Image restoration with generalized Gaussian mixture model patch priors

Charles-Alban Deledalle, Shibin Parameswaran, Truong Q. Nguyen

Image restoration with generalized Gaussian mixture model patch priors

Charles-Alban Deledalle*†, Shibin Parameswaran†, and Truong Q. Nguyen†

Abstract. Patch priors have became an important component of image restoration. A powerful approach in this category of restoration algorithms is the popular Expected Patch Log-likelihood (EPLL) algorithm. EPLL uses a Gaussian mixture model (GMM) prior learned on clean image patches as a way to regularize degraded patches. In this paper, we show that a generalized Gaussian mixture model (GGMM) captures the underlying distribution of patches better than a GMM. Even though GGMM is a powerful prior to combine with EPLL, the non-Gaussianity of its components presents major challenges to be applied to a computationally intensive process of image restoration. Specifically, each patch has to undergo a patch classification step and a shrinkage step. These two steps can be efficiently solved with a GMM prior but are computationally impractical when using a GGMM prior. In this paper, we provide approximations and computational recipes for fast evaluation of these two steps, so that EPLL can embed a GGMM prior on an image with more than tens of thousands of patches. Our main contribution is to analyze the accuracy of our approximations based on thorough theoretical analysis. Our evaluations indicate that the GGMM prior is consistently a better fit for modeling image patch distribution and performs better on average in image denoising task.

Key words. Generalized Gaussian distribution, Mixture models, Image denoising, Patch priors.

AMS subject classifications. 68U10, 62H35, 94A08

1. Introduction.

Image restoration is the process of recovering the underlying clean image from its degraded or corrupted observation(s). The images captured by common imaging systems often contain corruptions such as noise, optical or motion blur due to sensor limitations and/or environmental conditions. For this reason, image restoration algorithms have widespread applications in medical imaging, satellite imaging, surveillance, and general consumer imaging applications. Priors on natural images play an important role in image restoration algorithms. Image priors are used to denoise or regularize ill-posed restoration problems such as deblurring and super-resolution, to name just a few. Classical image priors include Gibbs distributions [23], total variation which imposes Laplacian [53, 60] or hyper-Laplacian [30] prior on image gradients, and generalized Gaussian [33, 39, 6, 14] or scaled mixture of Gaussian [47] priors for wavelet or curvelet coefficients [3]. Alternatively, modeling the distribution of patches of an image (*i.e.*, small windows usually of size 8×8) has proven to be a powerful solution. In particular, popular patch techniques rely on non-local self-similarity [5], Fields of experts [51], learned patch dictionaries [1, 19], sparse or low-rank properties of stacks of similar patches [9, 11, 32, 29], patch re-occurrence priors [36], Gaussian mixture model prior on image gradients [21] or image patches [62, 61, 26].

Of these approaches, a successful approach introduced by Zoran and Weiss [62] is to model patch priors of clean natural images using Gaussian Mixture Models (GMM). The agility of this model lies in the fact that a prior learned on clean image patches can be effectively

*Institut de Mathématiques de Bordeaux, CNRS, Université de Bordeaux, Bordeaux INP, Talence, France (charles-alban.deledalle@math.u-bordeaux.fr).

†Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, USA (cdeledalle@eng.ucsd.edu, sparames@ucsd.edu, tqn001@eng.ucsd.edu).

employed to restore a wide range of inverse problems. It is also easily extendable to include other constraints such as sparsity or multi-resolution patches [57, 43]. The use of GMMs for patch priors make these methods computationally tractable and flexible. Although GMM patch prior is effective and popular, in this article, we argue that a mixture of Generalized Gaussian distributions (GGMM) is a better fit for image patch prior modeling. Compared to Gaussian models, Generalized Gaussian distributions (GGD) have an extra degree of freedom controlling the shape of the distribution and they encompass Gaussian and Laplacian models.

The superior patch prior modeling capability of a GGMM over a GMM is illustrated in Figure 5. The figure shows histograms of six orthogonal 1-D projections of a clustered subset of a patch database onto the eigenvectors of its covariance matrix. To illustrate the difference in the shapes (ν) and scales (λ) of the distributions, we have chosen a few projections corresponding to both the most and the least significant eigenvalues. It can be seen that GGD is a better fit on the obtained histograms than a Gaussian model. Additionally, different dimensions of the patch follow a different GGD. Hence, it does not suffice to model all the feature dimensions of a patch database as Laplacian or Gaussian. Therefore, we propose to model patch priors as Generalized Gaussian Mixture models (GGMM) with a separate shape and scale parameters for each feature dimension. This differs from the recent related approach in [42] that considered GGMM where each component has a fix shape parameter for all directions.

