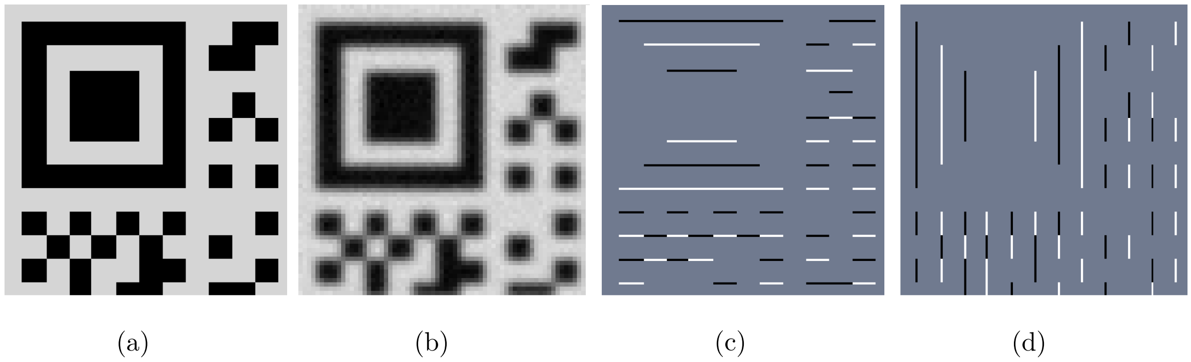
(j) (k) (l)

Figure 9.2: Reconstruction of the signal via gamma, inverse gamma, local hybrid and global hybrid hyperprior (left), the hyperparameter θ (center) and the CGLS iterations per each IAS iteration (right). For the gamma hyperprior in the top row the parameter values are $\eta = 10^{-2}$ and $\vartheta = 10^{-5}$, for the inverse gamma hypeprior in the second row $\eta = -4.5$ and $\vartheta = 10^{-5}$. The hybrid hyperpriors in the bottom rows inherit the parameters from the generative hyperpriors.



(a) (b) (c)

Figure 9.3: Indices of variances switching from a gamma to an inverse gamma hyperprior at each local hybrid IAS iteration (a), variances in the convexity region plotted in blue at iteration $\bar{t} - 1$ (b) and along all the iterations of global hybrid IAS (c). Figure (a) demonstrates that the index set I is monotonically increasing, indicating stable convexity without the need to force the bound constraint (9.4).



(a) (b) (c) (d)

Figure 9.4: Original test image $x \in \mathbb{R}^{136\times136}$ (a), observed data $b \in \mathbb{R}^{68\times68}$ corrupted by Gaussian blur and additive Gaussian noise (b), sparse vector of vertical (c) and horizontal increments (d) computed on the original image.

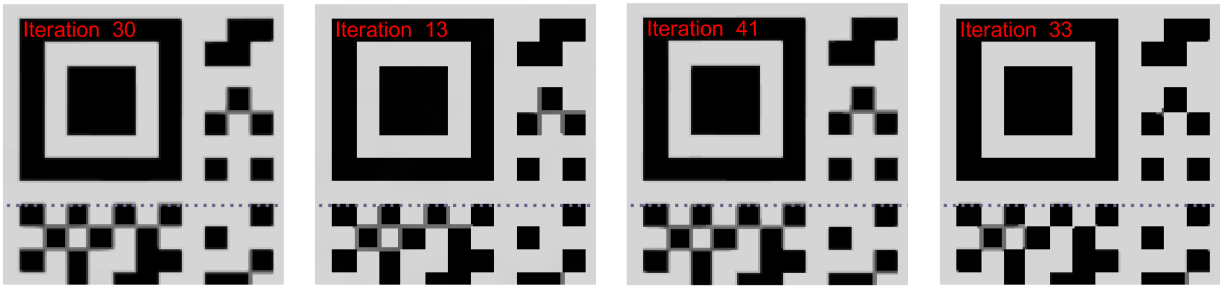


Figure 9.5: From left to right: restored images via gamma, inverse gamma, local hybrid and global hybrid hyperprior. For the gamma hyperprior the parameter values are $\eta = 10^{-4}$ and $\vartheta = 10^{-3}$, for the inverse gamma hypeprior $\eta = -6.5$ and $\vartheta = 10^{-4}$. The hybrid hyperpriors in the bottom rows inherit the parameters from the generative hyperpriors.

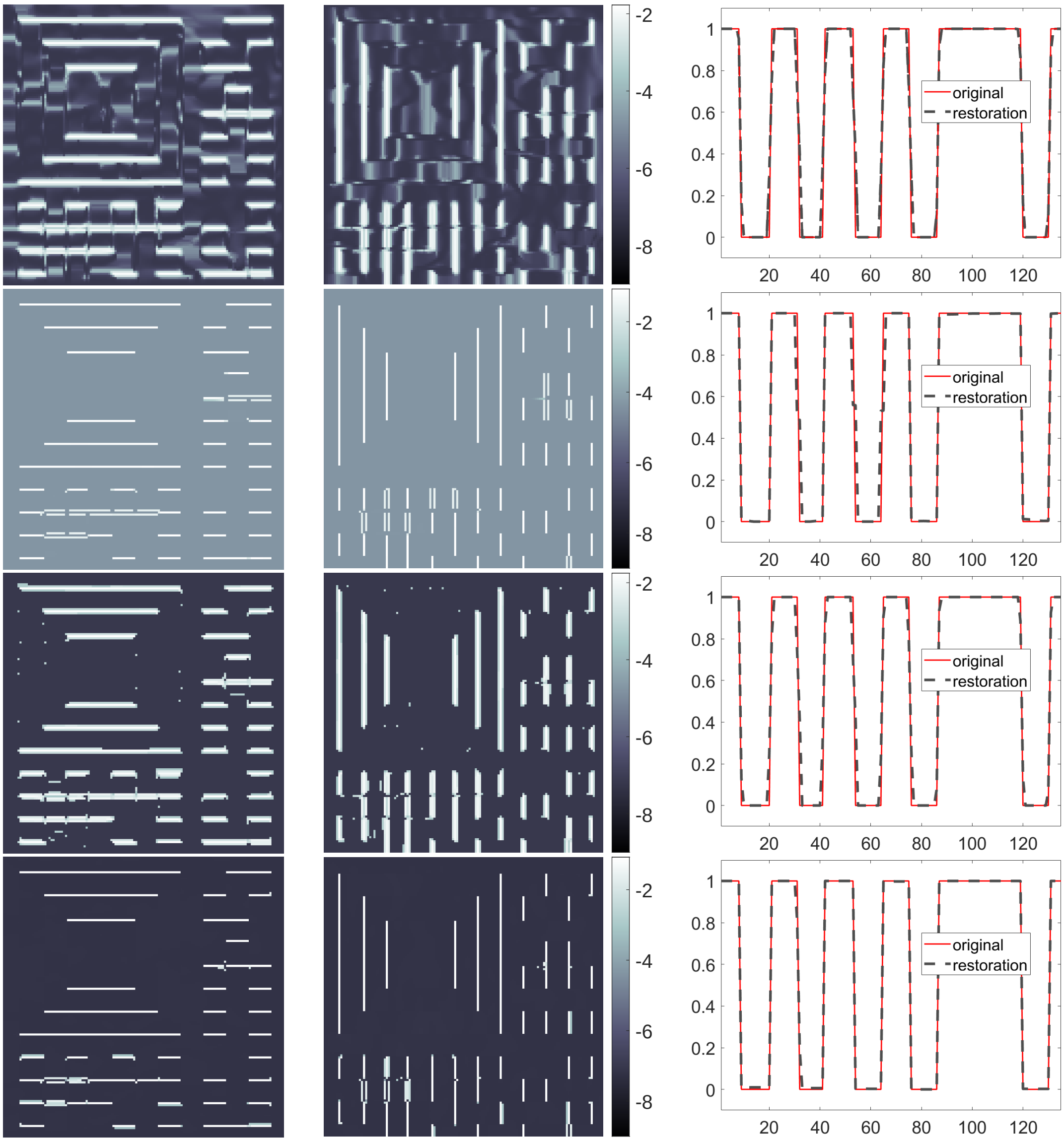


Figure 9.6: From top to bottom: logarithmic plot of variances corresponding to vertical and horizontal increments and one-dimensional profiles extracted from the restorations in Figure 9.5 for the gamma, inverse gamma, local hybrid and global hybrid hyperprior.

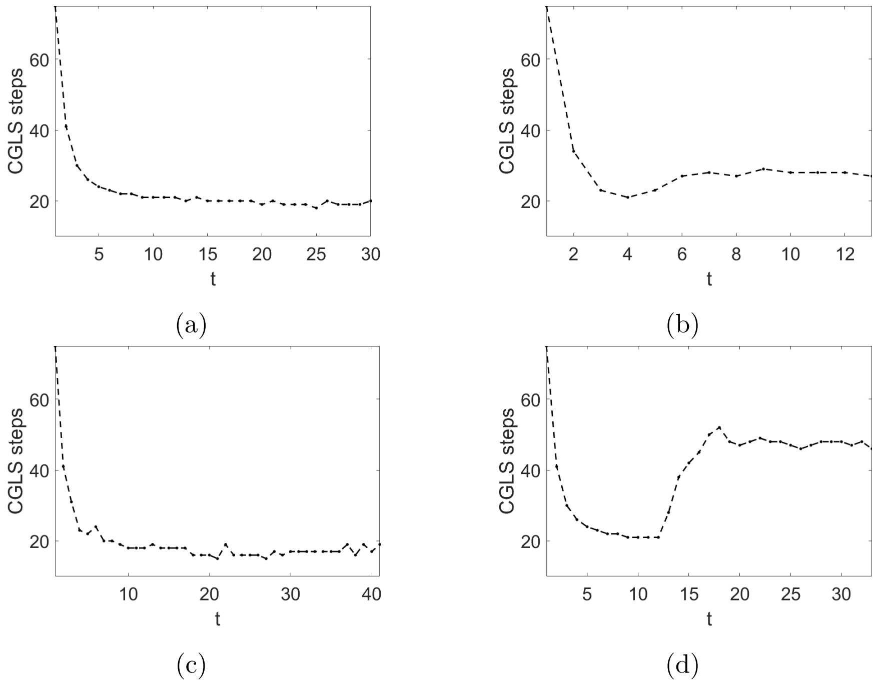


Figure 9.7: Number of CGLS steps per outer iteration for gamma (a), inverse gamma (b), local hybrid (c) and global hybrid hyperprior (d).

Finally, in Figures 9.8a-9.8d we show the indices of the variances of horizontal and vertical increments that at the last iteration of local hybrid IAS follow a gamma (yellow) and an inverse gamma (green) distribution. Figures 9.8b-9.8e and Figures 9.8c-9.8f show the indices behavior at the switching iteration $\bar{t}$ and at the last iteration of global hybrid IAS.

