Automatic fidelity and regularization terms selection in variational image restoration

BIT manuscript No.
(will be inserted by the editor)

Automatic fidelity and regularization terms selection in variational image restoration

A. Lanza · M. Pragliola · F. Sgallari

Received: date / Accepted: date

Abstract In this paper we study a class of variational models for the image restoration inverse problem. Our main assumption is that the additive noise model and the image gradient magnitudes follow a generalized normal (GN) distribution, whose very flexible probability density function (pdf) is characterized by two parameters - typically unknown in real world applications - determining its shape and scale. The unknown image and parameters, which are both modeled as random variables in light of the hierarchical Bayesian perspective adopted here, are jointly automatically estimated within a Maximum A Posteriori (MAP) framework. The hypermodels resulting from the selected prior, likelihood and hyperprior pdfs are minimized by means of an alternating scheme which benefits from a robust initialization based on the noise whiteness property. For the minimization problem with respect to the image, we employ the Alternating Direction Method of Multipliers (ADMM) algorithm which takes advantage of efficient procedures for the solution of proximal maps. Computed examples show that the proposed approach holds the potential to automatically detect the noise distribution, and it is also well-suited to process a wide range of images.

1 Introduction

In this paper, we are interested in the restoration of images undergoing a degradation process of the form

$$b = \mathsf{K}u + e, \quad e \text{ realization of } E, \tag{1.1}$$

where the matrix $\mathsf{K} \in \mathbb{R}^{n\times n}$, representing the discretized linear blurring operator, is a large severely ill-conditioned and possibly rank-deficient matrix, and $u, b, e \in \mathbb{R}^n$ are

A. Lanza
Department of Mathematics, University of Bologna
E-mail: alessandro.lanza2@unibo.it

M. Pragliola
Department of Mathematics, University of Bologna
E-mail: monica.pragliola2@unibo.it

F. Sgallari
Department of Mathematics, University of Bologna
E-mail: fiorella.sgallari@unibo.it

vectorized forms of the discrete sought uncorrupted image, observed degraded image and additive noise image, respectively; the n-variate random vector E models the inherently probabilistic nature of noise corruption, and its pdf $\mathrm{Pr}_E(e) : \mathbb{R}^n \to \mathbb{R}_+$, with $\mathbb{R}_+$ denoting the set of non-negative real numbers, represents the largest information one can hope to possess about the unknown noise realization e in (1.1).

Determining the uncorrupted image u from the observed degraded image b is a discrete linear inverse problem and the more information is available on the image degradation process (1.1) and on the characteristics of the sought image u, the higher the chances for an accurate estimate u^* will be.

In this paper, we address the restoration of images undergoing (1.1) under the assumption that the noise corruption is independently and identically distributed (IID) with zero-mean generalized normal (GN) pdf, in short IIDGN, with unknown pdf parameters. The class of zero-mean IIDGN noises is larger than it may seem at a first glance. In fact, as detailed in the paper, the zero-mean GN pdf is a flexible family of distributions characterized by two parameters, a scale and a shape parameter. It thus contains some popular noise distributions such as, e.g. the normal, Laplace, hyper-Laplace (or impulsive) and, as a limiting case, uniform distributions.

In order to tackle the considered ill-posed inverse problem, we adopt a variational framework where the image restoration task is recast as the problem of seeking an estimate u^* of the original u among the minimizers of a suitable cost functional $J : \mathbb{R}^n \to \mathbb{R}$. The typical variational model for image restoration reads

$$u^* \in \arg\min_{u \in \mathbb{R}^n} J(u)\,, \quad J(u) := R(u) + \mu\, F(u;\mathsf{K},b)\,, \tag{1.2}$$

where the functionals R and F and the positive real scalar μ are referred to as *regularization term*, *fidelity term* and *regularization parameter*, respectively. In particular, R encodes prior information or beliefs available on the sought image u, while F measures the likelihood of any u given the knowledge of the observed data b and of the observation (or degradation) process, namely of the blur matrix and the noise distribution. Finally, the regularization parameter μ allows to balance solution regularity and trust in the data and is of crucial importance for obtaining good quality restorations.

It is quite well known - and it will be demonstrated in the paper - that whenever, in accordance with our assumption, the additive noise realization e in (1.1) is drawn from an IIDGN noise distribution having zero-mean, shape parameter $q \in \mathbb{R}_{++}$ and standard deviation $\sigma_r \in \mathbb{R}_{++}$ - with $\mathbb{R}_{++} = \mathbb{R}_+ \setminus \{0\}$ - the most suitable choice for F from a Bayesian probabilistic perspective is the so-called L_q fidelity term, namely

$$F(u;\mathsf{K},b) = \mathrm{L}_q(u;\mathsf{K},b) := \|r(u;\mathsf{K},b)\|_q^q\,, \quad r(u;\mathsf{K},b) := \mathsf{K}u - b \in \mathbb{R}^n, \tag{1.3}$$

where $r(u;\mathsf{K},b)$ represents the restoration residual image.

For what concerns R, the choice of modeling the sought image u by a Markov random field (MRF) but with a special Gibbs prior which, in analogy with the considered noise distributions, is of zero-mean IIDGN type with shape parameter $p \in \mathbb{R}_{++}$ and standard deviation $\sigma_g \in \mathbb{R}_{++}$, leads to the class of so-called TV_p regularizers, namely

$$R(u) = \mathrm{TV}_p(u) := \|g(u)\|_p^p\,, \quad g(u) := \left(\|(\mathsf{D}u)_1\|_2, \ldots, \|(\mathsf{D}u)_n\|_2\right)^T \in \mathbb{R}_+^n\,, \tag{1.4}$$

where $g(u)$ represents the vector of image gradient magnitudes, the matrix $\mathsf{D} \in \mathbb{R}^{2n\times n}$ is defined by $\mathsf{D} := \left(\mathsf{D}_h^T, \mathsf{D}_v^T\right)^T$ with $\mathsf{D}_h, \mathsf{D}_v \in \mathbb{R}^{n\times n}$ coefficient matrices of finite difference operators discretizing the first-order horizontal and vertical partial derivatives of image u, respectively, and where, with a little abuse of notation, $(\mathsf{D}u)_i := \big((\mathsf{D}_h u)_i, (\mathsf{D}_v u)_i\big) \in \mathbb{R}^2$ indicates the discrete gradient of u at pixel i [1].

Despite the fixed choice of the gradient as the 'inner' linear operator, the model proposed in this paper is general and other linear operators might be chosen as well, such as, e.g., higher-order differential operators or any transform operator. Indeed, the TV_p class of regularizers in (1.4) is large enough to model effectively the features of a sufficiently wide set of images, ranging from piecewise constant to smooth images.

Replacing (1.3) and (1.4) into (1.2), one obtains the class of TV_p-L_q variational models for image restoration, which read

$$u^* \in \underset{u\in\mathbb{R}^n}{\arg\min}\, J_{p,q}(u)\,, \quad J_{p,q}(u) := \mathrm{TV}_p(u) + \mu\, \mathrm{L}_q(u;\mathsf{K},b), \quad (p,q) \in \mathbb{R}^2_{++}\,. \tag{1.5}$$

The class of TV_p-L_q models contains some very popular members such as, e.g. the TV-L_2 [27], TV-L_1 [20,16] and, as a limiting case, TV-L_∞ models [17] - where TV stands for TV_1 and represents the standard Total Variation semi-norm - but it is 'larger' than its renowned members and the two free shape parameters p,q hold the potential for changing the functional form of the objective $J_{p,q}$ so as to deal effectively with wider sets of target images and of noise corruptions.

However, selecting manually or also by heuristic approaches the triplet (p,q,μ) of shape and regularization parameters yielding optimal or even only good quality restorations, so as to fully exploit the TV_p-L_q model potentialities, is a very hard task. In this paper, we propose an effective fully automatic approach for selecting (p,q,μ) based on a hierarchical Bayesian formulation of the problem and on MAP estimation.

1.1 Related work

Strategies for the automatic selection of only the regularization parameter μ have been proposed in literature for few fixed values of the shape parameters p,q.

For $p=q=2$, the TV_p-L_q model reduces to a standard Tikhonov-regularized least-squares problem. For this class of quadratic models, many heuristic approaches have been proposed for automatically selecting μ, such as, e.g., L-curve [11] and generalized cross-validation (GCV) [6]; on the other hand, several methods exploiting the information on the noise corruption have been designed. Within this class, we mention the discrepancy principle (DP) [34,5,26], which can be adopted whenever the noise level is assumed to be known, and the residual whiteness principle (RWP) [2], consisting in selecting μ which minimizes the residual normalized auto-correlation.

The automatic DP and GCV strategies have been extended to the TV-L_2 model in [32,33] and the former has been applied in [23] to the larger class of TV_p-L_2 models. Recently, a fully automatic selection strategy for μ based on the RWP has been

[1] In fact, from definitions in (1.4), we have $\mathrm{TV}_p(u) := \|g(u)\|_p^p = \big\|\big(\|(\mathsf{D}u)_1\|_2, \ldots, \|(\mathsf{D}u)_n\|_2\big)^T\big\|_p^p$ $= \sum_{i=1}^n \big|\,\big\|(\mathsf{D}u)_i\big\|_2\big|^p = \sum_{i=1}^n \big\|(\mathsf{D}u)_i\big\|_2^p$, which is the standard TV_p regularizer definition (see, e.g., [23]).

proposed [24] for a wide class of R-L_2 models, with R a set of (convex) regularizers containing the TV_p terms with $p \geq 1$.

The problem of estimating the regularization parameter and, more in general, the parameters arising in the regularization term $R(u)$, has been extensively discussed within a Bayesian framework. The probabilistic paradigm relies on the well-established connection between the classical variational regularizer and the *prior* probability density function (pdf), encoding information or beliefs available *a priori* on u. The presence of unknown or poorly known parameters in the prior, i.e. in the regularizer, accounts for the lack of meaningful and/or precise prior information which prevents from designing a fully determined prior distribution. The original information can be thus enriched relying on two main strategies, namely empirical and hierarchical Bayesian techniques. Empirical methods aim to select the unknown parameters by exploiting the information encoded in the data b, i.e. in the *likelihood* pdf - see, e.g., the recent works [29,19] and references therein. On the other hand, hierarchical approaches overcome the uncertainty in the prior by layering it and introducing additional priors on the parameters, referred to as *hyperpriors* [35,14]. The hierarchical formulation has been employed for the automatic selection of the sole μ in presence of L_2 fidelity term and TV_p regularizers with $p = 2$ ([25]), $p = 1$ ([4]) and $p \in [1,2]$ ([3]). For sparse recovery problems, more advanced hierarchical models have been proposed in [10,12,15,13], where the authors introduced locally varying zero-mean conditionally Gaussian priors with unknown variances for which gamma and GG hyperpriors have been considered; the space-variant parameters resulting in the penalty term, which is coupled with the L_2 fidelity term, are automatically estimated based on an iterative alternating scheme for the MAP formulation.

Besides the former strategies, that can be classified as heuristic or based on deterministic information, and the latter Bayesian approaches, we also mention a hybrid line of research aimed at selecting the functional form of the regularizer [23,9,22,7,8], but only coupled with the L_2 data term. More specifically, the variational models of interest are derived based on a probabilistic MAP formulation. However, the regularizer parameters are not recast in a Bayesian framework and their estimation is carried out either once for all in a preliminary phase [23] or as a Maximum Likelihood step interlaced with the iterations of the minimization algorithm [9,22,7,8].

To the best of authors' knowledge, the joint automatic selection of the regularization parameter μ and the shape parameters p,q of the TV_p-L_q class of models has not been dealt with before either in an empirical or hierarchical Bayesian perspective.

1.2 Contribution

In this paper, we present a fully automatic approach stemming from a Bayesian modeling of the image restoration inverse problem when the acquisition process is as in (1.1) under the assumption of zero-mean IIDGN noise corruption with unknown scale and shape parameters. The noise-related likelihood and the assumed Markovian prior lead to the TV_p-L_q class of variational models, whereas the inclusion of informative hyperpriors on the free parameters of the likelihood and the prior, which represents the key conceptual contribution of this work, provides a probabilistically grounded and transparent machinery for automatically selecting all the free parameters μ, p, q in the TV_p-L_q model. This framework gives rise to two different hypermodels, one

following a fully hierarchical Bayesian paradigm, the other taking advantage of the interaction between the introduced hyperpriors and the RWP. We also propose an automatic, efficient strategy for initializing - i.e., setting once for all from the observed degraded image b before starting the restoration - the hyperpriors based on a recent approach presented in [24] for different purposes.

From a numerical point of view, we address the minimization of a cost function of both the sought image u and the hyperparameter vector $\theta := (\sigma_g, p, \sigma_r, q)$. The function is non-convex jointly in (u, θ), hence disjoint sets of global minimizers may exist and, more importantly, the iterative minimization algorithm might get trapped in bad local minimizers. We propose an alternating minimization (AM) approach which, coupled with choosing the output of the hyperpriors initialization as the initial iterate, converges towards good quality solutions. For the minimization problem with respect to u, the classical ADMM algorithm is enriched with efficient and robust procedures for the solution of the proximal maps arising in the resulting sub-problems.

The proposed approach provides valuable benefits from different perspectives:

(i) From an applicative point of view, the automatic selection of the free parameters p, q, μ in the TV$_p$-L$_q$ models allows to apply the proposed framework to the restoration of a wide spectrum of images corrupted by different IIDGN noises.
(ii) On the numerical side, the proposed initialization strategy allows for a stable recovery of u and θ, despite non-convexity of the objective function.
(iii) Finally, from a modeling perspective the proposed hierarchical framework can be easily extended to other classes of parametrized variational models, also aimed at solving inverse imaging problems different from restoration (such as, e.g., image inpainting, super-resolution and computed tomography reconstruction) and characterized by general - i.e. rectangular and/or singular - forward linear operators K. Moreover, the proposed approach can be generalized by replacing the first order difference matrix D in the regularizer R with a matrix L representing, e.g., higher order difference operators or a generic transform operator. In fact, the only condition required by all theoretical derivations in this paper is that the null spaces of the forward operator K and the regularization operator L have trivial intersection.

The paper is organized as follows. In Sec. 2, we introduce useful definitions and properties and we detail the hierarchical MAP formulation for the model of interest. Then, in Sec. 3 we design suitable hyperpriors for the unknown parameters and introduce the two proposed hypermodels, whose numerical solution is addressed in Sec. 4. In Sec. 5, our method is extensively tested on different images. Finally, we conclude with some outlook for future research in Sec. 6.

2 The proposed model by probabilistic Bayesian formulation

In this section, we first motivate and then illustrate in detail the proposed approach. In particular, after introducing some notations and recalling some useful preliminaries, we derive the proposed variational model by recasting in fully probabilistic terms the image restoration inverse problem (1.1) under the assumption of additive zero-mean IIDGN noise corruption with unknown scale and shape parameters.

2.1 Notations and preliminaries

In the paper, we indicate random variables and their realizations by capital letters and their corresponding lower case letters, e.g. X and x, and we denote by $\Pr_X$, η_X, m_X and σ_X the pdf, mean, mode and standard deviation of random variable X, respectively; we will omit the subscript X when not necessary. The characteristic function χ_S of a set S is defined by $\chi_S(x) = 1$ if $x \in S$, 0 otherwise, whereas the indicator function ι_S is defined by $\iota_S(x) = 0$ if $x \in S$, $+\infty$ otherwise. We thus have $\iota_S(x) = -\ln \chi_S(x)$. We denote by 0_n the n-d column vector of all zeros.