Organization. After explaining the considered patch prior based restoration framework in Subsection 2.1 and our motivations in Subsection 2.2, we derive our GGMM based restoration scheme in Section 3. Unlike [42], that incorporates a GGMM prior in a posterior mean estimator based on importance sampling, we directly extend the maximum *a posteriori* formulation of Zoran and Weiss [62] for the case of GGMM priors. While such a GGMM prior has the ability to capture the underlying distribution of clean patches more closely, we will show that it introduces two major computational challenges in this case. The first one can be thought of as a classification task in which a noisy patch is assigned to one of the components of the mixture. The second one corresponds to an estimation task where a noisy patch is denoised given that it belongs to one of the components of the mixture. Due to the interaction of the noise distribution with the GGD prior, we first show in Section 3 that these two tasks lead to a group of one-dimensional integration and optimization problems, respectively. Specifically, for $x \in \mathbb{R}$, these problems are of the following forms

$$\int_{\mathbb{R}} \exp\left(-\frac{(t-x)^2}{2\sigma^2} - \frac{|t|^\nu}{\lambda_\nu^\nu}\right) \mathrm{d}t \qquad \text{and} \qquad \operatorname*{argmin}_{t\in\mathbb{R}} \frac{(t-x)^2}{2\sigma^2} + \frac{|t|^\nu}{\lambda_\nu^\nu} \,. \tag{1}$$

for some $\nu > 0$, $\sigma > 0$ and $\lambda_\nu > 0$. These two problems are studied in Section 4 and Section 5, respectively. In general, they do not admit closed-form solutions but some particular solutions or approximations have been derived for the estimation/optimization problem [39, 7]. By contrast, up to our knowledge, little is known for approximating the classification/integration one (only crude approximations were proposed in [56]).

Contributions. A major contribution of this paper is to develop an accurate approximation for the classification/integration problem. In particular, we show that our approximation error vanishes for $x \to 0$ and $x \to \pm\infty$ when $\nu = 1$, see Theorem 1 and Theorem 2. We next generalize this result for $\frac{2}{3} < \nu < 2$ in Theorem 3 and Theorem 4. On top of that,

we prove that the two problems enjoy some important desired properties in Proposition 2 and Proposition 3. These theoretical results allow the two quantities to be approximated by functions that can be quickly evaluated in order to be incorporated in fast algorithms. Our last contribution is experimental and concerns the performance evaluation of the proposed model in image denoising scenario, see Section 6. Together with this paper, we have released our implementation at https://bitbucket.org/cdeledalle/ggmm-epll.

2. Background.

In this section we provide a brief background on the use of patch-based priors in image restoration and a quick overview of popular patch and image priors.

2.1. Image restoration with patch based priors.

We consider the problem of estimating an image $\boldsymbol{u} \in \mathbb{R}^N$ (N is the number of pixels) from noisy linear observations $\boldsymbol{v} = \mathcal{A}\boldsymbol{u} + \boldsymbol{w}$, where $\mathcal{A} : \mathbb{R}^N \to \mathbb{R}^M$ is a linear operator and $\boldsymbol{w} \in \mathbb{R}^M$ is a noise component assumed to be white and Gaussian with variance σ^2. In this paper, we will focus on standard denoising problems where $\mathcal{A}$ is the identity matrix, but in more general settings, it can account for loss of information such as blurring. Typical examples for operator $\mathcal{A}$ are: a low pass filter (for *deconvolution*), a masking operator (for *inpainting*), or a projection on a random subspace (for *compressive sensing*). To reduce noise and stabilize the inversion of $\mathcal{A}$, some *prior* information is used for the estimation of $\boldsymbol{u}$. Recent techniques [19, 62, 57] include this *prior* information as a model for the distribution of patches found in natural clean images. The Expected Patch Log-Likelihood (EPLL) algorithm [62] restores an image by maximum *a posteriori* estimation, corresponding to the following minimization problem:

$$\underset{\boldsymbol{u}\in\mathbb{R}^n}{\operatorname{argmin}} \frac{P}{2\sigma^2}\|\mathcal{A}\boldsymbol{u} - \boldsymbol{v}\|^2 - \sum_{i=1}^{N} \log p\left(\mathcal{P}_i\boldsymbol{u}\right) \tag{2}$$

where $\mathcal{P}_i : \mathbb{R}^N \to \mathbb{R}^P$ is the linear operator extracting a patch with P pixels centered at the pixel with location i (typically, $P = 8{\times}8$), and $p(.)$ is the *a priori* probability density function (*i.e.*, the statistical model of noiseless patches in natural images). While the first term in eq. (2) ensures that $\mathcal{A}\boldsymbol{u}$ is close to the observations $\boldsymbol{v}$ (this term is the negative log-likelihood under the white Gaussian noise assumption), the second term regularizes the solution $\boldsymbol{u}$ by favoring an image such that all of its patches fit the *prior* model of patches in natural images.