Example 3 In the third example, we consider the problem of estimating again the nearly black two-dimensional object. The generating model is obtained as in Example 2 in Section 8.7. In this example, we assume that the number of observation points is $m = 64 \times 64 = 4\,096$. The noiseless signal is then corrupted by scaled white noise with standard deviation approximately 1.8% of the maximum of the noiseless signal. Figure 9.9 shows the original impulse image characterized by $k = 80$ non-zero points, as well as the noisy image also corrupted by Gaussian blur with kernel width $w = 0.015$. The restored images are shown in Figure 9.10, together with the number of outer IAS iterations. The differences in the four algorithms is highlighted by estimate of the variance θ and by the one-dimensional profiles along the super-imposed black dashed lines. The weak sparsity promotion performed by gamma hyperprior leads to a loss of contrast in the image. On the other hand, the inverse gamma hyperprior returns a sharper restoration which is also more greedy; in fact, the second star of the local profile is not recovered. When considering hybrid hyperpriors, the intensity level increases for the local case when compared with the plain gamma IAS, whereas in global settings the both stars are sharply recovered.

Finally, in Figure 9.11 the behavior of the variances in terms of distribution is shown for the local hybrid case - at the final IAS iteration - and for the global hybrid case - at the switching and final IAS iteration.

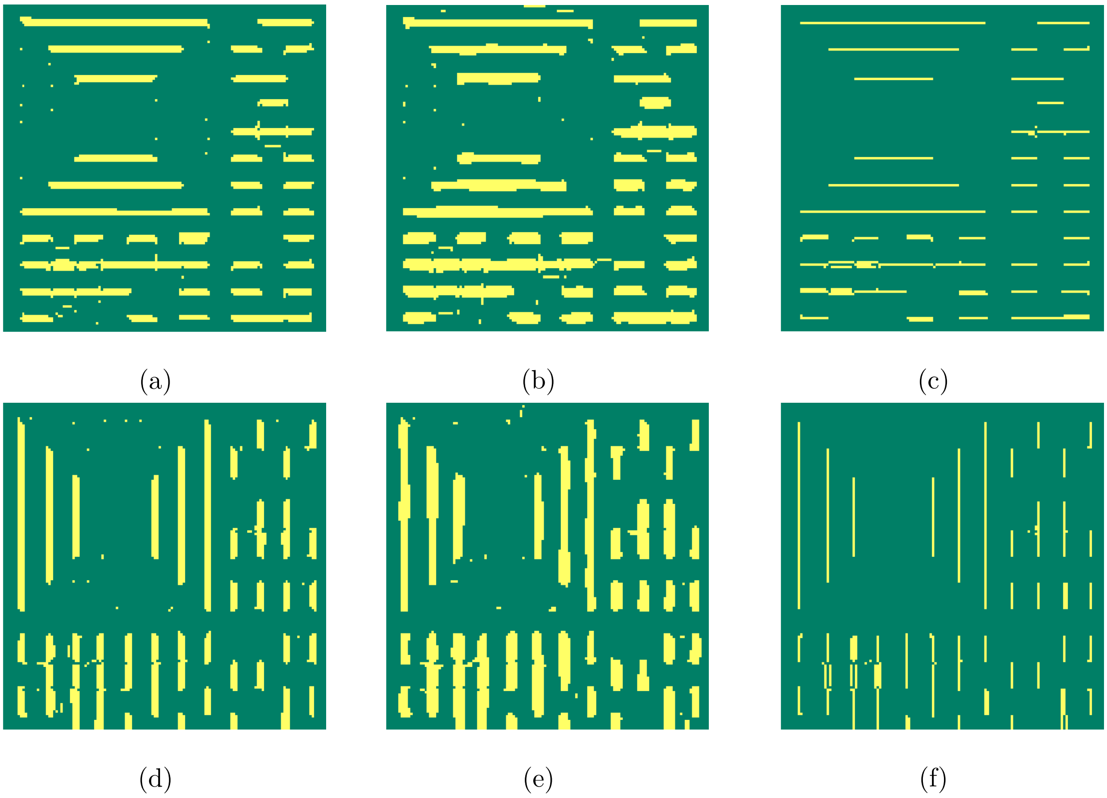


Figure 9.8: Image of the variances θ_j for vertical (top) and horizontal increments (bottom) with color coding indicating if $\theta_j < \bar{\theta}$ (green) or $\theta_j \geq \bar{\theta}$ (yellow). Figures (a)-(d) represent the final iteration of the local hybrid algorithm, (b)-(e) the iteration $\bar{t}-1$, right before the switch of the global hybrid algorithm, and (c)-(f) the final iteration of global hybrid algorithm.

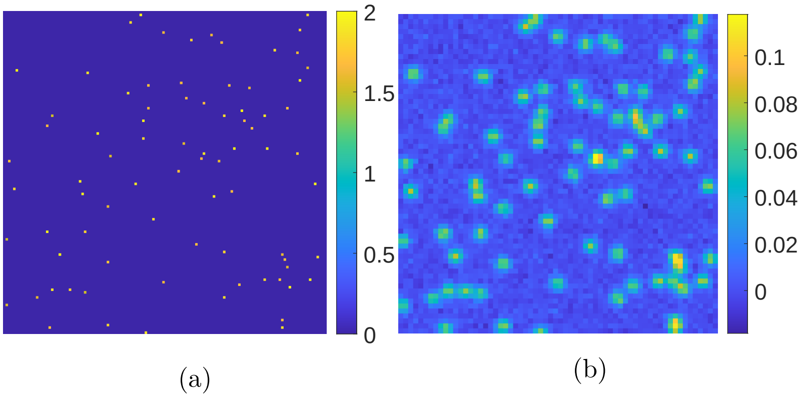


Figure 9.9: Original test image (128 × 128) (a), the 64 × 64 blurred and noisy observation, degraded by Gaussian blur and additive white Gaussian noise (b).

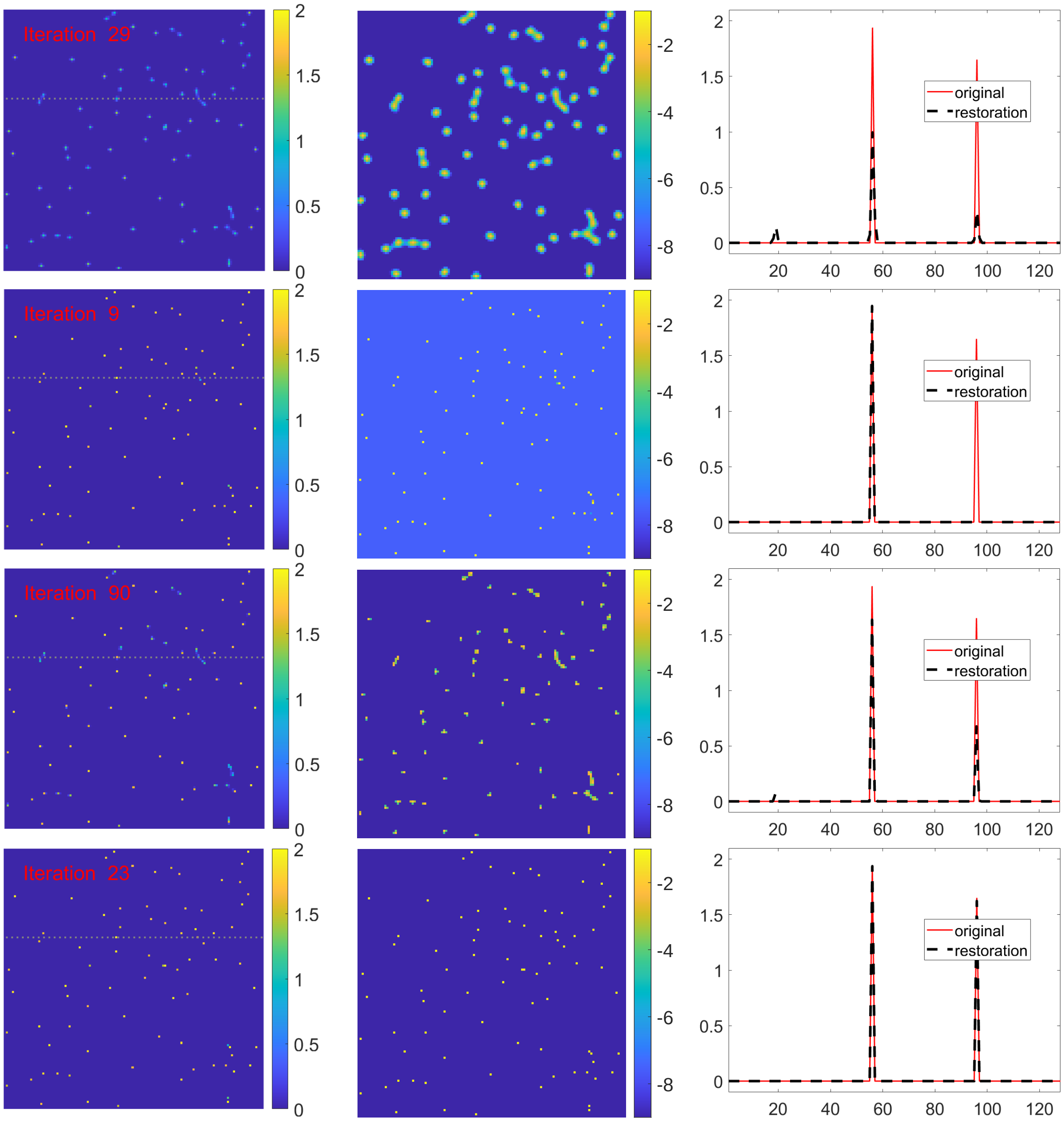


Figure 9.10: Restoration via gamma, inverse gamma, local hybrid and global hybrid hyperprior (first column), corresponding variances (middle column) and horizontal profiles (last column). For the gamma hyperprior in the top row the parameter values are $\eta = 10^{-5}$ and $\vartheta = 10^{-4}$, for the inverse gamma hypeprior in the second row $\eta = -4.5$ and $\vartheta = 10^{-6}$. The hybrid hyperpriors in the bottom rows inherit the parameters from the generative hyperpriors.

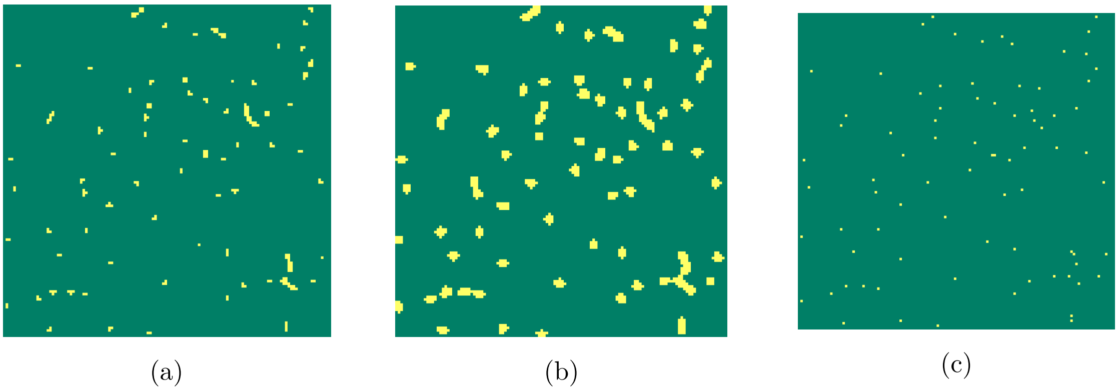


Figure 9.11: Image of the variances θ_j with color coding indicating if $\theta_j < \bar{\theta}$ (green) or $\theta_j \geq \bar{\theta}$ (yellow). Figure (a) represents the final iteration of the local hybrid algorithm, (b) the iteration $\bar{t}-1$, right before the switch of the global hybrid algorithm, and (c) the final iteration of global hybrid algorithm.