In what follows, we recall few well-known definitions and in Proposition 2.1 we also introduce a function ϕ arising in the paper - for the proof we rely on [1].

Proposition 2.1 *The gamma function $\Gamma : \mathbb{R}_{++} \to \mathbb{R}$, the log-gamma function $\Lambda : \mathbb{R}_{++} \to \mathbb{R}$ and the function $\phi : \mathbb{R}_{++} \to \mathbb{R}$ defined by*

$$\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt\,, \quad \Lambda(x) = \ln \Gamma(x)\,, \quad \phi(x) := \sqrt{\Gamma(1/x)/\Gamma(3/x)}\,, \tag{2.1}$$

satisfy the following properties:

$\Gamma, \Lambda, \phi \in C^\infty(\mathbb{R}_{++})$, Γ is strongly convex, Λ is convex,

$\lim_{x \searrow 0} \Gamma(x) = \lim_{x \searrow 0} \Lambda(x) = \lim_{x \nearrow \infty} \Gamma(x) = \lim_{x \nearrow \infty} \Lambda(x) = +\infty$, $\lim_{x \searrow 0} \phi(x) = 0$, $\lim_{x \nearrow \infty} \phi(x) = \sqrt{3}$.

Definition 2.1 (Generalized Normal distribution) A scalar random variable X is generalized normal-distributed with mean $\eta \in \mathbb{R}$, standard deviation $\sigma \in \mathbb{R}_{++}$ and shape parameter $s \in \mathbb{R}_{++}$, denoted by $X \sim GN(\eta, \sigma, s)$, if its pdf has the form

$$\Pr_X(x) = \Pr_{GN}(x|\eta, \sigma, s) := \frac{1}{2\sigma} \frac{s}{\Gamma(1/s)\,\phi(s)} \exp\left(- \left| \frac{x-\eta}{\sigma\,\phi(s)} \right|^s \right), \quad x \in \mathbb{R}, \tag{2.2}$$

with Γ and ϕ functions defined in (2.1). In particular, for any fixed $\eta \in \mathbb{R}$, $\sigma \in \mathbb{R}_{++}$, the pdf in (2.2) converges pointwise to a uniform distribution as $s \to +\infty$, namely

$$\lim_{s \to +\infty} \Pr_{GN}(x|\eta, \sigma, s) = \left(1/(2\sqrt{3}\sigma)\right) \chi_{[0, \sqrt{3}\sigma]}(|x-\eta|)\,.$$

Definition 2.2 (Generalized Gamma distribution) A scalar random variable X is generalized gamma-distributed with mode $m \in \mathbb{R}_{++}$ and shape parameters $d \in (1, +\infty)$, $s \in \mathbb{R}_{++}$, denoted by $X \sim GG(m, d, s)$, if its pdf has the form

$$\Pr_X(x) = \Pr_{GG}(x|m, d, s) := \frac{s}{m\Gamma(d/s)} \left(\frac{d-1}{s} \right)^{d/s} \left(\frac{x}{m} \right)^{d-1} \exp\left(- \frac{d-1}{s} \left(\frac{x}{m} \right)^s \right), \tag{2.3}$$

with Γ and ϕ functions defined in (2.1).

In Fig. 2.1 we report the graphs of some of the introduced functions and pdfs. In particular, the middle panel shows the large flexibility of the GN family, hence the large potential capability of the considered class of TV_p-L_q models in dealing effectively with images - shape parameter p - and noises - shape parameter q - of very different type. Finally, in the next subsection we report the analysis of the Maximum Likelihood Esimation (MLE) strategy for the unknown parameters of a zero-mean GN distribution, which will be useful for our discussion.

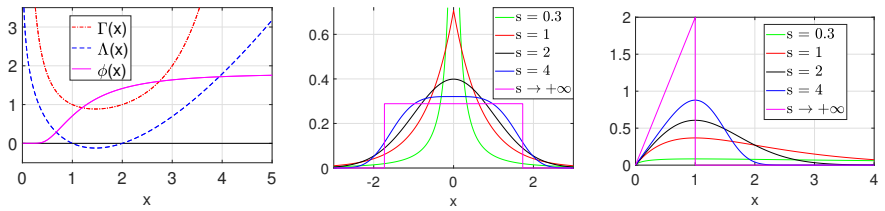

Fig. 2.1: From left to right, plots of functions Γ, Λ, ϕ defined in (2.1), of the GN pdf in (2.2) for $\eta = 0$, $\sigma = 1$ and some different values of the shape parameter s, of the GG pdf in (2.3) for $m = 1$, $d = 2$ and some different values of parameter s.

2.1.1 ML estimation of the unknown parameters of a zero-mean GN distribution

Let $X \sim GN(0, \sigma, s)$ with unknown scale and shape parameters $\sigma, s \in \mathbb{R}_{++}$, and let $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ be a vector of n independent realizations of X. According to the ML estimation approach, σ and s can be selected by maximizing the likelihood $\Pr(x \mid \sigma, s) = \prod_{i=1}^{n} \Pr_{GN}(x_i \mid 0, \sigma, s)$ or, equivalently, minimizing its negative logarithm. Recalling the GN pdf definition in (2.2), after some manipulations we have

$$\left\{\sigma^{\mathrm{ML}}, s^{\mathrm{ML}}\right\} \in \underset{(\sigma,s)\in\mathbb{R}^2_{++}}{\arg\min} \left\{G(\sigma, s; x) := n \ln[\Gamma(1+1/s)\phi(s)\sigma] + (\phi(s)\sigma)^{-s}\|x\|_s^s\right\}, \quad (2.4)$$

where $\Gamma(1+1/s) = (1/s)\Gamma(1/s)$. It comes from Proposition 2.1 that $G \in C^{\infty}(\mathbb{R}^2_{++})$ for any $x \in \mathbb{R}^n$ and that, for any fixed $s \in \mathbb{R}_{++}$, G is strictly convex and coercive in σ. By imposing $\partial G/\partial\sigma = 0$ and then replacing back into (2.4), it is a matter of simple algebra to verify that the ML estimates $\sigma^{\mathrm{ML}}, s^{\mathrm{ML}}$ in (2.4) are given by

$$s^{\mathrm{ML}} \in \underset{s\in\mathbb{R}_{++}}{\arg\min} \left\{ f^{\mathrm{ML}}(s) := \ln\Gamma(1+1/s) + (1/s)\left(1 + \ln s + \ln\left(\|x\|_s^s/n\right)\right) \right\}, \quad (2.5)$$

$$\sigma^{\mathrm{ML}} = (1/\phi(s^{\mathrm{ML}}))\left((s^{\mathrm{ML}}/n)\,\|x\|_{s^{\mathrm{ML}}}^{s^{\mathrm{ML}}}\right)^{1/s^{\mathrm{ML}}}. \quad (2.6)$$

Estimates $\sigma^{\mathrm{ML}}, s^{\mathrm{ML}}$ exist if (2.5) admits solutions. We note that $f^{\mathrm{ML}} \in C^{\infty}(\mathbb{R}_{++})$ and

$$\lim_{s\to+\infty} f(s) = \cancel{\ln\Gamma(1)} + \lim_{s\to+\infty}\left((1/s)\ln\left(\|x\|_s^s/n\right)\right) = \begin{cases} \ln\|x\|_\infty & \text{if } \|x\|_0 > 0 \\ -\infty & \text{otherwise.} \end{cases}$$

$$\lim_{s\searrow 0} f(s) = \lim_{s\searrow 0}\left(\ln\sqrt{2\pi/s} + (1/s)\ln(\|x\|_s^s/n)\right) = \begin{cases} +\infty & \text{if } \|x\|_0 = n \\ -\infty & \text{otherwise.} \end{cases}, \quad (2.7)$$

where the Stirling's series $\ln\Gamma(1/s) = (1/s)\ln(1/s) - s + \ln\sqrt{2\pi s} + O(s)$ has been used in (2.7). Hence, if $\|x\|_0 = n$, either the global minimizer(s) $s^{\mathrm{ML}} \in \mathbb{R}_{++}$ or a global infimizer is at $s^{\mathrm{ML}} = +\infty$. We can accept this latter scenario, as it indicates that samples are drawn from a uniform pdf, with $\sigma^{\mathrm{ML}} = \|x\|_\infty/\sqrt{3}$. Then, if $0 < \|x\|_0 < n$, function f in (2.5) is unbounded below as $\lim_{s\to 0} f^{\mathrm{ML}}(s) = -\infty$. This can be avoided by constraining $s \in [\underline{s}, +\infty)$, $\underline{s} \in \mathbb{R}_{++}$, as we will do in the paper. Finally, if $\|x\|_0 = 0$ then $\lim_{s\to 0} f^{\mathrm{ML}}(s) = \lim_{s\to+\infty} f^{\mathrm{ML}}(s) = -\infty$. However, this degenerate configuration, which arises when X follows a delta distribution, can be easily detected.

2.2 Model derivation by hierarchical Bayesian formulation and MAP estimation

A widely used approach for the derivation of variational models aimed at solving inverse problems consists in the fully probabilistic Bayesian formulation of the problem followed by a MAP estimation of the sought unknowns. The adoption of a Bayesian perspective requires to interpret all the unknowns involved in the image formation model (1.1) as random variables and the MAP estimation approach consists in deriving and then maximizing the *posterior* pdf of such unknowns. In order to get an analytical form for the posterior, we derive the expression of the *likelihood* from the probabilistic characteristics of the measurement model (1.1), assume a form for the *prior* and, eventually, for the *hyperprior*, and finally apply the Bayes' rule [28, 14].

Under our assumption that the additive noise in (1.1) is zero-mean IIDGN with the pair of scale-shape parameters given by $\theta_L := (\sigma_r, q) \in \mathbb{R}^2_{++}$, and recalling the definition of the residual image $r(u;\mathsf{K},b)$ in (1.3), the likelihood pdf takes the form

$$\Pr(b|u,\theta_L) = \left(\frac{1}{2\sigma_r}\frac{q}{\Gamma(1/q)\,\phi(q)}\right)^n \exp\left(-\left\|\frac{1}{\sigma_r\phi(q)}\,r(u;\mathsf{K},b)\right\|_q^q\right). \tag{2.8}$$

Modeling the unknown image u as a MRF with a Gibbs prior of IIDGN type with the pair of scale-shape parameters given by $\theta_P := (\sigma_g, p) \in \mathbb{R}^2_{++}$, and recalling the definition of the gradient norms vector $g(u)$ in (1.4), the prior pdf takes the form

$$\Pr(u|\theta_P) = \left(\frac{1}{2\sigma_g}\frac{p}{\Gamma(1/p)\,\phi(p)}\right)^n \exp\left(-\left\|\frac{1}{\sigma_g\phi(p)}\,g(u)\right\|_p^p\right). \tag{2.9}$$

Here, the so-called hyperparameter vector $\theta = (\theta_P, \theta_L) \in \mathbb{R}^4_{++}$ containing the free parameters of the likelihood and prior pdfs is unknown and has to be estimated jointly with the uncorrupted image u. The posterior pdf of (u,θ) given the data b reads

$$\begin{aligned}\Pr(u,\theta|b) = \Pr(u,\theta_P,\theta_L|b) &= \Pr(b|u,\theta_P,\theta_L)\Pr(u,\theta_P,\theta_L) \,/\, \Pr(b) &(2.10)\\ &= \Pr(b|u,\theta_P,\theta_L)\Pr(u|\theta_P,\theta_L)\Pr(\theta_P,\theta_L) \,/\, \Pr(b)\\ &= \Pr(b|u,\theta_L)\Pr(u|\theta_P)\Pr(\theta_P,\theta_L) \,/\, \Pr(b)\,, &(2.11)\end{aligned}$$

where (2.10) comes from applying the Bayes' rule and (2.11) from considering that b and θ_P are conditionally independent given u - hence $\mathrm{P}(b|u,\theta_P,\theta_L) = \mathrm{P}(b|u,\theta_L)$ - and that u and θ_L are conditionally independent given θ_P - hence $\mathrm{P}(u|\theta_P,\theta_L) = \mathrm{P}(u|\theta_P)$. Among the viable strategies explored in the literature and that could be applied here to extract meaningful information from the posterior - such as, e.g., the conditional mean or the minimum mean square error - we select as representative of $\Pr(u,\theta|b)$, according to the MAP approach, its mode, so that the joint estimates (u^*,θ^*) of the target image and the (prior and likelihood) hyperparameters are obtained by maximizing the posterior or, equivalently, minimizing its negative logarithm $-\ln\Pr(u,\theta|b)$:

$$\{u^*,\theta^*\} \in \underset{u\in\mathbb{R}^n,\,\theta\in\mathbb{R}^4_{++}}{\arg\min} \left\{-\ln\Pr(b|u,\theta_L) - \ln\Pr(u|\theta_P) - \ln\Pr(\theta_P,\theta_L)\right\}, \tag{2.12}$$

where (2.12) comes from (2.11) and from dropping the constant evidence term $\Pr(b)$. The MAP approach presents significative advantages in terms of computational efficiency. However, we point out that when the cost function in (2.12) is non-convex, the minimization algorithm may get trapped in bad local minima. Here, the typical downsides of MAP are held back by the design of a suitable initial guess - see Section 4.1 - which increases the algorithmic robustness.

Replacing (2.8) for the likelihood and (2.9) for the prior into the MAP inference formula (2.12), dropping the constants, rearranging and, then, recalling definitions (1.3) and (1.4) of the L_q fidelity and the TV_p regularizer, respectively, we obtain:

$$\{u^*,\theta^*\} \in \underset{u\in\mathbb{R}^n,\,\theta\in\mathbb{R}^4_{++}}{\arg\min} \{J_0(u,\theta) - \ln\Pr(\theta)\}, \tag{2.13}$$

where the hyperprior term $-\ln\Pr(\theta)$ will be made explicit in the next section, while the prior+likelihood function $J_0:\mathbb{R}^n\times\mathbb{R}^4_{++}\to\mathbb{R}$ reads

$$\begin{aligned} J_0(u,\theta) &= n\ln[\Gamma(1{+}1/p)\,\phi(p)\,\sigma_g] + (\phi(p)\,\sigma_g)^{-p}\,\mathrm{TV}_p(u) \\ &+ n\ln[\Gamma(1{+}1/q)\,\phi(q)\,\sigma_r] + (\phi(q)\,\sigma_r)^{-q}\,\mathrm{L}_q(u;\mathsf{K},b) \\ &= G(\sigma_g,p,g(u)) + G(\sigma_r,q,r(u)). \end{aligned} \tag{2.14}$$

In (2.14) we wrote the prior and likelihood terms in the same compact way based on function G defined in (2.4), with $x=g(u)$ or $x=r(u)$ - see definitions in (1.3)-(1.4) - regarded here as a third independent variable instead of a fixed parameter.

If no hyperpriors are considered or, equivalently, a flat (uniform) prior for the hyperparameter vector θ is used so that the term $-\ln\Pr(\theta)$ in (2.13) is constant, then the proposed model (2.13)-(2.14) reduces to minimizing the function J_0 only, jointly with respect to u and θ. However, it is easy to prove that J_0 is unbounded below, hence it does not admit global minimizers. In fact, for u a constant image, such that $\mathrm{TV}(u)=0$, J_0 tends to $-\infty$ as σ_g tends to 0. This represents a fatal weakness of the no-hyperprior model and indicates the necessity to introduce some hyperprior on θ.