Optimization with half-quadratic splitting. Problem (2) is a large optimization problem where $\mathcal{A}$ couples all unknown pixel values of $\boldsymbol{u}$ and the patch prior is often chosen non-convex. A classical technique, known as half-quadratic splitting [22, 30], introduces N auxiliary unknown vectors $\boldsymbol{z}_i \in \mathbb{R}^P$, and alternatively consider the penalized optimization problem, for $\beta > 0$, as

$$\underset{\substack{\boldsymbol{u}\in\mathbb{R}^n \\ \boldsymbol{z}_1,\ldots,\boldsymbol{z}_N\in\mathbb{R}^P}}{\operatorname{argmin}} \frac{P}{2\sigma^2}\|\mathcal{A}\boldsymbol{u} - \boldsymbol{v}\|^2 + \frac{\beta}{2}\sum_{i\in\mathcal{I}}\|\mathcal{P}_i\boldsymbol{u} - \boldsymbol{z}_i\|^2 - \sum_{i\in\mathcal{I}} \log p\left(\boldsymbol{z}_i\right). \tag{3}$$

When $\beta \to \infty$, the problem (3) is equivalent to the original problem (2). In practice, an increasing sequence of β is considered, and the optimization is performed by alternating the minimization for $\boldsymbol{u}$ and $\boldsymbol{z}_i$. Though little is known about the convergence of this algorithm, few iterations produce in practice remarkable results.

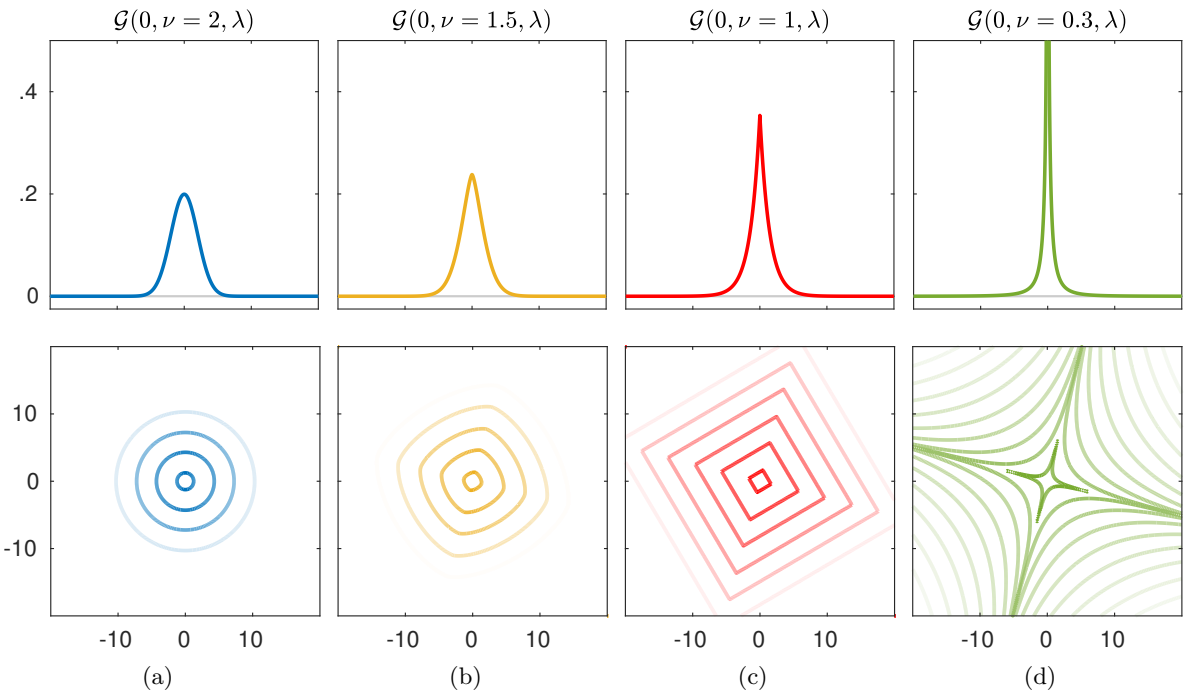


Figure 1. *Illustrations of zero-mean generalized Gaussian distributions for different values of the shape parameter ν. From left to right: $\nu = 2$, 1.5, 1, .3. Top: one-dimensional versions with variance $\lambda = 2$. Bottom: iso-contours of the two-dimensional versions with $\boldsymbol{W}$ being a rotation matrix .*