Chapter 10

Space-variance and overcomplete basis

In this chapter, the hierarchical Bayesian framework outlined so far will be applied to the recovery of sparse signals expressed in a redundant or over-complete basis. Many classical results in the compressed sensing field hold for signals which are sparse in the standard coordinate basis or sparse with respect to some other orthonormal basis. Nonetheless, in many cases the signal of interest cannot be sparsely represented in an orthonormal basis. The adoption of over-complete basis is also typically preferred since it leads to the reduction of artifacts that may arise during the reconstruction procedure - see [145, 147]. As a consequence, a huge effort has been done in order to extend the consolidated theory on orthonormal basis representation to more general settings [34, 107]. From the algorithmic point view, ℓ_1-based optimization approaches [132, 138, 164] and greedy approaches [130, 77] have been studied. Significant contributions have been also given within a Bayesian framework [10, 9, 141, 142].

A key role in our approach is played by the sensitivity scaling, first proposed in [32] for a gamma hyperprior and then extended to the wider class of generalized gamma distributions. In fact, scaling the columns of an over-complete basis by the sensitivities will prevent from favoring one representation instead of the others. In other words, we expect this procedure to compensate for a possible bias due to the nature of the bases forming a redundant dictionary. We also remark that the sensitivities, that are automatically set as in (8.11), are strictly related to the SNR, the latter encoding information that can be crucial in terms of quality of the final restoration - see discussion in Section 4.2.

10.1 Problem statement

Let $x \in \mathbb{R}^n$ be a signal and $\mathrm{W} \in \mathbb{R}^{n\times N}$, $N \gg n$, an over-complete basis. We denote by $\psi \in \mathbb{R}^N$ a vector representing x in the redundant basis given by the columns of W, the latter being the concatenation of k bases:

$$x = \mathrm{W}\psi = [\mathrm{W}_1 \; \mathrm{W}_1 \ldots \mathrm{W}_k] \begin{bmatrix} \psi_1 \\ \vdots \\ \psi_k \end{bmatrix} ,$$

where

$$\mathrm{W}_j \in \mathbb{R}^{n\times s_j}\,,\ \psi_j \in \mathbb{R}^{s_j} \text{ and } \quad \sum_{j=1}^{k} s_j = N\,.$$

The linear degradation model of reference thus reads

$$b = \mathrm{K}\mathrm{W}\psi + e\,, \quad e \sim \mathcal{N}(0,\Sigma),$$

where $\mathrm{K} \in \mathbb{R}^{m\times n}$, $m \leq n$, is the blur operator and $\Sigma \in \mathbb{R}^{m\times m}$ is the symmetric positive definite covariance matrix of the additive Gaussian noise that we assume to be known.

The numerical experiments will be performed under the adoption of a global hybrid IAS, i.e. the energy functional of reference will be

$$\mathcal{F}(\psi,\theta \mid r^{(1)},\vartheta^{(1)},\beta^{(1)}) = \frac{1}{2}\|\mathrm{K}\mathrm{W}\psi - b\|_2^2 + \sum_{j=1}^{n} \mathcal{P}_j^{(1)}\left(x_j,\theta_j \mid r^{(1)},\vartheta_j^{(1)},\beta^{(1)}\right),$$

with $r^{(1)} \geq 1$ in the first $\bar{t}$ iterations, and it will be changed into

$$\mathcal{F}(\psi,\theta \mid r^{(2)},\vartheta^{(2)},\beta^{(2)}) = \frac{1}{2}\|\mathrm{K}\mathrm{W}\psi - b\|_2^2 + \sum_{j=1}^{n} \mathcal{P}_j^{(2)}\left(x_j,\theta_j \mid r^{(2)},\vartheta_j^{(2)},\beta^{(2)}\right),$$

with $r^{(2)} < 1$, afterwards. Before presenting the experimental results that constitute the main part of the chapter we remark that here our point of view is dual. Namely, we address the problem in a compressed sensing perspective, that is, we are looking for the basis, or the bases, more sparsely representing a signal of which only few measurements are available. In addition, the same framework could be performed on image decomposition problems, also representing a very active field of research [146, 7, 122]. Hence, one could interpret again the problem in a space-variant perspective; in other words, when restoring an image, we look for the basis returning the sparsest representation of each local feature.

10.2 Computed Examples

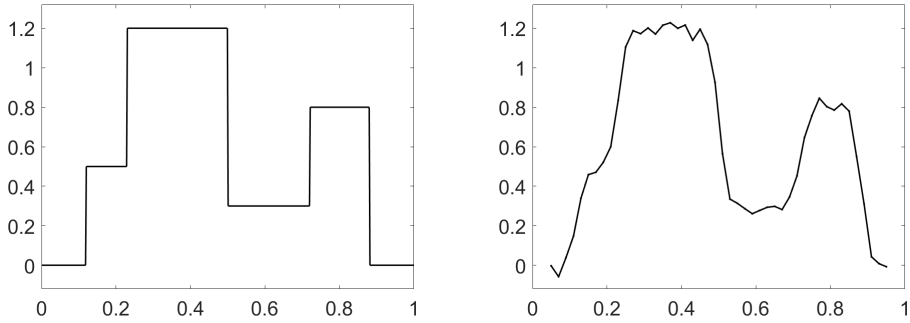
In the following examples the generative hyperpriors for the global hybrid IAS are the gamma hyperprior, $r^{(1)} = 1$, and the generalized gamma hyperprior with $r^{(2)} = 1/2$, while the switching iteration $\bar{t}$ is set equal to 20. We are denoting by,

$$\eta^{(1)} = r^{(1)}\beta^{(1)} - \frac{3}{2}\,, \quad \eta^{(2)} = r^{(2)}\beta^{(2)} - \frac{3}{2}$$

the hyperparameters corresponding to the generative hyperpriors. Moreover, as already remarked, we will make extensive use of sensitivities.

In the first example we will consider a one-dimensional deconvolution problems, where the signal of interest can be naturally represented in one of the two bases forming the over-complete dictionary. In Example 2, two natural bases for a blocky image will be considered and we will test the performance of the proposed algorithm in detecting the basis allowing for the sparsest representation. Finally, in the third example, we will address the task of restoring an image presenting different local features. We will thus look for the sparsest representation of each feature in one of the three bases forming the dictionary.

Furthermore, in the last two examples, the global hybrid IAS is performed with bound constraint according to the original image scale.



(a) Original x.

(b) Corrupted b.

Figure 10.1: The generative model (a) and the blurred and noisy data vector $b \in \mathbb{R}^{46}$ (b).

Example 1 In the first example, we consider the following over-complete representation for signal $x \in \mathbb{R}^n$:

$$x = \mathrm{W}\psi = [\mathrm{W}_1\ \mathrm{W}_2] \begin{bmatrix} \psi_1 \\ \psi_2 \end{bmatrix},$$

where the entries of $\psi_1 \in \mathbb{R}^n$ represent the increments of signal x, i.e., W_1 is the inverse of the increment matrix introduced in (8.18), while $\psi_2 \in \mathbb{R}^n$ represents x in the basis given by the columns of W_2 that here is the transpose of the cosine transform matrix.

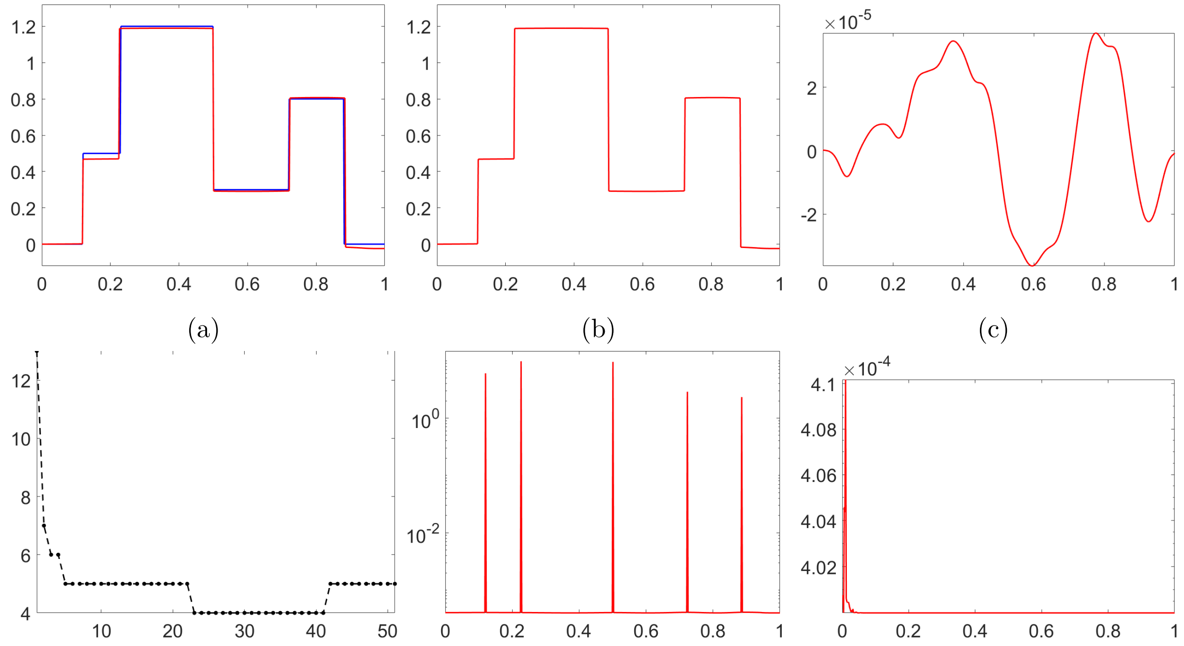
The generative model shown in Figure 10.1a has been obtained via the procedure detailed in Example 1, Section 8.7, using a dense discretization of the interval $[0,1]$ of $n_{dense} = 1253$ points. The signal has been corrupted by Gaussian blur of width $w = 0.02$ and additive scaled white Gaussian noise with standard deviation σ set to 2% of the maximum of the noiseless generated signal. Moreover, the number n of points in $[0,1]$ for the solution of the inverse problem has been set equal to 500 , while the number of observation points m is 46. The corrupted and down-sampled signal b is shown in Figure 10.1b.

The staircase signal x is naturally described by the increment basis. As a consequence, we do not expect ψ_2 to give any significant contribution in the representation of the final restoration. The restoration result obtained via the global hybrid IAS algorithm is shown in Figure 10.2a. The contributions of the two bases are encoded in the vectors $\mathrm{W}_1\psi_1$ and $\mathrm{W}_2\psi_2$, shown in Figure 10.2b-10.2c, respectively. In Figures 10.2e-10.2f, we also report the output variances corresponding to vectors ψ_1 and ψ_2 scaled by the sensitivities, i.e.

$$\xi_j = \frac{\theta_j}{\vartheta_j^{(2)}}.$$

According to the different order of magnitudes, we can conclude that, despite the high level of corruption and down-sampling in the observed data b, the hypermodel has detected the basis that can more sparsely and naturally describe the original signal. Finally, in Figure 10.2d, we show the number of CGLS iterations, whence one can again observe a connection between the dimension of the Krylov subspace generated by CGLS and the effective cardinality of the support of vector ψ.