3 Design of suitable hyperpriors

In this section, we define suitable hyperpriors $\Pr(\theta)$ for the proposed Bayesian MAP variational model (2.13)-(2.14). In particular, we design the hyperpriors with the threefold aim of (i) adjusting the fatal intrinsic weakness of the no-hyperprior model, (ii) allowing for an efficient solution and (iii) maximizing the beneficial effect of the hyperprior on the variational model solutions, i.e. on the quality of the attained restored images. The latter aim will be pursued by exploiting the peculiarities of the preliminary estimates provided by the initialization approach described in Sec. 4.1.

First, we rewrite the hyperprior $\Pr(\theta)$ in the following equivalent form

$$\Pr(\theta) = \Pr(\sigma_g,p,\sigma_r,q) = \Pr(\sigma_g,\sigma_r|p,q)\Pr(p,q). \tag{3.1}$$

In order to let the shape parameters to be selected in a fully free and automatic manner, we adopt a flat (uniform) hyperprior on the joint variable (p,q) over the set

$B = B_p \times B_q \subset \overline{\mathbb{R}}^2_{++}$, with $B_p = [\underline{p}, +\infty]$, $B_q = [\underline{q}, +\infty]$ and $\overline{\mathbb{R}}_{++} = \mathbb{R}_{++} \cup \{+\infty\}$:

$$\Pr(p,q) = \rho\, \chi_B(p,q) = \rho\, \chi_{B_p}(p) \chi_{B_q}(q)\,,\ \forall (p,q) \in \mathbb{R}^2_{++}\,,\ \rho \in \mathbb{R}_{++}\,. \tag{3.2}$$

Notice that we admit $q = +\infty$ - corresponding to a uniform noise - and $p = +\infty$ - which describes very peculiar configurations of the gradient magnitudes. In addition, as a general rule, $\underline{p}, \underline{q}$ are set in order to allow non-convex TV_p-L_q models, i.e. $\underline{p}, \underline{q} < 1$, which are known to be preferable in terms of sparsity promotion, and at the same time to avoid strongly non-convex configurations, i.e. $\underline{p}, \underline{q} \geq 0.5$, thus lowering the risk of getting stuck in local minima. Finally, notice that $-\ln \Pr(p,q) = \iota_B(p,q)$.

For what concerns the GN standard deviations σ_g, σ_r, two different approaches are discussed here. The first strategy is based on a fully hierarchical paradigm according to which σ_g and σ_r are modeled as independent random variables conditioned on p and q, respectively. In other words, the joint hyperprior in (3.1) takes the form

$$\Pr(\sigma_g, \sigma_r \mid p, q) = \Pr(\sigma_g \mid p) \Pr(\sigma_r \mid q)\,. \tag{3.3}$$

For both $\sigma_g \mid p$ and $\sigma_r \mid q$ we adopt GG hyperpriors - see definition (2.3) - with modes m_g and m_r and shape parameters d_g, p and d_r, q, respectively:

$$\Pr(\sigma_g \mid p) = \Pr_{GG}(\sigma_g \mid m_g, d_g, p)\,, \quad \Pr(\sigma_r \mid q) = \Pr_{GG}(\sigma_r \mid m_r, d_r, q)\,. \tag{3.4}$$

The family of GG distributions contains many notable distributions, such as, e.g., the gamma distribution for $p, q = 1$. In a number of recent works, the gamma and, more in general, the GG pdfs have been adopted as hyperpriors for the unknown variances of zero-mean Gaussian distributions used to model the entries of an assumed sparse signal [10, 12, 15, 13]. More specifically, the authors there exploit the heavy tail structure of the GG pdfs as it guarantees the realizations of outliers corresponding to the few non-zero entries of the signal. Here, our perspective is slightly different; in fact, we rely on a robust initialization for the parameters σ_g, σ_r that will be used to set the modes m_g, m_r. As a result, the parameters d_g, d_r will be fixed in order to tighten the GG pdfs around their modes, thus limiting the outcome of outliers. Hence, the choice of GG hyperpriors has to be interpreted here as a natural and - as explained next - advantageous way to constrain the unknown parameters within a small interval of the positive real line on which we rely with high confidence. More specifically, the advantage of a GG hyperprior is related to computational efficiency. In fact, as it will be illustrated in Sec. 4.2.2, this choice allows for independently updating the four hyperparameters p, σ_g, q, σ_r along the iterations of the proposed alternating minimization scheme, with explicit closed-form expressions for the scale parameters σ_g, σ_r.

By plugging (3.4) into (3.3), then (3.3) and (3.2) into (3.1), and finally taking the negative logarithm, the negative log-hyperprior takes the following compact form

$$-\ln \Pr(\theta) \;=\; H(\sigma_g, p; m_g, d_g) \;+\; H(\sigma_r, q; m_r, d_r) \;-\; \ln \rho\,, \tag{3.5}$$

with the (parametric) function $H(\sigma, s; m, d) : \mathbb{R}^2_{++} \to \mathbb{R}$ defined by

$$H(\sigma, s; m, d) = \ln\left(\frac{\sigma}{s}\Gamma\left(\frac{d}{s}\right)\right) + \frac{d}{s}\ln\frac{s}{d-1} - d\ln\frac{\sigma}{m} + \frac{d-1}{s}\left(\frac{\sigma}{m}\right)^s + \iota_B(s)\,, \tag{3.6}$$

where the last term $\iota_B(s)$, B being either B_p or B_q, accounts for the assumed constraints on the shape parameters p,q. Finally, replacing into (2.13) the expressions of J_0 in (2.14) and of $-\ln \Pr(\theta)$ in (3.5), our (first) complete hypermodel reads

$$\{u^*,\theta^*\} \in \underset{u\in\mathbb{R}^n,\,\theta\in\mathbb{R}^4_{++}}{\arg\min} \left\{J(u,\theta) := T(\sigma_g,p,g(u);m_g,d_g)+T(\sigma_r,q,r(u);m_r,d_r)\right\}, \quad (3.7)$$

where we dropped the constant $-\ln\rho$, we omitted the dependence of J on the constant parameters m_g,d_g,m_r,d_r, and function $T(\sigma,s,x;m,d) : \mathbb{R}^2_{++}\times\mathbb{R}^n \to \mathbb{R}$ reads

$$T(\sigma,s,x;m,d) := G(\sigma,s,x) + H(\sigma,s;m,d),$$

with functions G and H defined in (2.4) and (3.6), respectively.

As detailed in Sec. 4, problem (3.7), to which we refer as the H_1-TV_p-L_q hypermodel, can be solved by means of a very efficient strategy, which relies also on the existence of closed-form expressions for the unknowns σ_g, σ_r. However, as reported in Sec. 5, H_1-TV_p-L_q does not perform well on natural images characterized by textures and details at different scales. More specifically, the estimates of the GN shape parameters p,q are of good quality, while the scale parameters σ_g,σ_r are more likely to be misestimated, thus yielding a significant over-smoothing in the restorations.

To overcome this weakness, we introduce a second hypermodel H_2-TV_p-L_q which follows the outlined hierarchical Bayesian paradigm for obtaining the estimates u^*, p^*, q^*, σ_r^*, whereas σ_g^* is computed according to a bilevel optimization framework based on the RWP, so as to hold back the downsides of H_1-TV_p-L_q. In formulas:

$$\{u^*,p^*,q^*,\sigma_r^*\} = \left\{\tilde{u}(\sigma_g^*),\tilde{p}(\sigma_g^*),\tilde{q}(\sigma_g^*),\tilde{\sigma}_r(\sigma_g^*)\right\}, \text{ with} \quad (3.8)$$

$$\left\{\tilde{u}(\sigma_g),\tilde{p}(\sigma_g),\tilde{q}(\sigma_g),\tilde{\sigma}_r(\sigma_g)\right\} \in \underset{u\in\mathbb{R}^n,(p,q,\sigma_r)\in B\times\mathbb{R}_{++}}{\arg\min} J(u,p,q,\sigma_r;\sigma_g), \quad (3.9)$$

$$\text{and with } \sigma_g^* \in \underset{\sigma_g\in\mathbb{R}_{++}}{\arg\min} \left\{W(\sigma_g) := \mathscr{W}(\tilde{r}(\sigma_g))\right\}, \quad \tilde{r}(\sigma_g) = \mathsf{K}\tilde{u}(\sigma_g) - b. \quad (3.10)$$

In particular, in (3.9) the cost function J is the same defined in (3.7) for the first hypermodel H_1-TV_p-L_q but with σ_g regarded here as a free parameter. The residual whiteness function $\mathscr{W} : \mathbb{R}^n \to \mathbb{R}$ in (3.10) is the one introduced in [24] and leads to the function $W(\sigma) = (\mathscr{W}\circ\tilde{r})(\sigma) : \mathbb{R}_{++} \to \mathbb{R}$ whose explicit expression will be given in Sec. 4.1. The numerical examples in Sec. 5 will confirm the gain in terms of robustness attained with the introduction of this second hypermodel H_2-TV_p-L_q.

4 The Numerical Solution Algorithm

We now present an effective iterative method for the numerical solution of the proposed variational hypermodels in (3.7) and in (3.8)-(3.10), whose main steps are summarized in Alg. 1. The initialization strategy, which is common to the two approaches, is detailed in Sec. 4.1, while the u- and θ-update steps are discussed in Secs. 4.2 and 4.3, respectively. The algorithm relies on an alternating minimization (AM) scheme for which we do not provide a theoretical proof of convergence. However, in the numerical section we show clear empirical evidence of the good convergence behaviour

Algorithm 1 AM numerical algorithm for the solution of H-TV$_p$-L$_q$ hypermodels

inputs: observed image $b \in \mathbb{R}^n$, blurring operator $\mathsf{K} \in \mathbb{R}^{n\times n}$

outputs: estimated restored image $u^* \in \mathbb{R}^n$ and hyperparameters vector $\theta^* = (\sigma_g^*, p^*, \sigma_r^*, q^*) \in \mathbb{R}^4_{++}$

- **Initialization** (Sec. 4.1) **:**

· compute $u^{(0)} \in \mathbb{R}^n$, then $\theta^{(0)} \in \mathbb{R}^4_{++}$, then set hyperprior parameters $m_g, m_r \in \mathbb{R}_{++}$, $d_g, d_r \in (1, +\infty)$

- **AM for H$_1$-TV$_p$-L$_q$:**

for *k = 0, 1, 2, … until convergence* **do**:

· update $u^{(k+1)}$ (Sec. 4.2.1)

· update $\theta_P^{(k+1)} = (\sigma_g^{(k+1)}, p^{(k+1)})$ (Sec. 4.2.2)

· update $\theta_L^{(k+1)} = (\sigma_r^{(k+1)}, q^{(k+1)})$ (Sec. 4.2.2)

end for

- **AM for H$_2$-TV$_p$-L$_q$:**

for *k = 0, 1, 2, … until convergence* **do**:

· update $u^{(k+1)}, \sigma_g^{(k+1)}$ (Sec. 4.3)

· update $p^{(k+1)}$ (Sec. 2.1.1)

· update $\theta_L^{(k+1)} = (q^{(k+1)}, \sigma_r^{(k+1)})$ (Sec. 4.2.2)

end for

$u^* = u^{(k+1)}$, $\theta^* = \theta^{(k+1)}$

of the scheme. For the algorithm to be determined, we make the standard (and reasonable) assumption that the null spaces of matrices K and D have trivial intersection.

4.1 Robust initialization

Computing in a fully automatic and efficient manner an acceptable-quality initial estimate $u^{(0)}$ of u under the general assumption of IIDGN noise corruption with unknown scale and shape parameters is a hard task. However, it is of vital importance not only for setting suitable hyperprior parameters but also for the success of the alternating minimization scheme outlined in Alg. 1, which can get trapped in bad local minimizers as the joint optimization problem in (u, θ) is non-convex. Recently, a strategy for automatically selecting the regularization parameter of regularized least-square models for restoring images corrupted by additive white Gaussian noise has been proposed [24]. It is based on the RWP, i.e. on maximizing the whiteness of the restoration residual. The core block of this approach is applying the strategy to Tikhonov-regularized least-square (in short TIK-L$_2$) quadratic models.

The TV$_p$-L$_q$ model with $p = q = 2$ is a particular TIK-L$_2$ model, hence, also motivated by the fact that the noise whiteness property exploited by the RWP holds for any considered IIDGN noise corruption (note that IID property implies whiteness), we use the strategy in [24] for computing our initial estimate $u^{(0)}$. In formulas:

$$u^{(0)} = u_{\mathrm{Tik}}(\mu_W), \qquad u_{\mathrm{Tik}}(\mu) = \underset{u\in\mathbb{R}^n}{\arg\min} \left\{ \|\mathsf{D}u\|_2^2 + \mu \|\mathsf{K}u - b\|_2^2 \right\}, \tag{4.1}$$

with value μ_W automatically selected by minimizing the *whiteness measure function* $W : \mathbb{R}_{++} \to \mathbb{R}_+$ defined in [24]. Formally, μ_W is obtained by solving the 1-d problem

$$\mu_W \in \underset{\mu\in\mathbb{R}_{++}}{\arg\min} \left\{ W(\mu) = \left(\sum_i w_i^4(\mu)\right) / \left(\sum_i w_i^2(\mu)\right)^2 \right\}, \;\; w_i(\mu) = \varepsilon_i / (\eta_i + \zeta_i \mu), \tag{4.2}$$

with $\varepsilon_i, \eta_i, \zeta_i \in \mathbb{R}_+$, $i = 1, \dots, n$, defined in terms of the known quantities b, K, D [24]. Tests in Sec. 5.1 will confirm robustness of the RWP-based strategy in (4.1)-(4.2) when applied to images of different type corrupted by different IIDGN noises.

Given $u^{(0)}$, an initial estimate $\theta^{(0)} = (\theta_P^{(0)}, \theta_L^{(0)})$ is obtained by applying the MLE procedure in Sec. 2.1.1 to the data sets $g(u^{(0)}) \in \mathbb{R}^n_{++}$ for $\theta_P^{(0)}$ and $r(u^{(0)}) \in \mathbb{R}^n$ for $\theta_L^{(0)}$, respectively, with functions g, r defined in (1.3)-(1.4). More precisely, first we compute $p^{(0)}$ and $q^{(0)}$ by solving (2.5) with $x = g(u^{(0)})$ and $x = r(u^{(0)})$, respectively. The existence of solutions is guaranteed as we search for $p^{(0)}, q^{(0)}$ within the domains $B_p = [\underline{p}, +\infty]$, $B_q = [\underline{q}, +\infty]$, $\underline{p}, \underline{q} \in \mathbb{R}_{++}$, introduced in Section 3. Then, $\sigma_g^{(0)}, \sigma_r^{(0)}$ are obtained via (2.6) with $s^* = p^{(0)}$, $x = g(u^{(0)})$ and $s^* = q^{(0)}$, $x = r(u^{(0)})$, respectively. Finally, $\sigma_g^{(0)}$, $\sigma_r^{(0)}$ are used to set the modes of the GG hyperpriors on σ_g, σ_r for the first hypermodel and on the sole σ_r for the second one, whereas the two parameters d_r, d_g, accounting for the standard deviations of the GG hyperpriors, are set manually.

4.2 Alternating scheme for H_1-TV_p-L_q

Here, we discuss the u- and θ-updates for H_1-TV_p-L_q in the left side of Alg. 1.