Minimization with respect to $\boldsymbol{u}$. Considering all $\boldsymbol{z}_i$ to be fixed, optimizing (3) for $\boldsymbol{u}$ corresponds to solving a linear inverse problem with a Tikhonov regularization, and has an explicit solution known as the linear minimum mean square estimator (or often referred to as Wiener filtering):

$$\hat{\boldsymbol{u}} = \operatorname*{argmin}_{\boldsymbol{u} \in \mathbb{R}^n} \frac{P}{2\sigma^2} \|\mathcal{A}\boldsymbol{u} - \boldsymbol{v}\|^2 + \frac{\beta}{2} \sum_{i \in \mathcal{I}} \|\mathcal{P}_i \boldsymbol{u} - \hat{\boldsymbol{z}}_i\|^2 \tag{4}$$

$$= \left(\mathcal{A}^t \mathcal{A} + \frac{\beta \sigma^2}{P} \sum_{i \in \mathcal{I}} \mathcal{P}_i^t \mathcal{P}_i \right)^{-1} \left(\mathcal{A}^t \boldsymbol{v} + \frac{\beta \sigma^2}{P} \sum_{i \in \mathcal{I}} \mathcal{P}_i^t \hat{\boldsymbol{z}}_i \right), \tag{5}$$

where $\mathcal{P}_i^t \mathcal{P}_i$ is a diagonal matrix whose i-th diagonal element corresponds to the number of patches overlapping the pixel of index i.

Minimization with respect to $\boldsymbol{z}_i$. Considering $\boldsymbol{u}$ to be fixed, optimizing (3) for $\boldsymbol{z}_i$ leads to

$$\hat{\boldsymbol{z}}_i \leftarrow \operatorname*{argmin}_{\boldsymbol{z}_i \in \mathbb{R}^P} \frac{\beta}{2} \|\tilde{\boldsymbol{z}}_i - \boldsymbol{z}_i\|^2 - \log p(\boldsymbol{z}_i) \quad \text{where} \quad \tilde{\boldsymbol{z}}_i = \mathcal{P}_i \hat{\boldsymbol{u}} \ . \tag{6}$$

which corresponds to the maximum *a posteriori* denoising problem under the patch prior p of a patch $\tilde{\boldsymbol{z}}_i$ contaminated by Gaussian noise with variance $1/\beta$. The solution of this optimization problem strongly depends on the properties of the chosen patch prior.

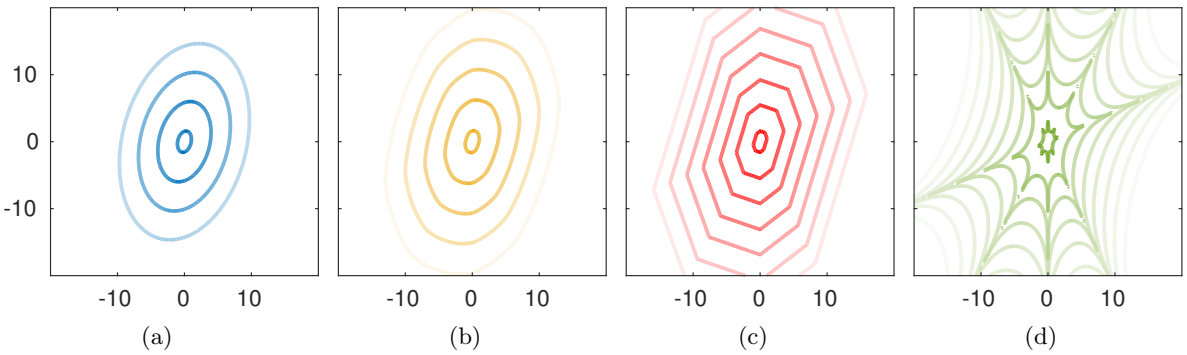


Figure 2. *Illustrations of the iso-lines of (a-d) four two-dimensional priors based on sparse analysis regularization with* $0 < \nu < 2$ *and a dictionary* $\boldsymbol{\Omega} \in \mathbb{R}^{4\times 2}$ *with four directions. From left to right: the shape parameter varies respectively as* $\nu = 2,\ 1.5,\ 1,\ .3$.