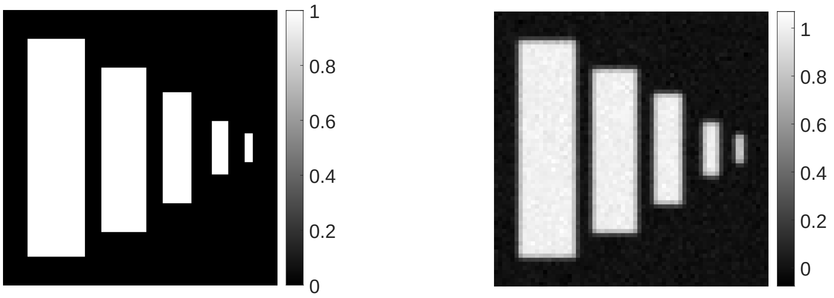
Example 2 In this example, we are testing the outlined framework on the recovery of a blocky image admitting a natural sparse representation both in the basis of the vertical and horizontal



(a) (b) (c)

(d) (e) (f)

Figure 10.2: Restoration via global hybrid IAS (a), contribution of the increment basis (b) and of the cosine transform matrix (c), number of CGLS steps per outer iteration (d), scaled variances corresponding to ψ_1 (e) and ψ_2 (f). The global hybrid CGLS has been performed with parameter $\eta_1 = 10^{-4}$ and $\eta_2 = 10^{-3}$.



(a) Original x.

(b) Corrupted b.

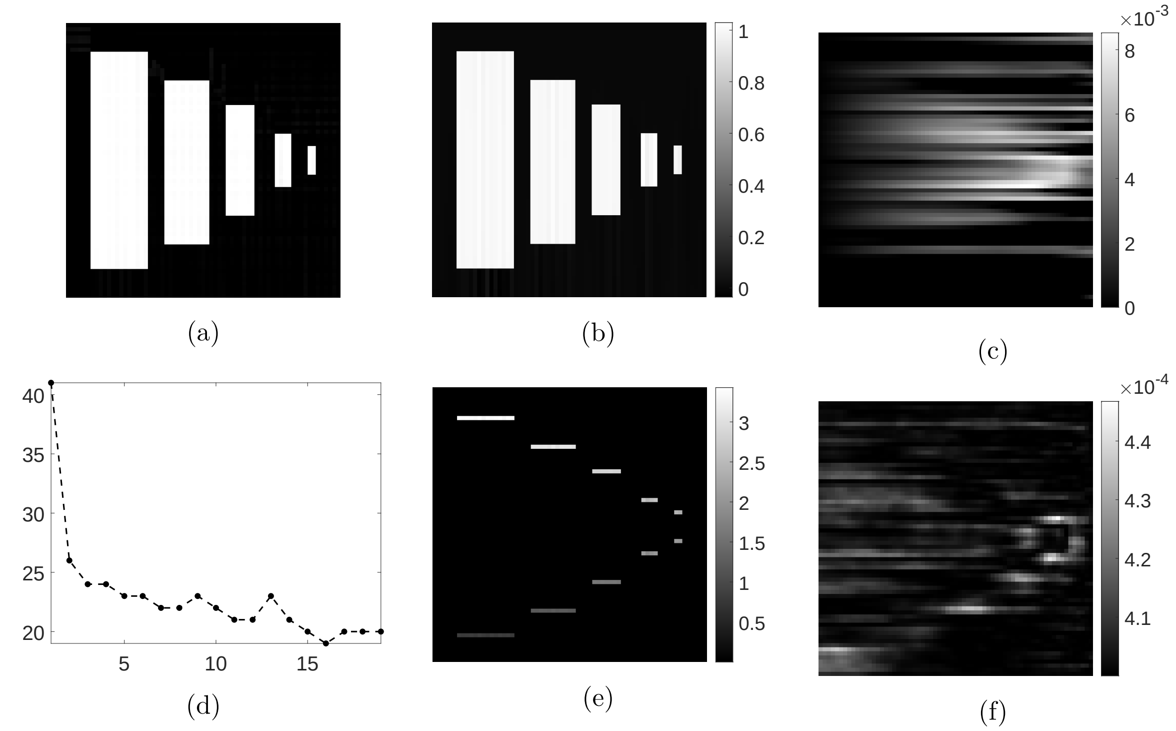
Figure 10.3: The original image (a) and the blurred and noisy image b (b).

increments. In formula,

$$x = [\mathrm{W}_1\mathrm{W}_2]\begin{bmatrix}\psi_1\\ \psi_2\end{bmatrix} \quad \text{with} \quad x \in \mathbb{R}^{n^2}, \psi_1, \psi_2 \in \mathbb{R}^{n^2} \text{ and } \mathrm{W}_1 = \mathrm{D}_v^{-1}, \ \mathrm{W}_2 = \mathrm{D}_h^{-1} \in \mathbb{R}^{n^2 \times n^2},$$

where D_v and D_h are defined as in (3.6). The original blocky image x in Figure 10.3a rearranged as a 67×67 image has been corrupted by Gaussian blur of width $w = 0.01$ and additive scaled white Gaussian noise with standard deviation set as the 2% of the maximum of the noiseless image. The observed data b is shown in Figure 10.3b.

Notice that the representation via vertical increments is more sparse.
The restoration via global hybrid IAS is shown in Figure 10.4a, together with the contribution given by the vertical and horizontal increment bases in Figure 10.4b-10.4c, respectively. As



(a) (b) (c)

(d) (e) (f)

Figure 10.4: *First row*: restored image (a), contribution of $\mathrm{W}_1\psi_1$ (b) and $\mathrm{W}_2\psi_2$ (c) in the final restoration. *Second row*: number of CGLS steps per outer iteration (d), scaled variances corresponding to ψ_1 (e) and ψ_2 (f). The global hybrid CGLS has been performed with parameters $\eta^{(1)} = 10^{-3}$ and $\eta^{(2)} = 10^{-2}$.

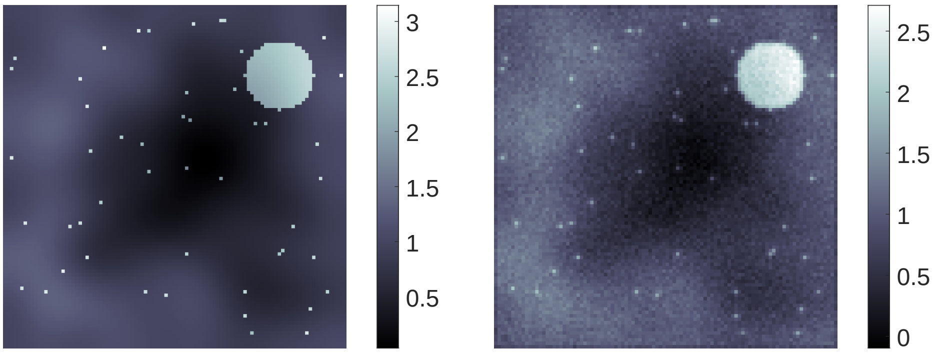
expected, the image is almost completely restored starting from the information encoded in the vertical increment representation. As a further evidence, we also report the output scaled variances corresponding to vectors ψ_1 and ψ_2 - see Figures 10.4e-10.4f - as well as the number of CGLS steps per outer iteration of the hybrid IAS - see Figure 10.4d.

Example 3 In this third example, we are finally addressing the restoration of an image presenting different local features, each of them admitting a sparse representation in one of the three bases forming the over-complete dictionary. More specifically, the original vectorized image x, shown in Figure 10.5a rearranged as 100×100 image, can be expressed as follows:

$$x = [\mathrm{W}_1\ \mathrm{W}_2\ \mathrm{W}_3] \begin{bmatrix} \psi_1 \\ \psi_2 \\ \psi_3 \end{bmatrix}, \quad \text{with} \quad x \in \mathbb{R}^{n^2}, \psi_j \in \mathbb{R}^{n^2},$$

where $\mathrm{W}_1 = \mathrm{I}_n$, W_2 is the transpose of the discrete cosine transform matrix and $\mathrm{W}_3 = \mathrm{D}_h$. Clearly, the smooth part of the image, i.e. the cloud, admits a sparse representation in the cosine transform basis, the blocky moon can be sparsely represented in the increment basis, while the stars are sparse in the standard coordinate basis. The original image has been corrupted by Gaussian blur of width $w = 0.006$ and additive scaled white Gaussian noise with standard deviation set as the 2% of the maximum of the noiseless image - see Figure 10.5b.

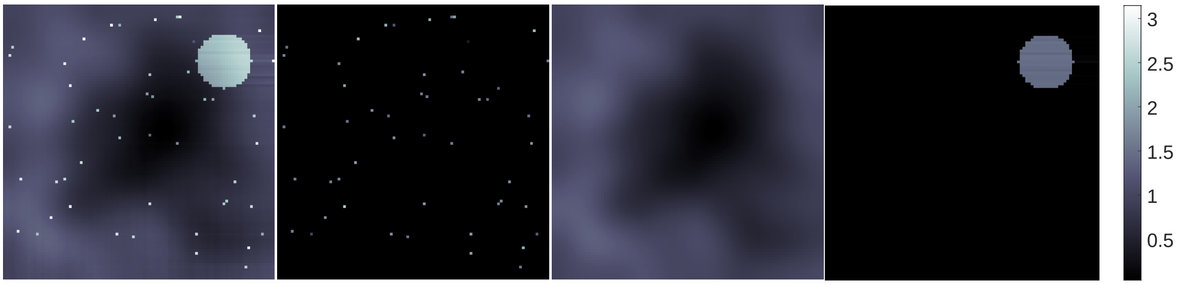
The restoration via global hybrid IAS is shown in Figure 10.6a, while the contributions given by the representation in the sub-bases of the over-complete dictionary are shown in Figures



(a) Original x.

(b) Corrupted b.

Figure 10.5: The original image (a) and the corrupted data (b).



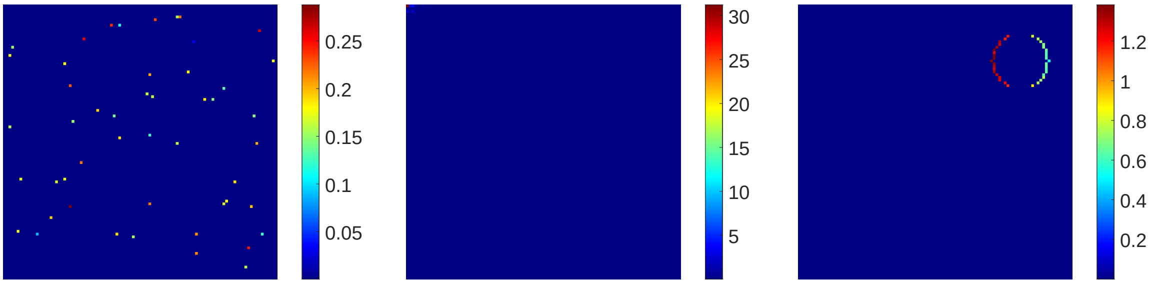
(a)

(b)

(c)

(d)

Figure 10.6: Final restoration (a), contribution of $\mathrm{W}_1\psi_1$ (b), $\mathrm{W}_2\psi_2$ (c) and $\mathrm{W}_3\psi_3$ (d) in the same scale. The global hybrid IAS has been performed with parameters $\eta^{(1)} = 10^{-3}$ and $\eta^{(2)} = 10^{-2}$.