4.2.1 Minimization with respect to u

Recalling the definition of $J(u, \theta)$ in (3.7), the u-update step for H_1-TV_p-L_q reads

$$
\begin{aligned}
u^{(k+1)} &\in \operatorname*{arg\,min}_{u \in \mathbb{R}^n} J\big(u, \theta^{(k)}\big) \\
&= \operatorname*{arg\,min}_{u \in \mathbb{R}^n} \left\{ \frac{\mathrm{TV}_{p^{(k)}}(u)}{\big(\sigma_g^{(k)} \phi\big(p^{(k)}\big)\big)^{p^{(k)}}} + \frac{\mathrm{L}_{q^{(k)}}(u; b, \mathsf{K})}{\big(\sigma_r^{(k)} \phi\big(q^{(k)}\big)\big)^{q^{(k)}}} \right\} \qquad (4.3) \\
&= \operatorname*{arg\,min}_{u \in \mathbb{R}^n} \left\{ J_{p,q}^{(k)}(u) := \sum_{i=1}^{n} \|(\mathsf{D}u)_i\|_2^{p^{(k)}} + \mu^{(k)} \|\mathsf{K}u - b\|_{q^{(k)}}^{q^{(k)}} \right\}, \qquad (4.4)
\end{aligned}
$$

where (4.4) comes from (4.3) by simple manipulations, by recalling definitions (1.3), (1.4) and by introducing the (current) regularization parameter $\mu^{(k)} \in \mathbb{R}_{++}$ with

$$
\mu^{(k)} := \big(\sigma_g^{(k)} \phi\big(p^{(k)}\big)\big)^{p^{(k)}} \big(\sigma_r^{(k)} \phi\big(q^{(k)}\big)\big)^{-q^{(k)}}. \qquad (4.5)
$$

In order to lighten the notation, in the rest of this section we drop the (outer) iteration index superscripts (k) and indicate by $\hat{u}$ - in place of $u^{(k+1)}$ - the sought solution. However, to avoid confusion, the (inner) iterations of the numerical algorithm proposed in the following for solving (4.4) are denoted by a different index j.

In Proposition 4.1, whose proof comes easily from Lemma 5.1 and Proposition 5.2 in [7] and from well-known arguments of convex analysis, we study $J_{p,q}$ in (4.4).

Proposition 4.1 *If the null spaces of matrices* K *and* D *have trivial intersection, then for any* $(\mu,p,q)\in\mathbb{R}^3_{++}$ *the function* $J_{p,q}$ *in (4.4) is continuous, bounded below by zero and coercive, hence it admits global minimizers. If* $p,q\geq 1$*, then* $J_{p,q}$ *is also convex, hence it admits a compact set of global minimizers. Finally, if* $p,q\geq 1$ *and* $pq>1$*, then* $J_{p,q}$ *is strictly convex, hence it admits a unique global minimizer.*

We solve (4.4) by means of an ADMM-based approach. First, we resort to the variable splitting strategy and rewrite (4.4) in the equivalent linearly constrained form

$$\{\hat{u},\hat{t},\hat{r}\}\in\underset{u,r\in\mathbb{R}^n,\,t\in\mathbb{R}^{2n}}{\arg\min}\left\{\sum_{i=1}^{n}\|t_i\|_2^p+\mu\,\|r\|_q^q\right\}\ \text{s.t.}:\ t=\mathsf{D}u,\ \ r=\mathsf{K}u-b, \tag{4.6}$$

where $t\in\mathbb{R}^{2n}$ and $r\in\mathbb{R}^n$ are the newly introduced variables and where we define $t_i:=((\mathsf{D}_h u)_i,(\mathsf{D}_v u)_i)^T\in\mathbb{R}^2$. The augmented Lagrangian function for (4.6) reads

$$\begin{aligned}\mathscr{L}(u,t,r,\lambda_t,\lambda_r)=&\sum_{i=1}^{n}\|t_i\|_2^p+\mu\,\|r\|_q^q-\langle\,\lambda_t\,,t-\mathsf{D}u\,\rangle+(\beta_t/2)\,\|\,t-\mathsf{D}u\,\|_2^2\\&-\langle\,\lambda_r\,,r-(\mathsf{K}u-b)\,\rangle+(\beta_r/2)\,\|\,r-(\mathsf{K}u-b)\,\|_2^2,\end{aligned} \tag{4.7}$$

where $\beta_t,\beta_r\in\mathbb{R}_{++}$ are penalty parameters and $\lambda_t\in\mathbb{R}^{2n}$, $\lambda_r\in\mathbb{R}^n$ are the vectors of Lagrange multipliers associated with the set of $2n+n$ linear constraints in (4.6).

Upon suitable initialization, and for any $j\geq 0$, the j-th iteration of the standard ADMM applied to computing a saddle-point of $\mathscr{L}$ defined in (4.7) reads as follows:

$$u^{(j+1)}=\left(\beta_t\mathsf{D}^T\mathsf{D}+\beta_r\mathsf{K}^T\mathsf{K}\right)^{-1}\left(\beta_t\mathsf{D}^T w^{(j)}+\beta_r\mathsf{K}^T v^{(j)}\right), \tag{4.8}$$

$$\text{with}:\ \ w^{(j)}=t^{(j)}-\frac{1}{\beta_t}\lambda_t^{(j)},\quad v^{(j)}=r^{(j)}-\frac{1}{\beta_r}\lambda_r^{(j)}+b,$$

$$t^{(j+1)}\in\underset{t\in\mathbb{R}^{2n}}{\arg\min}\left\{\sum_{i=1}^{n}\|t_i\|_2^p+\frac{\gamma_p}{2}\left\|t-y_p^{(j)}\right\|_2^2\right\}, \tag{4.9}$$

$$\text{with}:\ \ \gamma_p=\beta_t,\ \ y_p^{(j)}=\mathsf{D}u^{(j+1)}+\frac{1}{\beta_t}\lambda_t^{(j)},$$

$$r^{(j+1)}\in\underset{r\in\mathbb{R}^n}{\arg\min}\left\{\|r\|_q^q+\frac{\gamma_q}{2}\left\|r-y_q^{(j)}\right\|_2^2\right\}, \tag{4.10}$$

$$\text{with}:\ \ \gamma_q=\frac{\beta_r}{\mu},\ \ y_q^{(j)}=\mathsf{K}u^{(j+1)}-b+\frac{1}{\beta_r}\lambda_r^{(j)},$$

$$\lambda_t^{(j+1)}=\lambda_t^{(j)}-\beta_t\left(t^{(j+1)}-\mathsf{D}u^{(j+1)}\right).$$

$$\lambda_r^{(j+1)}=\lambda_r^{(j)}-\beta_r\left(r^{(j+1)}-\left(\mathsf{K}u^{(j+1)}-b\right)\right).$$

The sub-problem for u in (4.8) is a linear system which - under the assumption that the null spaces of K and D have trivial intersection - is solvable as the coefficient matrix is symmetric positive definite. Adopting periodic boundary conditions for u, $\mathsf{D}^T\mathsf{D}$ and $\mathsf{K}^T\mathsf{K}$ are block-circulant with circulant blocks matrices that can be diagonalized by the 2D discrete Fourier transform. The solution is thus obtained at the cost of $O(n\ln n)$ operations by using the 2D fast Fourier transform implementation.

The two remaining sub-problems for variables t and r in (4.9)-(4.10) can both be split into n independent (2-dimensional and 1-dimensional, respectively) minimization problems taking the form of proximity operators, in formula:

$$\begin{aligned} t_i^{(j+1)} &\in \underset{x\in\mathbb{R}^2}{\arg\min}\left\{ \|x\|_2^p + \tfrac{\gamma_p}{2}\left\|x - y_{p,i}^{(j)}\right\|_2^2 \right\} = \operatorname{prox}_{||x||_2^p}^{\gamma_p}\left(y_{p,i}^{(j)}\right) \\ r_i^{(j+1)} &\in \underset{x\in\mathbb{R}}{\arg\min}\left\{ |x|^q + \tfrac{\gamma_q}{2}\left(x - y_{q,i}^{(j)}\right)^2 \right\} = \operatorname{prox}_{|x|^q}^{\gamma_q}\left(y_{q,i}^{(j)}\right) \end{aligned}, \quad i = 1,\dots,n. \quad (4.11)$$

Hence, solving both sub-problems for t and r reduces to computing the proximal operator of the parametric function $f_s(x) := ||x||_2^s$, $s \in \mathbb{R}_{++}$, with $x \in \mathbb{R}^z$, $z \in \{1,2\}$. We remark that in the proposed AM approach - see Alg. 1 - the parameters p and q are automatically estimated at each outer iteration k, hence both p and q can assume whichever real positive value. We thus need a reliable and efficient strategy for computing the proximal map of f_s for any $s \in \mathbb{R}_{++}$. For the sake of better readability, here we only summarize the content of the derived results, which are rather technical and, for this reason, are reported in the Appendix. In particular, in Proposition A.1, we extend to the case $s > 2$ the results in Proposition 1 of [23], which in their turn extended to the z-d case the 1-d results presented in [36] for $s < 1$. We can conclude that, for any $s, \gamma \in \mathbb{R}_{++}$, $\operatorname{prox}_{f_s}^{\gamma}$ takes the form of a shrinkage operator, such that (4.11) reads

$$t_i^{(j+1)} = \xi_{p,i}^*\, y_{p,i}^{(j)}, \qquad r_i^{(j+1)} = \xi_{q,i}^*\, y_{q,i}^{(j)}, \qquad i = 1,\dots,n, \quad (4.12)$$

where, depending on p, q, the shrinkage factors $\xi_{p,i}^*$, $\xi_{q,i}^*$ in (4.12) can either be expressed in closed-form or be the unique solution of a non-linear equation - see Proposition A.1. For this latter case, in Corollary A.1 we prove that the Newton-Raphson scheme with suitable initialization can be applied with guarantee of convergence.

4.2.2 Minimization with respect to θ

Based on the definition of our cost function $J(u,\theta)$ in (3.7), the θ-update reads

$$\theta^{(k+1)} \in \underset{\theta\in\mathbb{R}_{++}^4}{\arg\min}\, J\left(u^{(k+1)},\theta\right) \Longleftrightarrow \begin{cases} \theta_P^{(k+1)} = \{\sigma_g^{(k+1)}, p^{(k+1)}\} = \{\sigma_g^{\mathrm{MAP}}, p^{\mathrm{MAP}}\} \\ \theta_L^{(k+1)} = \{\sigma_r^{(k+1)}, q^{(k+1)}\} = \{\sigma_r^{\mathrm{MAP}}, q^{\mathrm{MAP}}\} \end{cases},$$

where $\{\sigma^{\mathrm{MAP}}, s^{\mathrm{MAP}}\}$, either coinciding with $\theta_P^{(k+1)}$ or with $\theta_L^{(k+1)}$, is obtained as

$$\left\{\sigma^{\mathrm{MAP}}, s^{\mathrm{MAP}}\right\} \in \underset{(\sigma,s)\in\mathbb{R}_{++}^2}{\arg\min}\ \left\{ T(\sigma,s;x,m,d) = G(\sigma,s;x) + H(\sigma,s;m,d) \right\}, \quad (4.13)$$

with $x = g(u^{(k+1)})$ for $\theta_P^{(k+1)}$ and $x = r(u^{(k+1)})$ for $\theta_L^{(k+1)}$, and where functions G and H are defined in (2.4) and in (3.6), respectively. The MAP superscript indicates that (4.13) can be regarded as a MAP estimation problem for (σ,s), which reduces to the MLE problem (2.4) when the selected hyperprior is a uniform pdf. In analogy with the results reported for the MLE problem in Section 2.1.1, we can prove the following proposition on the existence of solutions for problem (4.13).

Proposition 4.2 *Let $T(\cdot;x,m,d):\mathbb{R}^2_{++}\to\mathbb{R}$ be the cost function of problem* (4.13)*, $x\in\mathbb{R}^n$ and $m\in\mathbb{R}_{++}, d\in(1,+\infty)$. Then, $T\in C^\infty(\mathbb{R}^2_{++})$. Furthermore, the minimizers of T can be obtained as*

$$s^{\mathrm{MAP}}\in\operatorname*{arg\,min}_{s\in B} f^{\mathrm{MAP}}(s)\,,\quad \sigma^{\mathrm{MAP}}=m\left(\frac{-\alpha+\sqrt{\alpha^2+4\beta m^{-s^{\mathrm{MAP}}}(\sigma^{\mathrm{ML}})^{s^{\mathrm{MAP}}}}}{2\beta}\right)^{1/s^{\mathrm{MAP}}},$$

with $B=[\underline{s},+\infty]$, $\underline{s}\in\mathbb{R}_{++}$, $\alpha:=(n-d+1)/n$, $\beta:=(d-1)/n$, σ_{ML} given in (2.7) *and*

$$f^{\mathrm{MAP}}(s)=n\ln[\Gamma(1+1/s)\phi(s)]+\ln[(1/d)\Gamma(1+d/s)]+(d/s)\ln(s/n\beta)$$
$$+(n/s)\big(\sqrt{\alpha^2+4\beta m^{-s}\sigma_{\mathrm{ML}}(s)^s}+\alpha\ln((-\alpha+\sqrt{\alpha^2+4\beta m^{-s}\sigma_{\mathrm{ML}}(s)^s})/(2\beta))\big)\,.$$

4.3 Alternating scheme for H_2-TV_p-L_q

We now discuss the numerical solution of H_2-TV_p-L_q defined in (3.8)-(3.10), whose main steps are outlined in the right side of Alg. 1. The bilevel structure of the model is simply dealt with by jointly updating the unknowns u and σ_g. More precisely, after the initialization step, at each outer AM iteration we first update for (u,σ_g) and then for (p,q,σ_r). The former update is carried out by solving problem (4.3)-(4.4) for different values of σ_g. For each considered σ_g, we compute the whiteness measure W as in (4.2), and select as updated $u^{(k+1)}$, $\sigma_g^{(k+1)}$ the ones for which W is the smallest. The latter update is performed by means of the procedure outlined in Sec. 4.2.2 for q and σ_r, while $p^{(k+1)}$ is sought as a solution of problem (2.4) constrained to the domain B_p with $\sigma_g=\sigma_g^{(k+1)}$ and $x=g(u^{(k+1)})$.

5 Computed examples

In this section, we assess the performance of the proposed fully automatic variational hypermodel in its two versions H_1-TV_p-L_q in (3.7) and H_2-TV_p-L_q in (3.8)-(3.10), solved by the iterative approaches outlined in Alg. 1. After defining the experimental setting, in Sec. 5.1 we evaluate the robustness of the initialization approach described in Sec. 4.1, then the two hypermodels are extensively tested and compared in Sec. 5.2. Finally, the performance of the H_2-TV_p-L_q hypermodel is evaluated in absolute terms in Sec. 5.3, also by comparing it with some popular (not fully automatic) competitors.