(a)

(b)

(c)

Figure 10.7: Output scaled variances corresponding to ψ_1 (a), ψ_2 (b) and ψ_3 (c).

10.6b-10.6d. The output variances scaled by the sensitivity factors are shown in Figure 10.7.

Conclusions

In this work, we introduced statistically-inspired and high-parametrized regularization terms for the restoration of natural and sparse signals. The flexibility of the proposed regularizers, i.e. of the prior from which they are derived, is coupled with automatic procedures for the estimation of the large number of parameters involved in their definition. Such procedures are based on the introduction of non-informative and informative hyperpriors that have been explored in Part II and Part III of the thesis, respectively.

In Part II, we proposed different space-variant regularizers, by sequentially adding a further degree of freedom in the distribution modeling the local behavior of the gradients in the image. This increasing generalization allowed us to analyze the contribution of each space-variance related to the local strength, type and orientation of the regularization. In order to determine which one influences the most the final restoration, one has to take into account the trade-off arising between benefits and computational costs. In this perspective, the HWTV-L$_2$ model presents two main advantages. Firstly, the convexity model, even if not ensuring the overall convergence of ADMM, prevents from hand-tuning the penalty parameters γ_v, γ_w whose setting is instead crucial for non-convex problems. Secondly, the local scale parameters in the WTV regularizer can be updated via a closed-formula that can thus be performed at each iteration of ADMM. When introducing a second space-variant parameter p in the expression of the $\mathrm{TV}_{p,\alpha}^{\mathrm{sv}}$ regularizer in Chapter 6, a non-convex minimization problem has to be addressed. In addition, updating the parameters at each iteration of the ADMM-based algorithm would slow down the computations, since no-closed formula is available for the p. The same considerations stand for the $\mathrm{DTV}_p^{\mathrm{sv}}$-L$_2$ model in Chapter 7. Hence, we can conclude that the last two models would certainly take advantage from a faster parameter estimation procedure, based, for instance, on the detection of patches in the image sharing the same properties, see e.g. [119], or on a shrinkage of the parameter domain inspired by the properties of the image of interest. Moreover, as a future study, we also plan to extend the proposed models, as well as the estimation procedure, to address other inverse problems in image processing, such as image reconstruction and inpainting, that can certainly benefit from a space-variant and directional approach.

In Part III, we proposed a detailed analysis of the behavior of the energy functional in the MAP estimation problem when different hyperpriors selected from the family of generalized gamma distributions are adopted. Starting from the results discussed in Chapter 8, we introduced two hybrid algorithms that have been designed with the purpose of combining the strong sparsity promotion of non-convex penalty terms with the guarantees of convex settings. Finally, the outlined framework has been applied to sparse recovery problems for signals represented in a redundant dictionary. In the future, we plan to answer to some of the questions that have been left open, concerning the convergence of the hybrid schemes proposed in Chapter 9 and the derivation of less restrictive conditions ensuring the stable convexity of the hybrid energy functional.

Moreover, motivated by the encouraging results shown in Chapter 10, we believe that the proposed model and algorithms are especially suited to deal with real worlds problems, such as medical imaging or surveillance problems, where meaningful information has to be extracted from degraded, down-sampled data, in which artifacts, possibly admitting a sparse representation in a given basis, may arise.

Appendix A

Probability preliminaries

Definition A.1. *Let Ω be a given set and $\mathcal{F}$ a nonempty collection of subsets of Ω. Then, $\mathcal{F}$ is a σ-algebra on Ω if*

(i) $\Omega \in \mathcal{F}$;

(ii) $F \in \mathcal{F}; \Rightarrow \overline{F} = \Omega \backslash F \in \mathcal{F}$

(iii) $\{F_n\}_n \subset \mathcal{F} \Rightarrow \bigcup_{n=1}^{\infty} F_n \in \mathcal{F}$.

The pair $(\Omega, \mathcal{F})$ is a measurable space.

$F \in \Omega$ such that $F \in \mathcal{F}$ is called $\mathcal{F}$-*measurable* or *event.*

We recall the following definition of probability introduced by Kolmogorov:

Definition A.2. *A function,*

$$\mathbb{P} : \mathcal{F} \to [0, 1]$$

is a probability measure on a measurable space $(\Omega, \mathcal{F})$ if,

(i) $\mathbb{P}(F) \geq 0, \quad \forall F \in \mathcal{F}$;

(ii) $\mathbb{P}(\Omega) = 1$;

(iii) denoting by $\{F_n\}_n$ a collection of mutually disjoint sets in $\mathcal{F}$, it holds $\mathbb{P}(\bigcup\limits_{n=1}^{\infty} F_n) = \sum\limits_{n=1}^{\infty} \mathbb{P}(F_n)$.

We refer to the triple $(\Omega, \mathcal{F}, \mathbb{P})$ as a probability space.

For sake of completeness, it is worth remarking that, beside the Kolmogorov definition, there are different ways to introduce the concept of probability. Other definitions proposed in literature are listed below:

1. **Classical probability**: The probability of an event is the number of favorable cases divided by the number of all possible cases (relative frequency of the event), in the hypothesis that each event is equally possible.

2. **Frequentist probability**: The probability of an event is the limit of its relative frequency in a large number of trials.

3. **Subjective probability**: The probability of an event is the level of confidence that an individual has about its occurrence. It depends on the knowledge and on the beliefs of the specific individual.

In particular, in the Bayesian framework, a subjective point of view is adopted.

Definition A.3. *Given $\mathcal{U}$, a family of subsets of Ω, we call σ-algebra generated by $\mathcal{U}$ the smallest σ-algebra containing $\mathcal{U}$, i.e. the intersection of all the sigma-algebras built on $\mathcal{U}$*

$$\sigma(\mathcal{U}) = \bigcap\{\mathcal{F} : \mathcal{F} \textit{ is a } \sigma\textit{-algebra on } \Omega,\, \mathcal{U} \subset \Omega\}\,.$$

If Ω is a topological space, the σ-algebra generated by the family of open sets in Ω is called the Borel σ-algebra on Ω and each $B \in \mathcal{B}$ is a Borel set.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and consider a function

$$Y : \Omega \to \mathbb{R}^n\,.$$

Y is said to be $\mathcal{F}$-*measurable* if

$$Y^{-1}(U) = \{\omega \in \Omega : Y(\omega) \in U\}\,,$$

for all U open sets in $\mathbb{R}^n$.
We can now give the definition of *random variable*, which will play a central role in our discussion.

Definition A.4. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, we say that a function*

$$X : \Omega \to \mathbb{R}^n\,,$$

is a (multivariate) random variable if X is $\mathcal{F}$-measurable. Function X assigns to each element $\omega \in \Omega$ a vector $x = X(\omega) \in \mathbb{R}^n$, with x being a realization of X.

Definition A.5. *Let X be a multivariate random variable and B a Borel set. The function*

$$\mu_X : \mathbb{R}^n \to [0, 1]\,,$$

such that

$$\mu_X(B) = \mathbb{P}(X \in B)\,,$$

is called probability distribution of X. Under the hypothesis of absolutely continuity of μ_X with respect to the Lebesgue measure, we can define the probability density function P_X:

$$\mathbb{P}(X \in B) = \mu_X(B) = \int_B d\mu_X(x) = \int_B \mathrm{P}_X(x)dx\,.$$

The integral of π_X over a generic Borel set B is a probability measure, and

$$\int_{\mathbb{R}^n} \pi_X(x)dx = 1\,.$$

Given two random variables $X, Y : \Omega \to \mathbb{R}^n$, we can define the *joint probability distribution* as

$$\mu_{XY}(A, B) = \mathbb{P}(X \in A, Y \in B)\,.$$

If μ_X is absolutely continuous with respect to the Lebesgue measure over $\mathbb{R}^n \times \mathbb{R}^n$, we have

$$\mu_{XY}(A, B) = \int\int_{A\times B} \mathrm{P}_{XY}(x, y) dxdy\,.$$

From the joint probability density $\pi(x, y)$, we can compute the *marginal density* of X and Y:

$$\mathrm{P}(x) = \int_{\mathbb{R}^n} \mathrm{P}(x, y) dy\,, \qquad \mathrm{P}(y) = \int_{\mathbb{R}^n} \mathrm{P}(x, y) dx\,.$$

Two random variables X and Y are said to be *independent* if their joint probability distribution and joint probability density can be factorized as follows:

$$\mu(x, y) = \mu_X(x)\mu_Y(y)\,, \qquad \mathrm{P}(x, y) = \pi_X(x)\mathrm{P}_Y(y)\,.$$

Given X and Y, we are interested in finding the probability density of X assuming that a realization for $Y = y$ is known. We have to define the *conditional probability density*:

$$\mathrm{P}(x \mid y) = \frac{\mathrm{P}(x, y)}{\mathrm{P}(y)}\,.$$

The roles of X and Y are clearly symmetric so:

$$\mathrm{P}(x \mid y)\pi(y) = \mathrm{P}(y \mid x)\mathrm{P}(x)\,.$$

Consider n random variables $X_1, \dots, X_n$. We can formulate the analogous of the *compound probability law* in terms of probability density:

$$\mathrm{P}(x_1, \dots, x_n) = \mathrm{P}(x_1 \mid x_2, \dots, x_n)\mathrm{P}(x_2 \mid x_3, \dots, x_n) \cdots \mathrm{P}(x_{n-1} \mid x_n)\mathrm{P}(x_n)\,.$$

One of the most useful tool in the following discussion is the *Bayes' formula*, which is an immediate consequence of the conditional density definition and the symmetry identity:

$$\mathrm{P}(x \mid y) = \frac{\mathrm{P}(y \mid x)\mathrm{P}(x)}{\mathrm{P}(y)}, \quad \mathrm{P}(y) \neq 0\,.$$

Let X be a random variable. If

$$\int_\Omega |x| d\mu(x) < \infty\,,$$

we can define the *mean* or *expectation* of X as

$$\mathbb{E}(X) = \overline{x} = \int_\Omega x d\mu(x) = \int_\Omega x\mathrm{P}(x) dx\,.$$

Clearly, $\mathbb{E}$ is a linear operator:

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]\,.$$

The *variance* of $X : \Omega \to \mathbb{R}$ is defined as

$$var(X) = \mathbb{E}[(X - \overline{x})^2] = \int_{\mathbb{R}} (x - \overline{x})^2 \mathrm{P}(x) dx\,.$$

The variance measures how far each value taken on by the random variable is from the mean $\overline{x}$. The *covariance* of $X, Y : \Omega \to \mathbb{R}$ is defined as

$$cov(X,Y) = \mathbb{E}[(X-\overline{x})(Y-\overline{y})] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Consider now a multivariate random variable $X : \Omega \to \mathbb{R}^n$. The *covariance matrix* of X is defined as

$$cov(X) = \int_{\mathbb{R}^n} (x-\overline{x})(x-\overline{x})^T \mathrm{P}(x)dx\,,$$

or, using a component-wise notation

$$cov(X)_{ij} = \int_{\mathbb{R}^n} (x_i-\overline{x}_i)(x_j-\overline{x}_j)\mathrm{P}(x)dx\,.$$

In particular $cov(X)_{ii} = var(X_i)$, $\forall i = 1,..,n$.
The covariance matrix is symmetric and positive semi-definite: this guarantees its invertibility. In the following, we show the semi-definiteness, for any $v \in \mathbb{R}^n$, $v \neq 0$:

$$\begin{aligned} v^T cov(X) v &= v^T \Big(\int_{\mathbb{R}^n} (x-\overline{x})(x-\overline{x})^T \mathrm{P}(x)dx \Big) v \\ &= \int_{\mathbb{R}^n} v^T (x-\overline{x})(x-\overline{x})^T v \mathrm{P}(x)dx \\ &= \int_{\mathbb{R}^n} [v^T(x-\overline{x})]^2 \mathrm{P}(x)dx \geq 0\,, \end{aligned}$$

due to the non-negativity of P.

Appendix B

ADMM

The Alternating Direction Method of Multipliers (ADMM) is a popular algorithm, which is well-established in the field of convex optimization. Moreover, in light of recent convergence results, it has started to be widely used to address also non-convex problems. ADMM was first introduced by Glowinski and Marrocco [78] and by Gabay and Mercier [75]. Nonetheless, it is closely related to many other algorithms, such as the Douglas-Rachford splitting method.

At the core of ADMM there is the idea of splitting the original problems into smaller ones that are easier to handle. The solution of the smaller sub-problems thus coordinate in order to find a solution for the large problem.

Here, we quickly recall the general form of ADMM for convex problems and mention some of the most remarkable results in terms of convergence. For further details, one can refer to the surveys [14, 124] and the references therein.

Let $f, g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be two closed proper convex functions and consider the following minimization problem,

$$\begin{aligned} &\min_{x,z} \{f(x) + g(z)\} \\ &\text{s.t. } \mathrm{A}x + \mathrm{B}z = c\,, \end{aligned} \tag{B.1}$$

where $x \in \mathbb{R}^n$, $z \in \mathbb{R}^m$, $\mathrm{A} \in \mathbb{R}^{p \times n}$, $\mathrm{B} \in \mathbb{R}^{p \times m}$, $c \in \mathbb{R}^p$. Notice that f and g are not required to be differentiable. Furthermore, their being extended-valued functions allows to consider the case in which they are indicator functions, which is a typical configuration is several applications. We thus introduce the Augmented Lagrangian associated with problem (B.1) :

$$\mathcal{L}(x, z, \lambda; \beta) = f(x) + g(z) - \lambda^T (c - \mathrm{A}x - \mathrm{B}z) + \frac{\beta}{2} \|\mathrm{A}x - \mathrm{B}z - c\|_2^2,$$

where we refer to β as the *penalty parameter*. The ADMM algorithm consists of the following steps:

$$x^{(k+1)} = \arg\min_x \mathcal{L}(x, z^{(k)}, \lambda^{(k)}; \beta) \tag{B.2}$$

$$z^{(k+1)} = \arg\min_z \mathcal{L}(x^{(k+1)}, z, \lambda^{(k)}; \beta) \tag{B.3}$$

$$\lambda^{(k+1)} = \mathrm{argmin}_\lambda\, \mathcal{L}(x^{(k+1)}, z^{(k+1)}, \lambda; \beta)\,, \tag{B.4}$$

that can be made explicit as follows:

$$\begin{align}
x^{(k+1)} =& \arg\min_x \left\{f(x) + \lambda^{(k)T}x + (\beta/2)\|x - z^{(k)}\|_2^2\right\} \tag{B.5}\\
z^{(k+1)} =& \arg\min_z \left\{g(z) + \lambda^{(k)T}z + (\beta/2)\|x^{(k+1)} - z\|_2^2\right\} \notag\\
\lambda^{(k+1)} =& \lambda^{(k)} - \beta(c - \mathrm{A}x - \mathrm{B}z). \tag{B.6}
\end{align}$$

Observe that the update in (B.4) reduces to a gradient ascent iteration with step-size equal to penalty parameter β in the Augmented Lagrangian. It is also worth remarking that the role of x and z *almost* symmetric; as a matter of fact, the inversion of the two updates in (B.2) and (B.3) may influence the convergence rate.
From (B.5)-(B.6), one can immediately derive the proximal formulation of ADMM, reading

$$\begin{aligned}
x^{(k+1)} =& \mathrm{prox}_{\frac{1}{\beta}f(\cdot)}\left(z^{(k)-\lambda^{(k)}}\right))\\
z^{(k+1)} =& \mathrm{prox}_{\frac{1}{\beta}g(\cdot)}\left(x^{(k+1)} + \lambda^{(k)}\right)\\
\lambda^{(k+1)} =& \lambda^{(k)} - \beta(c - \mathrm{A}x - \mathrm{B}z).
\end{aligned}$$

Adopting the proximal interpretation, it is clear the ADMM is very suitable when addressing problems such that the proximal operator of $f + g$ is difficult to obtain while the proximal operator of function f and g, separately, can be computed easily.

The convergence of ADMM in convex settings has been proved in [74, 65]:

Theorem B.1. *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and $g : \mathbb{R}^m \to \mathbb{R} \cup \infty$ be closed, proper and convex. Moreover, assume that the unaugmented Lagrangian,*

$$\mathcal{L}(x, z, \lambda; \beta) - \frac{\beta}{2}\|Ax - Bz - c\|_2^2$$

admits a saddle point. The iterative schemes (B.2)-(B.4) ensures that:

- *the residual $r^{(}k) = Ax^{(k)} + Bz^{(k)} - c \to 0$ as $k \to \infty$;*
- *the objective converges to its minimum $f(x^{(k)}) + g(z^{(k)}) \to f(x^*) + g(z^*)$, as $k \to \infty$;*
- *the dual variables converges $\lambda^{(k)} \to \lambda^*$, as $k \to \infty$.*

In the image restoration field, ADMM is usually applied to problems of the form,

$$u^* \in \arg\min_u \ \{\mathcal{R}(u) + \mu\mathcal{F}(u)\} \ .$$

When both $\mathcal{F}$ and $\mathcal{R}$ are convex and $\mu > 0$ is set, the convergence result in Theorem B.1 can be applied. In Chapter 5-7, we consider possibly convex models, namely when $p_i \geq 1$. Nevertheless, theorem B.1 can not be applied since in the designed models the regularization terms includes parameters updated along the iterations of the ADMM. Here we do not provide any convergence result, but we remark that in [89] the authors provide a proof of the convergence of ADMM when the regularization parameter μ is iteratively adapted along the iterations. As far as the non-convex case is concerned, we mention that a proof of the convergence of ADMM under certain assumptions of the objective function has been given in [158]. However, the convergence result in non-convex settings can not be applied here. Nevertheless, we experienced that the empirical convergence of ADMM is strictly related to the choice of suitable penalty parameters. In fact, their settings contribute to driving ADMM towards a local minimum or not. Hence, after a not trivial hand-tuning, we observed empirical convergence also in the non-convex case.

Bibliography

[1]

[2] R. Acar and C.R. Vogel. Analysis of bounded variation penalty methods for ill-posed problems. *Inverse Problems*, 10:1217–1229, 1994.

[3] M.V. Afonso, J.M. Bioucas-Dias, and M.A.T. Figueiredo. Fast image recovery using variable splitting and constrained optimization. *IEEE Transactions on Image Processing*, 19:2345–2356, 2010.

[4] C. Aguerrebere, A. Almansa, J. Delon, Y. Gousseau, and P. Musé. A Bayesian hyperprior approach for joint image denoising and interpolation, with an application to HDR imaging. *IEEE Transactions on Computational Imaging*, 3:633–646, 2017.

[5] L. Ambrosio, N. Fusco, and D. Pallara. Functions of bounded variation and free discontinuity problems. *Oxford Mathematical Monographs.*

[6] N. Antipa, G. Kuo, R. Heckel, B. Mildenhall, E. Bostan, R. Ng, and L. Waller. Diffusercam: Lensless single-exposure 3D imaging. *Optica*, 5, 2018.

[7] J.F. Aujol and A. Chambolle. Dual norms and image decomposition models. *International Journal of Computer Vision*, 63:85–104, 2005.

[8] S.D. Babacan, L. Mancera, R. Molina, and A.K. Katsaggelos. Non-convex priors in Bayesian compressed sensing. *2009 17th European Signal Processing Conference*, pages 110–114, 2009.

[9] S.D. Babacan, L. Mancera, R. Molina, and A.K. Katsaggelos. Non-convex priors in Bayesian compressed sensing. *European Signal Processing Conference*, pages 110–114, 2009.

[10] S.D. Babacan, R. Molina, and A.K. Katsaggelos. Fast Bayesian compressive sensing using laplace priors. *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2873–2876, 2009.

[11] I. Bayram and M.E. Kamasak. Directional Total Variation. *IEEE Signal Processing Letters*, 19:781–784, 2012.

[12] Z. Boukouvalas, S. Said, L. Bombrun, Y. Berthoumieu, and T. Adalı. A new Riemannian averaged fixed-point algorithm for mggd parameter estimation. *IEEE Signal Processing Letters*, 22:2314–2318, 2015.

[13] A.C. Bovik. *Handbook of Image and Video Processing (Communications, Networking and Multimedia)*. Academic Press, Inc., USA, 2005.

[14] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3:1–122, 2011.

[15] K. Bredies, K. Kunisch, and T. Pock. Total Generalized Variation. *SIAM Journal on Imaging Sciences*, 3:492–526, 2010.

[16] A. Buccini, Y. Park, and L. Reichel. Comparison of a-posteriori parameter choice rules for linear discrete ill-posed problems. *Journal of Computational and Applied Mathematics*, 2019.

[17] M. Burger, J. Flemming, and B. Hofmann. Convergence rates in ℓ_1-regularization if the sparsity assumption fails. *Inverse Problems*, 29, 2013.

[18] L. Calatroni, C. Cao, J.C. De Los Reyes, C.B. Schönlieb, and T. Valkonen. Bilevel approaches for learning of variational imaging models. *RADON book series, Variational methods*, 18.

[19] L. Calatroni, A. Lanza, M. Pragliola, and F. Sgallari. Adaptive parameter selection for weighted-TV image reconstruction problems. *arXiv:1905.11264*.

[20] L. Calatroni, A. Lanza, M. Pragliola, and F. Sgallari. A flexible space-variant anisotropic regularization for image restoration with automated parameter selection. *SIAM Journal on Imaging Sciences*, 12:1001–1037, 2019.

[21] D. Calvetti, H. Hakula, S. Pursiainen, and E. Somersalo. Conditionally Gaussian hypermodels for cerebral source localization. *SIAM Journal on Imaging Sciences*, 2:879–909, 2008.

[22] D. Calvetti, B. Lewis, and L. Reichel. GMRES, L-curves, and discrete ill-posed problems. *BIT Numerical Mathematics*, 42:44–65, 2002.

[23] D. Calvetti, A. Pascarella, F. Pitolli, E. Somersalo, and B. Vantaggi. A hierarchical Krylov–Bayes iterative inverse solver for MEG with physiological preconditioning. *Inverse Problems*, 31, 2015.