In order to highlight the great flexibility of the proposed fully automatic image restoration approach, we consider four test images characterized by very different properties; see Fig. 5.1. In particular, `qrcode` is a purely geometric, piecewise constant image, `sinusoid` a prototypical smooth image, `peppers` a piecewise smooth image and `skyscrapers`, the most difficult to be dealt with, is made by a mixture of piecewise constant, smooth and textured components. For each (uncorrupted) test image $\bar{u}$, we also report the quantities $\bar{p}$, $\bar{\sigma}_g$, representing the "target" values of the

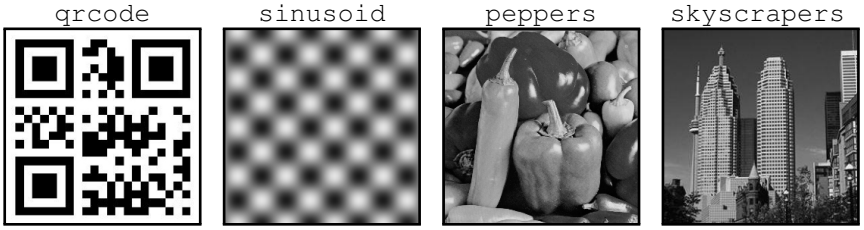

Fig. 5.1: Original test images $\bar{u}$: `qrcode` (256×256, $\bar{p} = 0.7$, $\bar{\sigma}_g = 0.2408$), `sinusoid` (200×200, $\bar{p} = 98$, $\bar{\sigma}_g = 0.0348$), `peppers` (256×256, $\bar{p} = 0.72$, $\bar{\sigma}_g = 0.1008$), `skyscraper` (256×256, $\bar{p} = 0.7$, $\bar{\sigma}_g = 0.1572$).

shape and scale parameters of the underlying GN prior pdf. These values are obtained by applying the MLE procedure for GN pdf parameters outlined in Sec. 2.1.1 to the data set $x = (\|(\mathsf{D}\bar{u})_1\|_2, \ldots, \|(\mathsf{D}\bar{u})_n\|_2)$, namely the gradient norms of the target uncorrupted image $\bar{u}$, where, here and in the following experiments, the s estimation problem - with $s = p, q$ - is addressed via gridsearch; $\bar{p}, \bar{\sigma}_g$ will serve as reference values with which to compare the obtained estimates p^*, σ_g^*. The $\bar{p}$ value ranges from 0.7 for the piecewise constant image `qrcode` to 98 for the smooth image `sinusoid`. The overbar notation is also used to indicate the true values $\bar{q}$, $\bar{\sigma}_r$ of the shape and scale parameters of the GN likelihood pdf (i.e., of the IIDGN noise pdf).

In the experiments, all test images have been corrupted by space-invariant Gaussian blur defined by a 2D discrete convolution kernel generated using the Matlab command `fspecial('gaussian',band,sigma)` with parameters `band` = 5 and `sigma` = 1. In particular, `band` represents the side length (in pixels) of the square support of the kernel, whereas `sigma` is the standard deviation of the circular, zero-mean, bivariate Gaussian pdf representing the kernel in the continuous setting. We assume periodic boundary conditions for image $\bar{u}$, hence the blur matrix $\mathsf{K} \in \mathbb{R}^{n \times n}$ is block-circulant with circulant blocks and can be diagonalized in $\mathbb{C}$ by the 2D discrete Fourier transform, thus allowing fast multiplication and storage saving. In accordance with the considered degradation model (1.1), after applying K to $\bar{u}$, the blurred image $\mathsf{K}\bar{u}$ is additively corrupted by IIDGN noise realizations $e \in \mathbb{R}^n$ from GN pdfs with standard deviation $\bar{\sigma}_r = 0.1$ and different shape parameters $\bar{q}$, ranging from 0.5, which is the case of a strongly impulsive GN noise, to $+\infty$, namely uniform noise, passing through the two notable intermediate values $\bar{q} = 1$ and $\bar{q} = 2$ associated with Laplacian and Gaussian noises, respectively. The blur- and noise-corrupted images $b = \mathsf{K}\bar{u} + e$ are shown in the first column of Fig. 5.3 for `qrcode` and `sinusoid` corrupted by uniform noise and in the first rows of Figs. 5.4-5.5 for `peppers` and `skyscraper` degraded by impulsive, Laplacian, Gaussian and uniform noises.

The quality of the obtained restored images u^* versus the associated true images $\bar{u}$ has been assessed by means of two scalar measures, the Improved Signal-to-Noise Ratio (ISNR), $\mathrm{ISNR}(b, \bar{u}, u^*) := 10 \log_{10} \left(\|b - \bar{u}\|_2^2 / \|u^* - \bar{u}\|_2^2 \right)$, and the Structural Similarity Index (SSIM) [31]. The larger the ISNR and SSIM values, the higher the restoration quality. For all tests, both the outer AM scheme iterations and the inner

ADMM iterations are stopped as soon as $\delta_u^{(k)} := \left\|u^{(k)} - u^{(k-1)}\right\|_2 / \left\|u^{(k-1)}\right\|_2 < 10^{-5}$, $k \in \mathbb{N} \setminus \{0\}$, and the ADMM penalty parameters β_t, β_r have been set manually.

5.1 Robustness evaluation of the RWP-based Tikhonov initialization

We analyze experimentally the automatic Tikhonov initialization approach outlined in Sec. 4.1 and based on the RWP. In particular, it is of crucial importance to show that the approach is effective for the wide range of images and noises considered. To this aim, in Fig. 5.2 we graphic, for all test images and some different $\bar{q}$ values, the whiteness measure function $W(\mu)$ in (4.2), but reparameterized as a function of the scalar quantity $\tau(\mu) := \hat{\sigma}_r(\mu)/\bar{\sigma}_r$, where $\bar{\sigma}_r$ is the true noise standard deviation and $\hat{\sigma}_r(\mu) := \| \mathsf{K} u_{\text{Tik}}(\mu) - b \|_2 / \sqrt{n}$ represents its μ-dependent estimate based on the solution $u_{\text{Tik}}(\mu)$ of the TIK-L_2 model in (4.1). Plotting W as a function of $\tau(\mu)$ (instead of μ) helps in detecting immediately the quality of the estimate $u^{(0)}$ in terms of associated discrepancy. In each plot of Fig. 5.2, the dashed colored vertical lines represent the optimal $\tau_W := \tau(\mu_W) = \hat{\sigma}_r(\mu_W)/\bar{\sigma}_r$ values selected according to the RWP (global minimizers of $W(\tau)$), while the black vertical line at $\tau = 1$ corresponds to a perfect estimate $\hat{\sigma}_r$ of $\bar{\sigma}_r$.

In analogy to what observed in [24], where the same study has been done only for $q = 2$, first we notice that the RWP tends to slightly under-estimate the true standard deviation $\bar{\sigma}_r$. In general, this circumstance has already been proved to be preferable since it leads to restorations with higher ISNR and SSIM values [9,24]. More important - if not crucial - to the purpose of this work, one can notice from Fig. 5.2 that, for each test image, the function W presents approximately the same behavior and, hence, optimal τ_W value, for any different $\bar{q}$ value. This reflects the expected power of the RWP which, relying on the residual whiteness property, allows to deal equally well with IID noises having very different distributions, like, e.g., the hyper-Laplacian, Laplacian, Gaussian and uniform distributions considered here.

In Tab. 5.1 we report the ISNR and SSIM values associated with the initialized image $u^{(0)} = u_{\text{Tik}}(\mu_W)$ as well as the hyperparameter estimates $p^{(0)}$, $q^{(0)}$, $\sigma_r^{(0)}$, $\sigma_g^{(0)}$ for all test images and noise shape parameters $\bar{q} \in \{0.5, 1, 2, +\infty\}$. We note that $q^{(0)}$ is a good approximation of the shape parameter $\bar{q}$ of the true underlying noise distribution, while $p^{(0)}$ is a less accurate estimate of the target value $\bar{p}$.

5.2 Comparative performance evaluation of the two variational hypermodels

We now evaluate the accuracy of the results obtainable by the two proposed hypermodels in (3.7) and (3.8)-(3.10), solved via the iterative schemes in Alg. 1. We remark that, once for all after the initialization phase, for the H_1-TV_p-L_q hypermodel the modes m_g, m_r of the two GG hyperpriors are set equal to the estimates $\sigma_g^{(0)}$, $\sigma_r^{(0)}$, respectively, while the shape parameters d_g, d_r are set manually in order to fix the standard deviation of the two GG hyperpriors to 10^{-3}. For H_2-TV_p-L_q hypermodel, only σ_r and d_r are set in this way, as σ_g is automatically selected based on the RWP.

In Tab. 5.2 we report the obtained quantitative results, i.e. the ISNR and SSIM

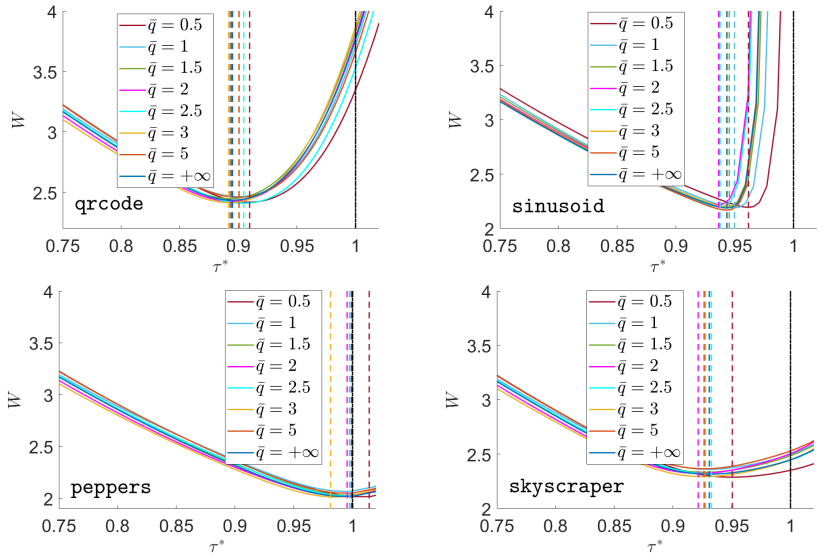

Fig. 5.2: *Tikhonov initialization*. Whiteness measure function W in (4.2) - as function of $\tau(\mu)$ - for the four test images corrupted by blur and additive IIDGN noises of standard deviation $\bar{\sigma}_r = 0.1$ for some different values of the noise shape parameter $\bar{q}$.

$\bar{q}$	0.5	1	2	$+\infty$	0.5	1	2	$+\infty$
	qrcode				**sinusoid**			
ISNR	4.2194	2.1826	2.1643	2.1572	11.0900	10.9092	10.9621	10.8743
SSIM	0.5703	0.5624	0.5604	0.5571	0.9048	0.9033	0.9025	0.9004
$p^{(0)}$	1.2507	1.3102	1.3499	1.4094	3.7667	5.6396	6.0459	5.9667
$q^{(0)}$	0.8352	1.2478	2.0955	4.1485	0.6602	1.0853	1.9730	7.7994
$\sigma_g^{(0)}$	0.1864	0.1846	0.1846	0.1844	0.0417	0.0416	0.0416	0.0395
$\sigma_r^{(0)}$	0.0898	0.0894	0.0896	0.0889	0.0966	0.0943	0.0945	0.0936
	peppers				**skyscraper**			
ISNR	5.0205	4.9249	4.9513	4.9207	2.6645	2.6528	2.6538	2.6828
SSIM	0.6995	0.7036	0.6953	0.6963	0.5072	0.4944	0.4867	0.4854
$p^{(0)}$	1.3486	1.3685	1.3586	1.4279	2.2009	2.3495	2.5775	2.5676
$q^{(0)}$	0.7477	1.1728	1.9980	4.5986	0.8102	1.2103	2.0105	4.3486
$\sigma_g^{(0)}$	0.0425	0.0423	0.0425	0.0419	0.0755	0.0818	0.0818	0.0832
$\sigma_r^{(0)}$	0.1007	0.1001	0.0997	0.0999	0.0935	0.0929	0.0928	0.0921

Table 5.1: *Tikhonov initialization*. Quantitative results for the four test images corrupted by blur and different IIDGN noises with standard deviation $\bar{\sigma}_r = 0.1$.

values of the restored images u^* and the estimated hyperparameters p^*, q^*, σ_g^*, σ_r^* where, for each test, the highest ISNR and SSIM values are in boldface. We also show the (percentage) variations ΔISNR and ΔSSIM with respect to the ISNR and SSIM values achieved after the initialization phase (see Tab. 5.1).

On the piecewise constant `qrcode` and the smooth `sinusoid` test images, the two hypermodels perform similarly well, as evidenced by the reported ISNR and

		H_1-TV_p-L_q				H_2-TV_p-L_q			
	$\bar{q}$	0.5	1	2	$+\infty$	0.5	1	2	$+\infty$
qrcode	ISNR	8.3696	**9.6049**	9.1617	**11.2711**	**9.2474**	9.5596	**9.2441**	11.0704
	SSIM	**0.9568**	0.9361	0.9126	0.9694	0.9308	**0.9379**	**0.9275**	**0.9716**
	ΔISNR	+98%	+340%	+323%	+422%	+119%	+338%	+327%	+413%
	ΔSSIM	+67%	+66%	+63%	+74%	+63%	+66%	+65%	+74%
	p^*	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000
	q^*	0.7000	0.9638	2.0188	14.4154	0.7000	0.9765	1.9719	24.0312
	σ_g^*	0.1752	0.1724	0.1711	0.1726	0.4664	0.1921	0.2500	0.1851
	σ_r^*	0.1015	0.0996	0.0995	0.0996	0.1004	0.0996	0.0990	0.1000
sinusoids	ISNR	14.0339	13.7271	13.8246	13.6854	**15.1811**	**14.3889**	**13.8540**	**14.4425**
	SSIM	**0.9774**	0.9776	**0.9775**	0.9770	0.9768	**0.9782**	0.9774	**0.9807**
	ΔISNR	+26.5%	+25.8%	+26.1%	+25.8%	+36.9%	+31.9%	+26.4%	+32.8%
	ΔSSIM	+8%	+8.2%	+8.3%	+8.5%	+7.9%	8.3%	+8.3%	+8.9%
	p^*	31.5928	15.7289	35.7675	56.6410	8.8413	8.2456	8.2456	9.2384
	q^*	0.7000	0.9783	2.0916	7.6578	0.7000	0.9987	2.0090	11.2241
	σ_g^*	0.0329	0.0300	0.0323	0.0329	0.0380	0.0368	0.0579	0.0350
	σ_r^*	0.1002	0.0980	0.0985	0.0984	0.0995	0.1010	0.0985	0.0985
peppers	ISNR	4.0386	2.3401	2.3229	3.2742	**6.7796**	**5.8839**	**5.5538**	**6.1540**
	SSIM	0.6891	0.6104	0.6058	0.6351	**0.7775**	**0.7334**	**0.7113**	**0.7369**
	ΔISNR	-19.5%	-52.5%	-53.1%	-33.4%	+35%	+19.4%	+12.1%	+25.1%
	ΔSSIM	-1.5%	-13.2%	-12.8%	-8.8%	+11.1%	+4.2%	+2.3%	+5.8%
	p^*	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000
	q^*	0.8879	1.2275	1.7550	3.0738	0.7000	0.9212	1.9719	10.8754
	σ_g^*	0.0403	0.0391	0.0393	0.0393	0.2319	0.2682	0.4899	0.2188
	σ_r^*	0.1104	0.1159	0.1162	0.1108	0.1018	0.1004	0.0997	0.1004
skyscraper	ISNR	1.5536	0.9225	0.9585	1.2245	**4.0255**	**2.9595**	**2.8088**	**3.1076**
	SSIM	0.5787	0.5219	0.5220	0.5450	**0.6876**	**0.6285**	**0.6363**	**0.7094**
	ΔISNR	-41.7%	-65%	-63.8%	-54.3%	+51.1%	+11.5%	+5.8%	+15.6%
	ΔSSIM	+14.1%	+5.5%	+7.2%	+12.3%	+35.6%	+27.1%	+30.7%	+46.1%
	p^*	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000
	q^*	0.7000	1.2181	1.7362	2.9020	0.7000	0.9765	2.0593	162.3645
	σ_g^*	0.0688	0.0757	0.0759	0.0774	0.5727	0.6366	1.2570	1.6093
	σ_r^*	0.1102	0.1132	0.1134	0.1104	0.1002	0.1023	0.0966	0.1014

Table 5.2: *The two variational hypermodels.* Quantitative results for the four test images corrupted by different IIDGN noises with standard deviation $\bar{\sigma}_r = 0.1$.

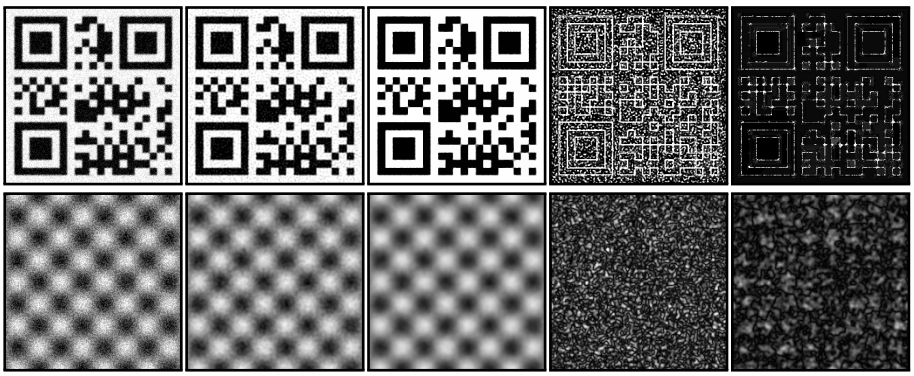

Fig. 5.3: From left to right: data corrupted by IIDGN with $\bar{q} = +\infty$, output of TIK-L_2, restoration by H_1-TV_p-L_q and absolute errors (x10) for `qrcode` and `sinusoid`.

SSIM values. The improvements ΔISNR and ΔSSIM are remarkable on both images and for all noise corruptions, with those on `qrcode` being an order of magnitude larger. This difference is explained by considering that the initial image $u^{(0)}$ is attained by means of a quadratically-regularized model - which is particularly suitable for smooth images like `sinusoid` - hence $u^{(0)}$ for `sinusoid` is already of good quality. Since the two hypermodels perform similarly on `qrcode` and `sinusoid` for all noise types, in Fig. 5.3 we only show some visual results in the case of IIDGN uniform noise ($\bar{q} = +\infty$). More specifically, we report the image $u^{(0)}$ computed by the Tikhonov initialization (second column), the output u^* of the H_1-TV_p-L_q model (third column), and the absolute error images for $u^{(0)}$ and u^* (fourth and fifth columns, respectively), scaled by a factor of 10 to facilitate the visualization. As already indicated by the values reported in Table 5.2, the improvement of u^* with respect to $u^{(0)}$ is particularly significant for the `qrcode` test image.

The similarity in the performance of the two hypermodels does not hold anymore when natural images such as `peppers` and `skyscraper` are processed. In fact, as reported in Tab. 5.2, the ISNR and SSIM for H_1-TV_p-L_q are significantly lower than for H_2-TV_p-L_q, with ISNR values also lower than those achieved after the initialization phase (see the negative ΔISNR values). The estimated noise standard deviations σ_r^* for H_1-TV_p-L_q are in fact significantly larger than the true value $\bar{\sigma}_r$ and thus correspond to over-smoothed restored images u^* - see second rows of Figs. 5.4-5.5, where we show the visual results for `peppers` and `skyscraper`. The estimation errors are also reflected in the histograms of the residual images shown in the third rows of Figs. 5.4-5.5, where the pdfs corresponding to the Tikhonov initialization (dashed magenta lines) are closer to the true ones (solid green lines) when compared to the pdfs resulting from the overall H_1-TV_p-L_q approach (dashed red lines). On the other hand, one can notice from Tab. 5.2 and from the forth and fifth rows of Figs. 5.4-5.5 that the estimated noise standard deviations for H_2-TV_p-L_q result to be closer to the true value $\bar{\sigma}_r = 0.1$ and high-quality restorations are attained.

The observed behavior is motivated by the fact that adopting the H_2-TV_p-L_q hypermodel can be equivalently interpreted as a way to set a hyperprior on the regularization parameter μ in the TV_p-L_q model (1.3)-(1.5), with μ expressed as in (4.5). This guarantees a more direct monitoring of the balancing between the regularization and fidelity terms. We also remark that the mentioned balancing is particularly delicate when the selected regularizer is not expected to be flexible enough to model the actual image properties, as in the case of `peppers` and `skyscrapers` test image that would certainly benefit from using a more sophisticated regularizer.

As a result of this comparison, we can conclude that the H_2-TV_p-L_q hypermodel is the best performing one and, hence, hereinafter we focus on it.

In Fig. 5.6 we provide some evidences for the good convergence behaviour of the overall alternating minimization approach outlined in Alg. 1 for the H_2-TV_p-L_q hypermodel. In particular, for `qrcode` image corrupted by blur and additive IIDGN noise with $\bar{q} = 2$, $\bar{\sigma}_r = 0.1$, we monitor the hyperparameter vector $\theta^{(k)}$ and the restoration quality measures $\mathrm{ISNR}\big(u^{(k)}\big)$, $\mathrm{SSIM}\big(u^{(k)}\big)$ along the outer iterations of Alg. 1. The reported values for iteration index $k = 0$ correspond to the outputs $u^{(0)}$, $\theta^{(0)}$ of the Tikhonov initialization. We note that all the monitored quantities stabilize after very

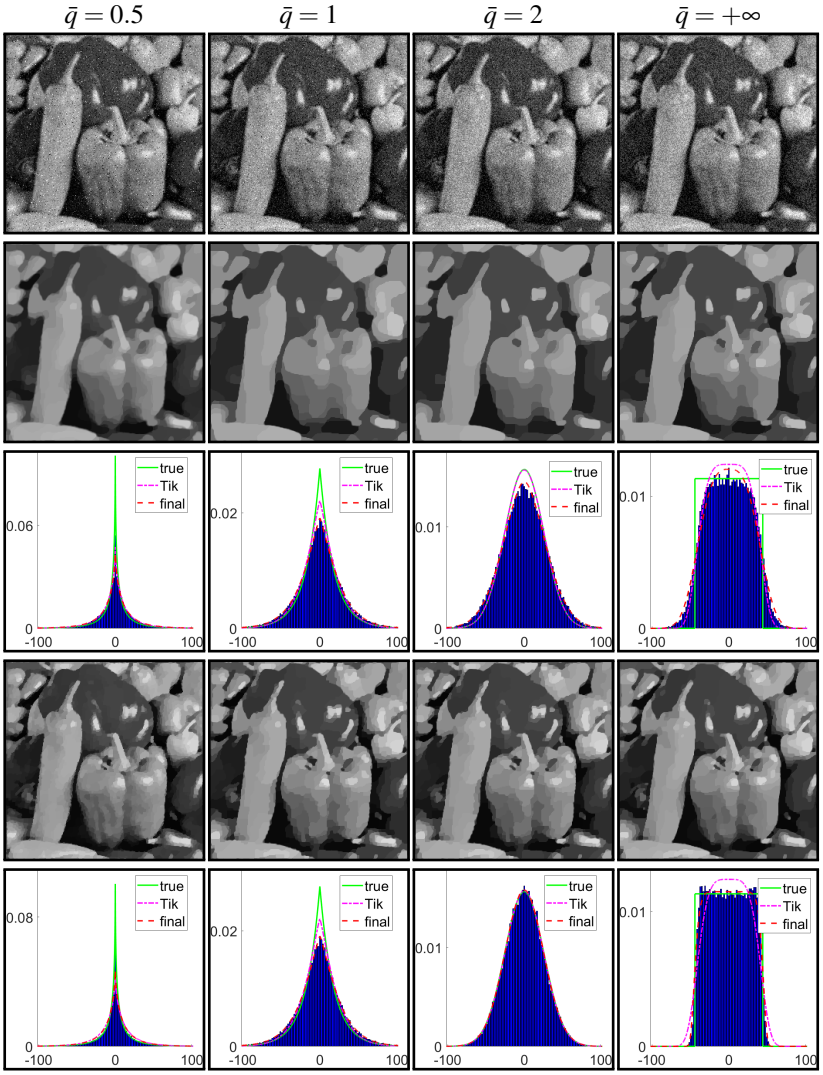

Fig. 5.4: *Hypermodels on* `peppers`. From top to bottom: corrupted images, restored images by H_1-TV_p-L_q and H_2-TV_p-L_q, and their residual histograms.

few iterations, even in this experiment where not only the cost function J is jointly non-convex in u and the hyperparameters but also its restriction to u (minimized by ADMM) becomes non-convex after the first outer iteration - in fact, $p^{(k)} < 1$ for $k \geq 1$.

5.3 Performance evaluation of the H_2-TV_p-L_q hypermodel

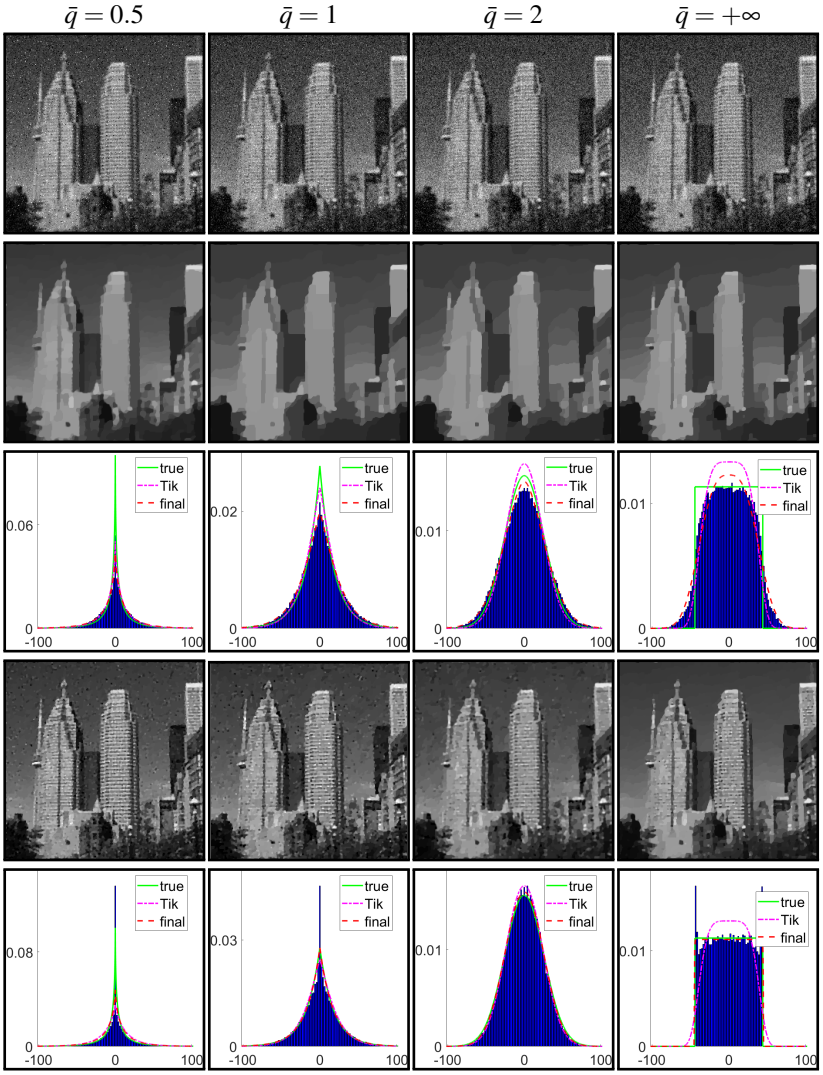

Fig. 5.5: *Hypermodels on* `skyscraper`. From top to bottom: corrupted images, restored images by H_1-TV_p-L_q and H_2-TV_p-L_q, and their residual histograms.

The key novelty of the proposed method relies on the illustrated hierarchical Bayesian framework which allows to automatically pick up and use for restoration one variational model among the infinity contained in the considered TV_p-L_q class, parametrized by the two shape parameters p,q and the regularization parameter μ. This leads to a completely unsupervised restoration approach which, as evidenced by the results reported in previous section for the H_2-TV_p-L_q hypermodel, is capable of achieving good-quality restorations for a large class of images and of noise corruptions

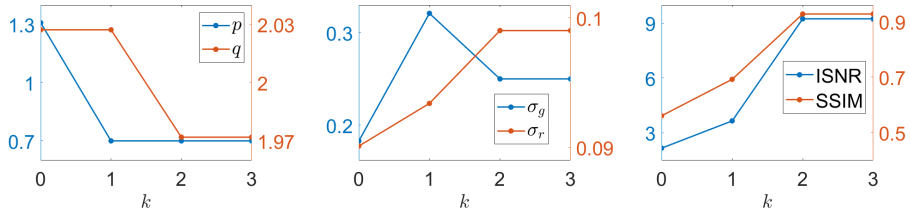

Fig. 5.6: From left to right: trajectories of $p^{(k)}$, $q^{(k)}$, of $\sigma_g^{(k)}$, $\sigma_r^{(k)}$, accuracy measures $\mathrm{ISNR}\left(u^{(k)}\right)$ and $\mathrm{SSIM}\left(u^{(k)}\right)$, along the outer iterations of the AM scheme in Alg. 1.

To the best of author's knowledge, there not exist in literature other approaches of this type for the automatic selection of the three hyperparameters of the TV_p-L_q class of models. Furthermore, the proposed hierarchical framework could be used for other classes of parametrized variational models as well, also larger than TV_p-L_q or even containing TV_p-L_q as a subset. Hence, we think it is not meaningful here to compare the results reported in previous section with those obtainable by using regularizers not belonging to the TV_p class or, even more, fidelity terms suitable for noises other than the considered additive IIDGN class. Instead, we believe it is of crucial importance in order to validate the proposed automatic framework and evaluate its potential practical appeal, to compare its accuracy with the best results achievable by the whole class of TV_p-L_q models and, in particular, by the most popular members of the TV_p-L_q class, such as the TV-L_2, TV-L_1,TIK-L_2 and TIK-L_1 models.

To this aim, we carry out the following experiment, whose quantitative results are reported in Tab. 5.3. We consider the most severe test `skyscraper` corrupted by the same blur as in previous sections and by additive IIDGN noise with standard deviation $\bar{\sigma}_r = 0.1$ and shape parameter $\bar{q} = 3$. We run our fully automatic restoration H_2-TV_p-L_q hypermodel and report the obtained ISNR and SSIM accuracy results in the first row of Tab. 5.3. Then, we compare these results with those achievable by a set of non-automatic meaningful TV_p-L_q models, with a priori fixed p and q values reported in the second and third column of Tab. 5.3. In particular, for each of these tests, the regularization parameter μ of the TV_p-L_q model has been set manually so as to achieve the highest ISNR value (shown in the fourth column of Tab. 5.3, with the associated SSIM in the fifth column) and the highest SSIM value (in the sixth column, with the associated ISNR in the last column). For a fair comparison, the considered non-automatic TV_p-L_q models have been solved via the ADMM algorithm outlined in Sec. 4.2.1, using the same initial iterate (image $u^{(0)}$ output of the proposed Tikhonov initialization) and stopping criterion ($\delta_u^{(k)} < 10^{-5}$) as our hypermodel.