[24] D. Calvetti, A. Pascarella, F. Pitolli, E. Somersalo, and B. Vantaggi. Brain activity mapping from MEG data via a hierarchical Bayesian algorithm with automatic depth weighting. *Brain Topography*, 32:363–393, 2017.

[25] D. Calvetti, F. Pitolli, J. Prezioso, E. Somersalo, and B. Vantaggi. Priorconditioned CGLS-based quasi-MAP estimate, statistical stopping rule, and ranking of priors. *SIAM Journal on Scientific Computing*, 39, 2017.

[26] D. Calvetti, F. Pitolli, E. Somersalo, and B. Vantaggi. Bayes meets Krylov: Statistically inspired preconditioners for CGLS. *SIAM Review*, 60:429–461, 2018.

[27] D. Calvetti, M. Pragliola, E. Somersalo, and A. Strang. Sparse reconstructions from few noisy data: analysis of hierarchical Bayesian models with generalized gamma hyperpriors. *Inverse Problems*, 2019.

[28] D. Calvetti and E. Somersalo. A Gaussian hypermodel to recover blocky objects. *Inverse Problems*, 23:733–754.

[29] D. Calvetti and E. Somersalo. *Introduction to Bayesian Scientific Computing: Ten Lectures on Subjective Computing (Surveys and Tutorials in the Applied Mathematical Sciences).* Springer-Verlag, Berlin, Heidelberg, 2007.

[30] D. Calvetti and E. Somersalo. Hypermodels in the Bayesian imaging framework. *Inverse Problems*, 24, 2008.

[31] D. Calvetti and E. Somersalo. Inverse problems: from regularization to Bayesian inference. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10, 2018.

[32] D. Calvetti, E. Somersalo, and A. Strang. Hierarchical Bayesian models and sparsity: ℓ_2-magic. *Inverse Problems*, 35, 2008.

[33] E. J. Candès, M. B. Wakin, and S. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications*, 14:877–905, 2007.

[34] E.J. Candès, Y.C. Eldar, and D. Needell. Compressed sensing with coherent and redundant dictionaries. *Applied and Computational Harmonic Analysis*, 31:59–73, 2010.

[35] E.J. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. 2006.

[36] E.J. Candès, J.K. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52:489–509, 2004.

[37] E.J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51:4203–4215, 2005.

[38] S. Serra Capizzano. A note on antireflective boundary conditions and fast deblurring models. *SIAM Journal on Scientific Computing*, 25:1307–1325, 2004.

[39] V. Caselles, A. Chambolle, and M. Novaga. The discontinuity set of solutions of the TV denoising problem and some extensions. *Multiscale Modeling & Simulation*, 6:879–894, 2007.

[40] L. Chaâri, H. Batatia, N. Dobigeon, and J.Y. Tourneret. A hierarchical sparsity-smoothness Bayesian model for $\ell_0 + \ell_1 + \ell_2$ regulariz ation. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[41] A. Chambolle. An algorithm for Total Variation minimization and applications. *Journal of Mathematical Imaging and Vision*, 20:89–97, 2004.

[42] A. Chambolle, S. Levine, and B.J. Lucier. An upwind finite-difference method for Total Variation-based image smoothing. *SIAM Journal on Imaging Sciences*, 4:277–299, 2011.

[43] A. Chambolle and P.L. Lions. Image recovery via Total Variation minimization and related problems. *Numerische Mathematik*, 76:167–188, 1997.

[44] T. Chan and J. Shen. *Image Processing And Analysis: Variational, PDE, Wavelet, And Stochastic Methods.* Society for Industrial and Applied Mathematics, USA, 2005.

[45] R. Chartrand. Exact reconstruction of sparse signals via nonconvex minimization. *IEEE Signal Processing Letters*, 14:707–710, 2007.

[46] R. Chartrand and W. Yin. Iteratively reweighted algorithms for compressive sensing. *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3869–3872, 2008.

[47] G.A. Chen and D. Needell. Compressed sensing and dictionary learning. 2016.

[48] Y. Chen, S. Levine, and M. Rao. Variable exponent, linear growth functionals in image restoration. *SIAM Journal on Applied Mathematics*, 66:1383–1406, 2006.

[49] C.G. Chung, J.C. de los Reyes, and C.B. Schöenlieb. Learning optimal spatially-dependent regularization parameters in Total variation image denoising. *Inverse Problems*, 33, 2017.

[50] R. Ciak. *Coercive functions from a topological viewpoint and properties of minimizing sets of convex functions appearing in image restoration.* PhD thesis, Technische Universität Kaiserslautern, 2015.

[51] L. Condat. Discrete Total Variation: New definition and minimization. *SIAM Journal on Imaging Sciences*, 10:1258–1290, 2017.

[52] P. Craven and G. Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31:377–403, 1978.

[53] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. 2003.

[54] I. Daubechies, M. Defrise, and C. De Mol. Sparsity-enforcing regularisation and ISTA revisited. *Inverse Problems*, 32, 2016.

[55] I. Daubechies, R. Devore, M. Fornasier, and C. Sinan Gunturk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 63, 2010.

[56] I. Daubechies, R. Devore, M. Fornasier, and C.S. Gunturk. Iteratively re-weighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 63:1–38, 2010.

[57] D.C. Dobson and C.R. Vogel. Convergence of an iterative method for Total Variation denoising. 1997.

[58] Y.Q. Dong, M. Hintermüller, and M.M. Rincon-Camacho. Automated regularization parameter selection in a multi-scale variation model for image restoration. *Journal of Mathematical Imaging and Vision*, 40.

[59] D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52:1289–1306, 2004.

[60] D.L. Donoho. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Communicationson Pure and Applied Mathematics*, 59:797–829, 2006.

[61] D.L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via l minimization. *Proceedings of the National Academy of Sciences of the United States of America*, 100 5:2197–202, 2003.

[62] A. Durmus, E. Moulines, and M. Pereyra. Efficient Bayesian computation by proximal Markov Chain Monte Carlo: When Langevin meets Moreau. *SIAM Journal on Imaging Sciences*, 11:473–506, 2016.

[63] J.C. Pesquet E. Chouzenoux, A. Jezierska and H. Talbot. A majorize-minimize subspace approach for $\ell_2 - \ell_0$ image regularization. *SIAM Journal on Imaging Science*, 6:563–591, 2013.

[64] G.R. Easley, D. Labate, and F. Colonna. Shearlet based Total Variation for denoising. *IEEE Transactions on Image Processing*, 18:260–268, 2009.

[65] J. Eckstein and D.P. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55:293–318, 1992.

[66] D. Brie E.J. Duan, C. Soussen and J. Idier. A continuation approach to estimate a solution path of mixed $\ell_2 - \ell_0$ minimization problems. *Proc. Signal Processing with Adaptive Sparse Structured Representations Workshop*, 2009.

[67] M. Elad, P. Milanfar, and R. Rubinstein. Analysis versus synthesis in signal priors. *2006 14th European Signal Processing Conference*, 2006.

[68] H.W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems.* Springer Science & Business Media, 1996.

[69] Y. Fan and J. Nagy. Synthetic boundary conditions for image deblurring. *Linear Algebra and its Applications*, 434.

[70] C. Fenu, L. Reichel, and G. Rodriguez. GCV for Tikhonov regularization via global golub-kahan decomposition. *Numerical Lin. Alg. with Applic.*, 23:467–484, 2016.

[71] M.A.T. Figueiredo. Synthesis versus analysis in patch-based image priors. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1338–1342, 2017.

[72] M.A.T. Figueiredo, R. Nowak, and S. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing*, 1:586–597, 2007.

[73] M. Fornasier and H. Rauhut. Iterative thresholding algorithms. *Applied and Computational Harmonic Analysis*, 25:187–208, 2008.

[74] D. Gabay. Applications of the method of multipliers to variational inequalities. *Studies in Mathematics and Its Applications*, 15:299–331, 1983.

[75] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers and Mathematics with Application*, pages 17–40, 1976.

[76] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6:721–741, 1984.

[77] R. Giryes and D. Needell. Near oracle performance and block analysis of signal space greedy methods. *Journal of Approximation Theory*, 194:157–174, 2014.

[78] R. Glowinski and A. Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires. *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique*, 9:41–76, 1975.

[79] I.F. Gorodnitsky and B.D. Rao. Sparse signal reconstruction from limited data using focuss: a re-weighted minimum norm algorithm. *IEEE Trans. Signal Processing*, 45:600–616, 1997.

[80] M. Grasmair. Well-posedness and convergence rates for sparse regularization with sublinear ℓ^q penalty term. *Inverse Problems & Imaging*, 3:383–387, 2009.

[81] M. Grasmair, M. Haltmeier, and O. Scherzer. Sparse regularization with ℓ^q penalty term. *Inverse Problems*, 24, 2008.

[82] P.C. Hansen. The truncatedsvd as a method for regularization. *BIT Numerical Mathematics*, 27:534–553, 1987.

[83] P.C. Hansen. *Discrete Inverse Problems: Insight and Algorithms.* Society for Industrial and Applied Mathematics, USA, 2010.

[84] P.C. Hansen, T. Koldborg Jensen, and G. Rodriguez. An adaptive pruning algorithm for the discrete L-curve criterion. *Journal of Computational and Applied Mathematics*, 198:483–492, 2007.

[85] W. Hare and C. Sagastizábal. Computing proximal points of nonconvex functions. *Mathematical Programming*, 116:221–258, 2009.

[86] S. Häuser and G. Steidl. Convex multiclass segmentation with shearlet regularization. *International Journal of Computer Mathematics*, 90:62–81, 2011.

[87] B. He and X. Yuan. On the $o(1/n)$ convergence rate of the Douglas-Rachford Alternating Direction Method. *SIAM Journal on Numerical Analysis*, 50:700–709, 2012.

[88] C. He, C.H. Hu, W. Zhang, and B. Shi. A fast adaptive parameter estimation for Total Variation image restoration. *IEEE Transactions on Image Processing*, 23:4954–4967, 2014.

[89] C. He, Chang-Hua Hu, W. Zhang, and B. Shi. A fast adaptive parameter estimation for Total Variation image restoration. *IEEE Transactions on Image Processing*, 23:4954–4967, 2014.

[90] M.R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. 1952.

[91] M. Hintermüller, Rautenberg, and K. Papafitsoros. Analytical aspects of spatially adapted Total Variation regularisation. *Journal of Mathematical Analysis and Applications*, 454.

[92] M. Hintermüller and C.N. Rautenberg. Optimal selection of the regularization function in a weighted Total Variation model. part I: Modelling and theory. *Journal of Mathematical Imaging and Vision*, 59.

[93] M Hintermüller, C.N. Rautenberg, T. Wu, and L. Langer. Optimal selection of the regularization function in a weighted Total Variation model. part II: Algorithm, its analysis and numerical tests. *Journal of Mathematical Imaging and Vision*, 59.

[94] M. Hintermüller and T. Wu. Bilevel optimization for calibrating point spread functions in blind deconvolution. *Inverse Problems & Imaging*, 9:1139–1169, 2015.

[95] M.E. Hochstenbach, L. Reichel, and G. Rodriguez. Regularization parameter determination for discrete ill-posed problems. *J. Computational Applied Mathematics*, 273:132–149, 2014.

[96] M. Hong, Z. Luo, and M. Razaviyayn. Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. *SIAM Journal on Optimization*, 26:337–364, 2016.