In the second row of Tab. 5.3 we report the results of using the final shape parameter estimates $p^* = 0.7$ and $q^* = 3.08$ provided by our hypermodel. The $\mathrm{ISNR}_{\max}$ and $\mathrm{SSIM}_{\max}$ values are slightly higher than the hypermodel but the associated SSIM and ISNR values are slightly lower. In the third row we consider the target shape parameter values $\bar{p} = 0.7$ and $\bar{q} = 3$ and the obtained results are very little worse than using p^* and q^*. In the last six rows we show the results for six very popular members of the TV_p-L_q class corresponding to $p = 1, 2$, $q = 1, 2, +\infty$. All the six non-automatic mod-

`skyscraper`: $\bar{p} = 0.7$, $\bar{q} = 3$

model	p	q	ISNR$_{max}$	SSIM	ISNR	SSIM$_{max}$
H_2-TV$_p$-L$_q$	//	//	2.9321	//	//	0.6386
TV$_{p^*}$-L$_{q^*}$	0.7	3.08	2.9821	0.6287	2.7952	0.6538
TV$_{\bar{p}}$-L$_{\bar{q}}$	0.7	3	2.9777	0.6151	2.7917	0.6537
TV$_{\hat{p}}$-L$_{\hat{q}}$	0.7	3	2.9777	0.6151	2.7917	0.6537
TV-L$_1$	1	1	1.9887	0.5262	1.8163	0.5785
TV-L$_2$	1	2	2.8355	0.5778	2.3678	0.6279
TV-L$_\infty$	1	∞	1.9851	0.5501	1.7835	0.5675
TIK-L$_1$	2	1	1.9088	0.4584	1.3592	0.4997
TIK-L$_2$	2	2	2.6508	0.4884	1.7267	0.5415
TIK-L$_\infty$	2	∞	1.8932	0.5256	1.64334	0.5504

Table 5.3: *Quantitative performance evaluation of the H_2-TV$_p$-L$_q$ hypermodel.*

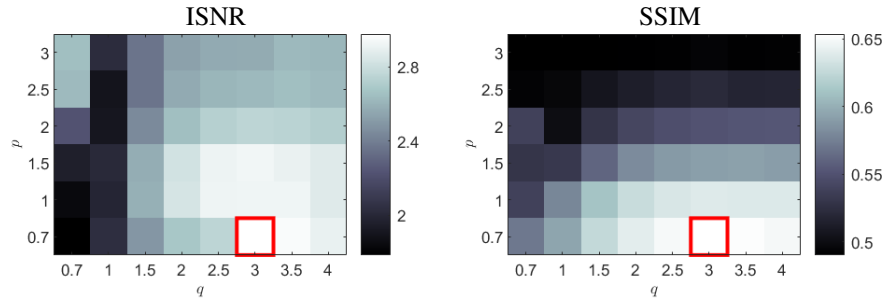

Fig. 5.7: ISNR and SSIM values achieved by TV$_p$-L$_q$ model for different (p,q) values. The largest ISNR and SSIM in the red box are attained at $p = 0.7$ and $q = 3$.

els perform worse than our automatic hypermodel. Finally, the results reported in the fourth row of Tab. 5.3 are the most meaningful as they clearly validate the proposed hierarchical Bayesian approach. The $\hat{p}$ and $\hat{q}$ values are in fact those providing the best results among the TV$_p$-L$_q$ class. To obtain $\hat{p}$ and $\hat{q}$, we ran the TV$_p$-L$_q$ model for a grid of different (p,q) values and then selected the pair $(\hat{p},\hat{q})$ yielding the highest ISNR and the highest SSIM. The results of this experiment are shown in Fig. 5.7.

6 Conclusions

The present article discusses the introduction of a probabilistic hierarchical approach for the fully automatic solution of the image restoration inverse problem when both the noise and the image gradient magnitudes are assumed to follow a GN distribution. The joint estimation of the original image and of the unknown parameters determining the prior and the likelihood pdfs relies on the introduction of two classes of hyperpriors. The resulting hypermodels are addressed by means of an alternating minimization scheme, for which a robust initialization based on the noise whiteness

property is provided. The proposed machinery leads to a completely unsupervised restoration approach which is capable of achieving good-quality restorations for a large class of images and of noise corruptions. As future work, more sophisticated regularizers enforcing the robustness of the algorithm in presence of high level of corruptions can be explored. In addition, the capability of the proposed framework to effectively address the case of mixed noise can be also investigated.

Acknowledgments Research was supported by the "National Group for Scientific Computation (GNCS-INDAM)" and by ex60 project by the University of Bologna "Funds for selected research topics".

References

1. Abramowitz, M. *Handbook of Mathematical Functions, With Formulas, Graphs, and Mathematical Tables*. 1974.
2. Almeida, M. S. C. and Figueiredo, M. A. T. *Parameter estimation for blind and non-blind deblurring using residual whiteness measures*. IEEE T. Image Process., 22: 2751-2763, 2013.
3. Babacan, S. D., Molina, R. and Katsaggelos, A. *Generalized Gaussian Markov random field image restoration using variational distribution approximation*. 2008 IEEE International Conference on Acoustics, Speech and Signal Processing:1265–1268, 2008.
4. Babacan, S. D., Molina, R. and Katsaggelos, A. *Parameter Estimation in TV Image Restoration Using Variational Distribution Approximation*. IEEE Transactions on Image Processing, 17: 326-339 , 2008.
5. Bauer, F. and Lukas, M. A. *Comparing parameter choice methods for regularization of ill-posed problem*. Math. Comput. Simulation, 81: 1795–1841, 2011.
6. Fenu, C, Reichel, L., Rodriguez, G. and Sadok, H. *GCV for Tikhonov regularization by partial SVD*. BIT, 57: 1019–1039, 2017.
7. Calatroni, L., Lanza, A., Pragliola, M. and Sgallari, F. *A Flexible Space-Variant Anisotropic Regularization for Image Restoration with Automated Parameter Selection*. SIAM J. Imaging Sci., 12(2): 1001-1037, 2019.
8. Calatroni, L., Lanza, A., Pragliola, M. and Sgallari, F. *Space-Adaptive Anisotropic Bivariate Laplacian Regularization for Image Restoration*. Tavares J., Natal Jorge R. (eds) VipIMAGE 2019. Lect. Notes Comput. Vis. Biomech., 34, 2019.
9. Calatroni, L., Lanza, A., Pragliola, M. and Sgallari, F. *Adaptive parameter selection for weighted-TV image reconstruction problems*. J. Phys. Conf. Ser., 1476, 2020.
10. Calvetti, D., Hakula, H., Pursiainen, S., and Somersalo, E. *Conditionally Gaussian Hypermodels for Cerebral Source Localization*. SIAM J. Imaging Sci.,2: 879–909, 2009.
11. Calvetti, D., Hansen, P. C. and Reichel, L. *L-curve curvature bounds via Lanczos bidiagonalization*. Electron. Trans. Numer. Anal., 14: 20-35, 2002.
12. Calvetti, D., Pascarella, A., Pitolli, F., Somersalo, E. and Vantaggi, B. *A hierarchical Krylov-Bayes iterative inverse solver for MEG with physiological preconditioning*. Inverse Problems, 31: 125005, 2015.
13. 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 36:025010, 2020.
14. Calvetti, D. and Somersalo, E. *Introduction to Bayesian Scientific Computing: Ten Lectures on Subjective Computing (Surveys and Tutorials in the Applied Math. Sciences)*. Springer-Verlag, 2007.
15. Calvetti, D., Somersalo, E., Strang, A. *Hierachical Bayesian models and sparsity: ℓ_2 -magic*. Inverse Problems 35:035003, 2019.
16. Chan, R. H., Dong, Y. and Hintermüller, M. *An efficient two-phase L_1-TV method for restoring blurred images with impulse noise*. IEEE T. Image Process., 65(5): 1817–1837, 2005.
17. Clason, C. *L∞ fitting for inverse problems with uniform noise*. Inverse Problems, 28(10), 2012.
18. Godefroy, M. B., *La fonction Gamma: theorie, histoire, bibliographie*. Cornell Univer. Library, 1901.
19. De Bortoli, V., Durmus, A., Pereyra, M., Vidal, A. F. *Maximum Likelihood Estimation of Regularization Parameters in High-Dimensional Inverse Problems: An Empirical Bayesian Approach Part II: Theoretical Analysis*. SIAM J. Imaging Sci. 13: 1990-2028, 2020.

20. Guo, X., Li, F. and Ng, M. K. *A Fast ℓ_1-TV Algorithm for Image Restoration*. SIAM J. Sci. Comput., 31(3): 2322-2341, 2009.
21. He, C., Hu, C., Zhang, W. Shi, B. *A Fast Adaptive Parameter Estimation for Total Variation Image Restoration*. IEEE Trans. on Image Process., 23(12): 4954-4967, 2014.
22. Lanza, A., Morigi, S., Pragliola, M. and Sgallari, F. *Space-variant generalised Gaussian regularisation for image restoration*. Comput. Method. Biomec., 7: 490-503, 2019.
23. Lanza, A., Morigi, S., and Sgallari, F. *Constrained TV_p-ℓ_2 Model for Image Restoration*. J. Sci. Comput., 68: 64-91, 2016.
24. Lanza, A., Pragliola, M. and Sgallari, F. *Residual whiteness principle for parameter-free image restoration*. Electron. T. Numer. Ana., 53: 329:351, 2020.
25. Molina, R., Katsaggelos, A. and Mateos, J. *Bayesian and regularization methods for hyperparameter estimation in image restoration*. IEEE transactions on image processing : a publication of the IEEE Signal Processing Society, 8: 231–246, 1999.
26. Reichel, L. and Rodriguez, G. *Old and new parameter choice rules for discrete ill-posed problems*. Numer. Algorithms, 63: 65-87, 2013.
27. Rudin, L. I., Osher, S. and Fatemi, E. *Nonlinear Total Variation Based Noise Removal Algorithms*. Phys. D, 60: 259-268, 1992.
28. Stuart, A. M. *Inverse problems: A Bayesian perspective*. Acta Numer., 19: 451-559, 2010.
29. Vidal, A. F., De Bortoli, V. Pereyra, M., Durmus, A. *Maximum Likelihood Estimation of Regularization Parameters in High-Dimensional Inverse Problems: An Empirical Bayesian Approach Part I: Methodology and Experiments*. SIAM J. Imaging Sci. 13: 1945-1989, 2020.
30. Wang, Y., Yin, W. and Zeng, J. *Global Convergence of ADMM in Nonconvex Nonsmooth Optimization*. J. Sci. Comput., 78: 29-63, 2019.
31. Wang, Z., Bovik, A. C., Sheikh, H. R. and Simoncelli, E. P. *Image Quality Assessment: From Error Visibility to Structural Similarity*. IEEE T. Image Process., 4: 600-612, 2004.
32. Wen, Y. and Chan, R. H. *Parameter selection for total-variation-based image restoration using discrepancy principle*. IEEE T. Image Process., 21(4): 1770-1781, 2012.
33. Wen, Y. and Chan, R. H. *Using generalized cross validation to select regularization parameter for Total Variation regularization problems*. Inverse Probl. Imaging, 12: 1103-1120, 2018.
34. Park, Y., Reichel, L., Rodriguez, G. and Yu, X. *Parameter determination for Tikhonov regularization problems in general form*. J. Comput. Appl. Math., 343: 12-25, 2018.
35. Robert, C. *The Bayesian choice : from decision-theoretic foundations to computational implementation*. Springer, 2007.
36. Zuo, W., Meng, D., Zhang, L., Feng X. and Zhang, D. *A Generalized Iterated Shrinkage Algorithm for Non-convex Sparse Coding*. IEEE Int Conf Comput Vis.: 217-224, 2013.

A Proofs of the results

Proposition A.1 *Let $s,\gamma \in \mathbb{R}_{++}$, $z \in \mathbb{N}$ be given constants, let $f_s : \mathbb{R}^z \to \mathbb{R}$ be the (parametric, not necessarily convex) function defined by $f_s(x) := \|x\|_2^s$ and let $\operatorname{prox}_{f_s}^{\gamma} : \mathbb{R}^z \rightrightarrows \mathbb{R}^z$ be the proximal operator of function f_s with proximity parameter γ, defined as the z-dimensional minimization problem*

$$x^* \in \operatorname{prox}_{f_s}^{\gamma}(y) := \underset{x\in\mathbb{R}^z}{\arg\min} \left\{ \|x\|_2^s + \frac{\gamma}{2}\|x-y\|_2^2 \right\}, \quad y \in \mathbb{R}^z .$$

Then, for any $z \in \mathbb{N}$, $s,\gamma \in \mathbb{R}_{++}$, $y \in \mathbb{R}^z$, $\operatorname{prox}_{f_s}^{\gamma}$ takes the form of a shrinkage operator:

$$\operatorname{prox}_{f_s}^{\gamma}(y) = \xi^*\big(s, \rho\,(s,\gamma,\|y\|_2)\big)\, y, \quad \text{with } \xi^* : \mathbb{R}_{++} \times \mathbb{R}_+ \rightrightarrows [0,1),$$

and with function

$$\rho : \mathbb{R}^2_{++} \times \mathbb{R}_+ \to \mathbb{R}_+ \;\; defined\ by \;\; \rho(s,\gamma,\|y\|_2) = \begin{cases} (\gamma/s)\|y\|_2^{2-s} & \text{if } \|y\|_2 > 0, \\ 0 & \text{if } \|y\|_2 = 0, \end{cases}$$

In particular, for any $z \in \mathbb{N}$, the shrinkage coefficient function $\xi^*(s,\rho)$ satisfies:

$$\begin{array}{ll} s>0, \rho=0 & \Longrightarrow \xi^*(s,\rho) = 0, \\ s\le 1, \rho < \bar\rho(s) & \Longrightarrow \xi^*(s,\rho) = 0, \\ s\le 1, \rho = \bar\rho(s) & \Longrightarrow \xi^*(s,\rho) \in \{0, \bar\xi(s)\}, \\ s\le 1, \rho > \bar\rho(s) & \Longrightarrow \xi^*(s,\rho) = \textit{unique sol. in } (\bar\xi(s),1) \textit{ of } h(\xi;s,\rho) = 0, \\ s>1, \rho>0 & \Longrightarrow \xi^*(s,\rho) = \textit{unique sol. in } \quad (0,1) \textit{ of } h(\xi;s,\rho) = 0, \end{array} \tag{A.1}$$

with functions $\bar{\rho}:(0,1]\to[1,\infty)$, $\bar{\xi}:(0,1]\to[0,1)$, *and* $h:(0,1]\to\mathbb{R}$ *defined by*

$$\bar{\rho}(s)=\frac{(2-s)^{2-s}}{s\,(2-2s)^{1-s}}\,,\quad \bar{\xi}(s)=2\,\frac{1-s}{2-s}\,,\quad h(\xi;s,\rho)=\xi^{s-1}+\rho\,(\xi-1)\,. \tag{A.2}$$

Finally, the function ξ^* *exhibits the following regularity and monotonicity properties:*

$$\begin{gathered}\xi^*\in C^\infty(S),\ \ S=S_0\cup S_1,\ \ S_0=(0,1]\times(\bar{\rho}(s),\infty),\ S_1=(1,\infty)\times\mathbb{R}_{++},\\ \frac{\partial\xi^*}{\partial s}(s,\rho)>0\,,\quad \frac{\partial\xi^*}{\partial\rho}(s,\rho)>0\,,\quad \forall(s,\rho)\in S\,.\end{gathered} \tag{A.3}$$

Proof The proof of statement (A.1)-(A.2) comes straightforwardly from Proposition 1 in [23]. In fact, (A.1)-(A.2) is there demonstrated for $s<2$, but the proof of case c) in that proposition can be seamlessly extended to cover the case $s\geq 2$.