[97] H.Rauhut, J. K. Romberg, and J. A. Tropp. Restricted isometries for partial random circulant matrices. *Applied and Computational Harmonic Analysis*, 32, 2012.

[98] Bangti J., Mass P., and Scherzer O. Sparsity regularization in inverse problems. *Inverse Problems*, 2017.

[99] K. Jalalzai. Some remarks on the staircasing phenomenon in Total Variation-based image denoising. *Journal of Mathematical Imaging and Vision*, 54:256–268, 2014.

[100] E.J. Candès J.L. Starck and D.L. Donoho. The curvelet transform for image denoising. *IEEE Transactions on Image Processing*, 11:670–684, 2002.

[101] J. Kaipio and E. Somersalo. Statistical inverse problems: Discretization, model reduction and inverse crimes. *Journal of Computational and Applied Mathematics*, 198:493–504, 2007.

[102] J.P. Kaipio and E. Somersalo. *Statistical and computational inverse problems.* Springer, 2005.

[103] R.D. Kongskov and Y. Dong. Directional Total Generalized Variation regularization for impulse noise removal. In *Scale Space and Variational Methods in Computer Vision*, pages 221–231, Cham, 2017. Springer International Publishing.

[104] R.D. Kongskov, Y. Dong, and K. Knudsen. Directional Total Generalized Variation regularization. *BIT Numerical Mathematics*, 59:903–928, 2019.

[105] M. Kowalski. Sparse regression using mixed norms. *Applied and Computational Harmonic Analysis*, 27:303–324, 2009.

[106] Torrésani B Kowalski, M. Sparsity and persistence: mixed norms provide simple signal models with dependent coefficients. *Signal, Image and Video Processing*, 3:251–264, 2009.

[107] F. Krahmer, D. Needell, and R. Ward. Compressive sensing with redundant dictionaries and structured measurements. *SIAM Journal on Mathematical Analysis*, 47:4606—4629, 2015.

[108] G. Kutyniok and D. Labate. *Shearlets: Multiscale Analysis for Multivariate Data.* Birkhäuser Basel, 2012.

[109] M. J. Lai, Y. Xu, and W. Yin. Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization. *SIAM Journal on Numerical Analysis*, 51:927–957, 2013.

[110] A. Lanza, S. Morigi, M. Pragliola, and F. Sgallari. Space-variant generalised Gaussian regularisation for image restoration. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 7:490–503, 2019.

[111] A. Lanza, S. Morigi, and F. Sgallari. Constrained TV_p-ℓ^2 model for image restoration. *Journal of Scientific Computing*, 68:64–91, 2016.

[112] F. Li, Z. Li, and L. Pi. Variable exponent functionals in image restoration. *Applied Mathematics and Computation*, 216:870 – 882, 2010.

[113] D.A. Lorenz. Convergence rates and source conditions for Tikhonov regularization with sparsity constraints. *Journal of Inverse and Ill-Posed Problems*, 16, 2008.

[114] D.A. Lorenz and E. Resmerita. Flexible sparse regularization. *Inverse Problems*, 33, 2016.

[115] Y. Lou, P. Yin, Q. He, and J. Xin. Computing sparse representation in a highly coherent dictionary based on difference of l_1 and l_2. *J. Sci. Comput.*, 64:178–196, 2015.

[116] J.J. Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien. *Comptes Rendus de l'Académie des Sciences - Mathematics*, 255:2897–2899, 1962.

[117] V.A. Morozov. On the solution of functional equations by the method of regularization, 1966.

[118] B.K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24:227–234, 1995.

[119] M. Niknejad, J.M. Bioucas-Dias, and M.A.T. Figueiredo. External patch-based image restoration using importance sampling. *IEEE Transactions on Image Processing*, 28:4460–4470, 2018.

[120] M. Nikolova. Local strong homogeneity of a regularized estimator. *SIAM Journal of Applied Mathematics*, 61:633–658, 2000.

[121] P. Ochs, R. Ranftl, T. Brox, and T. Pock. Techniques for gradient-based bilevel optimization with non-smooth lower level problems. *Journal of Mathematical Imaging and Vision*, 56:175–194, 2016.

[122] S. Osher, A.F. Solé, and L.A. Vese. Image decomposition and restoration using Total Variation minimization and the H^{-1} norm. 2003.

[123] K. Papafitsoros and C.B. Schönlieb. A combined first and second order variational approach for image reconstruction. *Journal of Mathematical Imaging and Vision*, 48:308–338, 2013.

[124] N. Parikh and S.P. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1:127–239, 2013.

[125] S. Parisotto. *Anisotropic variational models and PDEs for inverse imaging problems.* PhD thesis, 2019.

[126] S. Parisotto, J. Lellmann, S. Masnou, and C.B. Schönlieb. Higher-order Total Directional Variation. Part I: Imaging applications. 2018.

[127] S. Parisotto, S. Masnou, and C.B. Schönlieb. Higher-order Total Directional Variation. Part II: Analysis. 2018.

[128] Y. Park, L. Reichel, G. Rodriguez, and X. Yu. Parameter determination for Tikhonov regularization problems in general form. *J. Computational Applied Mathematics*, 343:12–25, 2018.

[129] F. Pascal, L. Bombrun, J. Tourneret, and Y. Berthoumieu. Parameter estimation for Multivariate Generalized Gaussian distributions. *IEEE Transactions on Signal Processing*, 61:5960–5971, 2013.

[130] T. Peleg and M. Elad. Performance guarantees of the thresholding algorithm for the cosparse analysis model. *IEEE Transactions on Information Theory*, 59:1832—1845, 2013.

[131] M. Pereyra. Proximal markov chain monte carlo algorithms. *Statistics and Computing*, 26:745–760, 2013.

[132] M. Elad R. Gribonval M.E. Davies R. Giryes, S. Nam. Greedy-like algorithms for the cosparse analysis model. *Linear Algebra and its Applications*, 441:22–60, 2014.

[133] R. Ramlau and E. Resmerita. Convergence rates for regularization with sparsity constraints. *Electronic Transactions on Numerical Analysis*, 37, 2010.

[134] L. Reichel and G. Rodriguez. Old and new parameter choice rules for discrete ill-posed problems. *Numerical Algorithms*, 63:65–87, 2012.

[135] W. Ring. Structural properties of solutions to Total Variation regularization problems. *ESAIM: Mathematical Modelling and Numerical Analysis*, 34:799–810, 2000.

[136] A. Roussos and P. Maragos. Tensor-based image diffusions derived from generalizations of the Total Variation and Beltrami functionals. pages 4141–4144, 2010.

[137] L.I. Rudin, S. Osher, and E. Fatemi. Nonlinear Total Variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60:259 – 268, 1992.

[138] M. Elad S. Nam, M. E. Davies and R. Gribonval. Greedy-like algorithms for the cosparse analysis model. *Applied and Computational Harmonic Analysis*, 34:30–56, 2013.

[139] R. Saab, R. Chartrand, and Ö. Yilmaz. Stable sparse approximations via nonconvex optimization. *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3885–3888, 2008.

[140] H. Scharr, M. J. Black, and H. W. Haussecker. Image statistics and anisotropic diffusion. 2:840–847, 2003.

[141] J.G. Serra, M. Testa, R. Molina, and A.K. Katsaggelos. Bayesian K-SVD using fast variational inference. *IEEE Transactions on Image Processing*, 26:3344–3359, 2017.

[142] J.G. Serra, S. Villena, R. Molina, and A.K. Katsaggelos. Greedy Bayesian double sparsity dictionary learning. *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1935–1939, 2017.

[143] K. Sharifi and A. Leon-Garcia. Estimation of shape parameter for generalized Gaussian distributions in subband decompositions of video. *IEEE Transactions on Circuits and Systems for Video Technology*, 5:52–56, 1995.

[144] K.S. Song. A globally convergent and consistent method for estimating the shape parameter of a generalized Gaussian distribution. *IEEE Transactions on Information Theory*, 52:510–527, 2006.

[145] J.L. Starck, M. Elad, and D.L. Donoho. Redundant multiscale transforms and their application for morphological component separation. *Advances in Imaging and Electron Physics Series*, 2004.

[146] J.L. Starck, M. Elad, and D.L. Donoho. Image decomposition via the combination of sparse representations and a variational approach. *IEEE Transactions on Image Processing*, 14:1570–1582, 2005.

[147] J.L. Starck, M.J. Fadili, and F. Murtagh. The undecimated wavelet decomposition and its reconstruction. *IEEE Transactions on Image Processing*, 16:297–309, 2007.

[148] G. Steidl and T. Teuber. Diffusion tensors for processing sheared and rotated rectangles. *IEEE Transactions on Image Processing*, 18:2640–2648, 2009.

[149] D.M. Strong, J.F. Aujol, and T. Chan. Scale recognition, regularization parameter selection, and Meyer's G norm in Total Variation regularization. *Multiscale Modelling and Simulation*, 5.

[150] A.M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

[151] A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation.* Society for Industrial and Applied Mathematics, USA, 2004.

[152] A. Tarantola and B. Valette. Inverse problems - quest for information. *Journal of Geophysics*, pages 159–170, 1982.

[153] H.L. Taylor, S.C. Banks, and J.F. McCoy. Deconvolution with the ℓ^1 norm. *Geophysics*, 44:39–52, 1979.

[154] R. Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B*, 73:273–282, 2011.

[155] A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-posed problems.* W.H. Winston, 1977.

[156] A.N. Tikhonov. On the stability of inverse problems. *Proceedings of the USSR Academy of Sciences*, 39:195–198, 1943.

[157] A.N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. 1963.

[158] Y. Wang, W. Yin, and J. Zeng. Global convergence of ADMM in nonconvex nonsmooth optimization. *Journal of Scientific Computing*, 78:29–63, 2019.

[159] J. Weickert. *Anisotropic Diffusion in Image Processing.* B.G. Teubner, Stuttgart, 1998.

[160] J. Weickert and H. Scharr. A scheme for coherence-enhancing diffusion filtering with optimized rotation invariance. *J. Visual Communication and Image Representation*, 13:103–118, 2002.

[161] J. Weickert and T.Brox. Diffusion and regularization of vector- and matrix-valued images. In *Inverse Problems, Image Analysis, and Medical Imaging*, pages 251–268. AMS, 2002.

[162] L.R. Welch. Lower bounds on the maximum cross correlation of signals. *IEEE Trans. Information Theory*, 20:397–399, 1973.

[163] F. Wen, L. Chu, P. Liu, and R. Caiming Qiu. A survey on nonconvex regularization-based sparse and low-rank recovery in signal processing, statistics, and machine learning. *IEEE Access*, 6:69883–69906, 2018.

[164] T. Mi Y. Liu and S. Li. Compressed sensing with general frames via optimal-dual-based ℓ_1-analysis. *IEEE Transactions on Information Theory*, 58:4201–4214, 2012.

[165] P. Yin, E. Esser, and J. Xin. Ratio and difference of ℓ_1 and ℓ_2 norms and sparse representation with coherent dictionaries. *Communications of the Association for Information Systems*, 14:87–109, 2014.

[166] C. Zarzer. On Tikhonov regularization with non-convex sparsity constraints. *Inverse Problems*, 25, 2009.

[167] W. Zhou, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004.