To prove (A.3), first we notice from (A.1)-(A.2) that $\xi^*(s,\rho)\in(0,1)\ \forall(s,\rho)\in S$ and introduce the function $f\colon S_f\to\mathbb{R},\ S_f=(0,1)\times S\subset\mathbb{R}^3_{++}$, defined by $f(\xi,s,\rho):=h(\xi;s,\rho)=\xi^{s-1}+\rho\,(\xi-1)$, $\forall(\xi,s,\rho)\in S_f$. The function f clearly satisfies

$$f\in C^\infty(S_f)\,,\quad \frac{\partial f}{\partial\xi}(\xi,s,\rho)=(s-1)\,\xi^{s-2}+\rho=0\ \text{ for } s<1\wedge\xi^{s-2}=\rho/(1-s).$$

We now demonstrate by contradiction that that if $s<1$ then $\xi^{s-2}\neq\rho/(1-s)$ for any $(\xi,s,\rho)\in S_f$. In fact, according to the definition of set S_f given above, we have that $\xi>\bar{\xi}(s)$ and $\rho>\bar{\rho}(s)$ for $s<1$, with functions $\bar{\xi},\bar{\rho}$ defined in (A.2). But we have

$$\xi>\bar{\xi}(s)\iff\xi^{s-2}<\left(\bar{\xi}(s)\right)^{s-2}\iff\frac{\rho}{1-s}<\left(2\,\frac{1-s}{2-s}\right)^{s-2}\iff\rho<\frac{s}{2}\,\bar{\rho}(s)\,.$$

Hence, $\partial f/\partial\xi\neq0\ \forall(\xi,s,\rho)\in S_f$ and it follows form the implicit function theorem that, for any $(s,\rho)\in S$, the shrinkage coefficient $\xi^*\in(0,1)$ is given by the infinitely many times differentiable function of (s,ρ), denoted $\xi^*(s,\rho)$, solution of equation

$$f(\xi^*(s,\rho),s,\rho)=0\iff(\xi^*(s,\rho))^{s-1}+\rho\,(\xi^*(s,\rho)-1)=0\,. \tag{A.4}$$

Taking the partial derivatives of both sides of (A.4) with respect to s and ρ, and recalling that for any function $c(x)=w(x)^{a(x)}$ with $w(x)>0\ \forall x\in\mathbb{R}$, it holds that

$$c'(x)=c(x)\left[a'(x)\ln w(x)+a(x)w'(x)/w(x)\right],$$

after simple manipulations we have

$$\frac{\partial}{\partial s}f(\xi^*(s,\rho),s,\rho)=0\iff\frac{\partial\xi^*}{\partial s}(s,\rho)=\frac{-(\xi^*(s,\rho))^{s-1}\ln\xi^*(s,\rho)}{(s-1)\,(\xi^*(s,\rho))^{s-2}+\rho}\,, \tag{A.5}$$

$$\frac{\partial}{\partial\rho}f(\xi^*(s,\rho),s,\rho)=0\iff\frac{\partial\xi^*}{\partial\rho}(s,\rho)=\frac{1-\xi^*(s,\rho)}{(s-1)\,(\xi^*(s,\rho))^{s-2}+\rho}\,. \tag{A.6}$$

Since $(s,\rho)\in S\Longrightarrow\rho>0$, $\xi^*(s,\rho)\in(0,1)$, then both the numerators in the definitions of $\partial\xi^*/\partial s$ and $\partial\xi^*/\partial\rho$ in (A.5)-(A.6) are positive quantities. The denominator in (A.5)-(A.6) is also clearly positive for $s\geq1$; for $s<1$ it is positive for

$$(\xi^*(s,\rho))^{2-s}>(1-s)/\rho\Longleftarrow(\xi^*(s,\rho))^{2-s}>(1-s)/\bar{\rho}(s)=(s/2)\left(\bar{\xi}(s)\right)^{2-s},$$

which is always verified since $\xi^*(s,\rho)>\bar{\xi}(s)$ for $s<1$. This proves (A.3).□

Corollary A.1 *Under the setting of Proposition A.1, let $(s,\rho)\in S$. Then, the Newton-Raphson iterative scheme applied to the solution of $h(\xi;s,\rho)=0$, namely*

$$\xi_{k+1} = \xi_k - \frac{h(\xi_k;s,\rho)}{h'(\xi_k;s,\rho)} = \frac{\rho+(s-2)\,\xi_k^{s-1}}{\rho+(s-1)\,\xi_k^{s-2}}, \quad k=1,2,\ldots, \tag{A.7}$$

converges to $\xi^*(s,\rho)$ if the initial iterate ξ_0 is chosen, dependently on s,ρ, as follows:

$$\begin{array}{llll} s\le 1, & \rho>\bar\rho(s) & \Longrightarrow \xi_0 \in \left[\bar\xi(s),1\right], \\ s\in(1,2), & \rho>2-s & \Longrightarrow \xi_0 \in (0,1], \\ s\in(1,2), & \rho\le 2-s & \Longrightarrow \xi_0 \in \left(0,\xi^*(s,\rho)\right], \\ s\ge 2 & \rho>0 & \Longrightarrow \xi_0 \in [0,1]. \end{array} \tag{A.8}$$

Hence, based on (A.3), ξ_0 for computing $\xi^*(s,\rho)$ by (A.7) can be chosen among the solutions $\xi^*(\bar s,\rho)$ for different s values according to the following strategy:

$$\begin{array}{llll} s\le 1, & \rho>\bar\rho(s) & \Longrightarrow \xi_0 = \xi^*(\bar s,\rho), & \bar s\ge s, \\ s\in(1,2), & \rho>2-s & \Longrightarrow \xi_0 = \xi^*(\bar s,\rho), & \bar s>1, \\ s\in(1,2), & \rho\le 2-s & \Longrightarrow \xi_0 = \xi^*(\bar s,\rho), & \bar s\in(1,s], \\ s\ge 2 & \rho>0 & \Longrightarrow \xi_0 = \xi^*(\bar s,\rho), & \bar s>0. \end{array} \tag{A.9}$$

Proof According to Proposition A.1, for any given pair $(s,\rho)\in S$, the shrinkage coefficient $\xi^*(s,\rho)$ is given by the unique root of nonlinear equation $h(\xi;s,\rho)=0$ in the open interval $(\bar\xi(s),1)$ for $s\le 1$, $(0,1)$ for $s>1$. Convergence of the Newton-Raphson method applied to finding such roots depends on the initial guess as well as on the first- and second-order derivatives of the function h, which read

$$h'(\xi;s,\rho) = (s-1)\,\xi^{s-2}+\rho, \quad h''(\xi;s,\rho) = (s-1)(s-2)\,\xi^{s-3}.$$

In particular, it is immediate to verify that $h\in C^\infty((0,1])$ for any $(s,\rho)\in S$ and that, depending on s, the function h and its derivatives h',h'' satisfy

$$\begin{cases} s<1 \\ \rho>\bar\rho(s) \end{cases} \Rightarrow \begin{cases} h\in C^\infty\left(\left[\bar\xi(s),1\right]\right),\ h\left(\bar\xi(s)\right)=\frac{s}{2-s}\left(\bar\rho(s)-\rho\right)<0,\ h(1)=1, \\ h'\left(\bar\xi(s)\right)=\rho+\frac{s}{2}\bar\rho(s),\ h'(1)=\rho+s-1,\ h'(\xi),h''(\xi)>0\ \forall\xi\in[\bar\xi(s),1], \end{cases}$$

$$\begin{cases} s\in(1,2) \\ \rho>0 \end{cases} \Rightarrow \begin{cases} h\in C^0\left([0,1]\right)\cap C^\infty\left((0,1]\right),\ h(0)=-\rho<0,\ h(1)=1,\ h'(0^+)=+\infty, \\ h'(1)=\rho+s-1,\ h'(\xi)>0\ \forall\xi\in(0,1],\ h''(\xi)<0\ \forall\xi\in(0,1], \end{cases}$$

$$\begin{cases} s>2 \\ \rho>0 \end{cases} \Rightarrow \begin{cases} h\in C^1\left([0,1]\right)\cap C^\infty\left((0,1]\right),\ h(0)=-\rho<0,\ h(1)=1,\ h'(0)=\rho, \\ h'(1)=\rho+s-1,\ h'(\xi)>0\ \forall\xi\in[0,1],\ h''(\xi)>0\ \forall\xi\in(0,1], \end{cases}$$

Properties of h,h',h'' for $s<1\wedge\rho>\bar\rho(s)$ indicate that in this case the iterative scheme (A.7) is guaranteed to converge to the unique root $\xi^*(s,\rho)$ of $h(\xi;s,\rho)=0$ in the open interval $(\bar\xi(s),1)$ if $\xi_0\in\left[\xi^*(s,\rho),1\right]$. But, since it can be proved (we omit the proof for shortness) that setting $\xi_0=\bar\xi(s)$ the first iteration of (A.7) yields $\xi_1\in\left[\xi^*(s,\rho),1\right)$, then (A.7) converges under the milder condition $\xi_0\in[\bar\xi(s),1]$.

Properties of h,h',h'' for $s\in(1,2)\wedge\rho>0$ lead to the convergence condition $\xi_0\in\left(0,\xi^*(s,\rho)\right]$ for (A.7), with $\xi_0=0$ excluded since $h'(0^+)=+\infty\Longrightarrow\xi_k=0\ \forall k$. It can be proved that setting $\xi_0=1$, then (A.7) yields $\xi_1\in\left(0,\xi^*(s,\rho)\right]$ if and only if $\rho>2-s$. Hence, for $\rho>2-s$ we have the milder convergence condition $\xi_0\in(0,1]$.

Properties of h,h',h'' for $s>2\wedge\rho>0$ indicate that (A.7) is guaranteed to converge to the desired $\xi^*(s,\rho)$ if $\xi_0\in\left[\xi^*(s,\rho),1\right]$. However, as it is easy to prove that $\xi_0=0$ in (A.7) yields $\xi_1=1$, then (A.7) converges for any $\xi_0\in[0,1]$ in this case.

This completes the proof of (A.8), whereas (A.9) comes easily from (A.8) and from the monotonicity properties of function h given in (A.3). □

Proof of Proposition 4.2 Based on Proposition 2.1, the function T is infinitely differentiable. Hence, we impose a first order optimality condition on T with respect to σ:

$$\frac{\partial T}{\partial\sigma} = (n-d+1)/\sigma - s(\phi(s))^{-s}\|x\|_s^s(1/\sigma)^{s+1} + (d-1)m^{-s}\sigma^{s-1} = 0. \tag{A.10}$$

Equation (A.10) admits the following closed-form solution:

$$\sigma^{\mathrm{MAP}}(s) = m\left[\left(-\alpha+\sqrt{\alpha^2+4\beta m^{-s}(\sigma^{\mathrm{ML}})^s}\right)/(2\beta)\right]^{1/s}, \tag{A.11}$$

where the expression of $\sigma^{\mathrm{ML}}(s)$, for a generic $s\in\mathbb{R}_{++}$, is given in (2.7). Note that (A.11) can be further manipulated so as to give

$$\sigma^{\mathrm{MAP}}(s) = m\alpha^{1/s}[(-1+\sqrt{1+c(s)})/(2\beta)]^{1/s}, \quad \text{with} \quad c(s) := 4(\beta/\alpha^2)(m\phi(s))^{-s}(s/n)\|x\|_s^s. \tag{A.12}$$

It is easy to verify that the second derivative of T with respect to σ computed at $\sigma^{\mathrm{MAP}}(s)$ is strictly positive, hence the stationary point in (A.10) is a minimum. Finally, plugging (A.12) into the expression of function T in (4.13), after a few simple manipulations, the s estimation problem takes the form:

$$s^{\mathrm{MAP}} \in \arg\min_{s\in B}\Big\{f^{\mathrm{MAP}}(s) := \underbrace{n\ln[\Gamma(1+1/s)\phi(s)]+\ln[(1/d)\Gamma(1+d/s)]+(d/s)\ln(s/(n\beta))}_{:=f_1(s)} + \underbrace{(n\alpha/s)\big(\sqrt{1+c(s)}+\ln(\alpha(-1+\sqrt{1+c(s)})/(2\beta))\big)}_{:=f_2(s)}\Big\}$$

Based on Proposition 2.1, it is easy to verify that $f^{\mathrm{MAP}}\in C^\infty(B)$ and that the following limits hold:

$$\lim_{s\to+\infty} f_1(s) = n\ln\sqrt{3}-\ln d\in\mathbb{R}, \quad \lim_{s\to+\infty}\|x\|_s/(m\phi(s)) = \|x\|_\infty/(m\sqrt{3}) =: \nu.$$

Depending on ν, the three following scenarios arise:

(a) $\nu = 1$. Based on the properties of function ϕ and $\|\cdot\|_\infty$, ν tends to 1 from the right. If $\nu\to 1^+$ slower than $s\to+\infty$, case (a) leads to case (c), otherwise, $c(s)\sim O(s)$ and $f_2(s)$ vanish so that $f^{\mathrm{MAP}}(s)$ tends to a finite value when s goes to $+\infty$.

(b) $\nu < 1$. For large s, the first term in $f_2(s)$ vanishes, while for the logarithmic term we consider the Maclaurin series expansion of $\sqrt{1+c(s)}$, so that:

$$\begin{aligned}\lim_{s\to+\infty} f_2(s) &= \lim_{s\to+\infty}(n\alpha/s)[\ln(\alpha/(2\beta))+\ln(-1+1+(1/2)c(s)+o(c(s)))]\\ &= \lim_{s\to+\infty}(n\alpha/s)\ln((1/2)c(s)) = \lim_{s\to+\infty}(n\alpha/s)\ln(2(\beta/\alpha^2)(s/n))+(n\alpha)\ln\nu = n\alpha\ln\nu\in\mathbb{R}.\end{aligned}$$

Therefore, function $f^{\mathrm{MAP}}(s)$ tends to a finite value when $s\to+\infty$.

(c) $\nu > 1$. For large s, we have $(n\alpha/s)\ln(-1+\sqrt{1+c(s)})\sim(n\alpha/s)\ln\sqrt{c(s)}\to k\in\mathbb{R}_+$ as $s\to+\infty$. Hence:

$$\lim_{s\to+\infty} f_2(s) = \lim_{s\to+\infty}(n\alpha/s)\sqrt{1+c(s)}+k = +\infty \longrightarrow \lim_{s\to+\infty} f^{\mathrm{MAP}}(s) = +\infty.$$

We conclude that f^{MAP} admits either a finite minimizer, that can be plugged into (A.11) thus returning σ^{MAP}, or $s^{\mathrm{MAP}} = +\infty$, i.e. the samples in x are drawn from a uniform distribution with $\sigma^{\mathrm{MAP}} = m\nu$. □