2.2. Overview of image/patch priors.

Whitening priors. A popular attempt in designing image priors, that can be applied to patch priors, considers a whitening transform $\boldsymbol{W} \in \mathbb{R}^{P\times P}$, *i.e.*, an orthogonal transform that decorrelated entries of $\boldsymbol{z}$. Typically, the transform $\boldsymbol{W}$ can be chosen by principal component analysis on a dataset of clean patches [11], or based on some prior knowledge, *e.g.*, by assuming decorrelations in the Fourier or wavelet domain [33]. Considering decorrelated and independent transform coefficients allows for the prior distribution to be separable in the transformed domain (*i.e.*, with independent marginals). Early works on wavelet representations of images (*e.g.*, Mallat, 1989 [33] and Moulin and Liu, 1999 [39]) suggested modeling such coefficients by a zero-mean generalized Gaussian distribution (GGD) of the form

$$p(\boldsymbol{z}) \propto \exp\left(-\rho^{\nu}\|\boldsymbol{W}\boldsymbol{z}\|_{\nu}^{\nu}\right) \tag{7}$$

where $\rho > 0$ and $\nu \geqslant 0$. Here, choosing $1 < \nu \leqslant 2$ models each coefficient by a bell shape distribution with small tails (the Gaussian distribution for $\nu = 2$), see first row on Figure 1(a-b). Image patches are thus assumed to be concentrated on a convex cluster (an ellipsoid for $\nu = 2$), see second row on Figure 1(a-b). In this setting, the coefficients tend to get closer and closer to zero (shrinkage) as ρ increases (becomes a linear shrinkage for $\nu = 2$). Consequently, this leads to diminishing edge features which yields over-smoothed images. To circumvent this issue $\nu \leqslant 1$ is considered. It corresponds to modeling each coefficient by a distribution with a peak at zero (non differentiability) and large tails, see first row on Figure 1(c-d). Image patches are thus assumed to be spread on an union of orthogonal subspaces aligned with the directions of the rows of $\boldsymbol{W}$, see second row on Figure 1(c-d). In this case, the shrinkage behaves as a thresholding operator such that most coefficients are set to zeros – *sparsity* – and a few non-zero coefficients are (more or less) preserved – *robustness*. As a consequence, under this scheme, edges can be expected to be preserved and resulting images are sharper. With such a prior, the maximum *a posteriori* problem corresponding to (2) is referred to as ℓ_{ν} regularization on the orthogonal dictionary whose atoms are the rows of $\boldsymbol{W}$. The subsequent optimization problem is well-known to be convex for $\nu \geqslant 1$, and thus $\nu = 1$ is a popular choice for practical applications.

Synthesis sparse priors. More recent works have shown that the distribution of clean image patches can be better modeled by assuming they are sparse combinations of the columns of a redundant dictionary $\boldsymbol{D} \in \mathbb{R}^{P \times Q}$, $Q \geqslant P$ (*i.e.*, composed of linearly dependent atoms). This is typically achieved with undecimated (*a.k.a,* shift-invariant) wavelets [54, 8] or, more generally, frames [17, 34]. A dictionary can also be learned on a large dataset of clean patches with the k-SVD algorithm [1]. This is to suggest that the coefficients of an image are spread on a union of non-orthogonal subspaces aligned with the columns of $\boldsymbol{D}$, see Figure 2. The so-called synthesis regularization framework [20], applied to image patches in [19, 57], corresponds to the following choice of prior distribution

$$p(\boldsymbol{z}) \propto \exp\left(-\rho^{\nu} \|\boldsymbol{\alpha}\|_{\nu}^{\nu}\right) \quad \text{subject to} \quad \boldsymbol{z} = \boldsymbol{D}\boldsymbol{\alpha} \;, \tag{8}$$

where $\rho > 0$ and $\nu \geqslant 0$. Similar to the whitening framework, choosing $\nu \leqslant 1$ enforces sparsity without penalizing large non-zero coefficients. A difficulty with such an approach is to deal with the non-orthogonality of the dictionary in the non-convex setting $\nu < 1$. Typically, the authors of [19, 57] consider the case $\nu = 0$ for which greedy techniques such as orthogonal matching pursuit can be employed [10].

Analysis sparse priors. Alternatively, the distribution of clean images can also be captured by modeling the correlations of its patches with the rows of a redundant dictionary $\boldsymbol{\Omega} \in \mathbb{R}^{Q \times P}$ that does not require to span the set of clean images ($\boldsymbol{\Omega}$ can be rank deficient). This is the case with the Total-Variation model [53] that considers the gradient of an image $\boldsymbol{\Omega} = \nabla$, and any type of filter bank analysis. Similar to the synthesis framework, the dictionary can also be learned on a large dataset of clean patches with the analysis k-SVD algorithm [52]. This suggests that the correlations of an image are spread on a union of non-orthogonal subspaces aligned with the rows of $\boldsymbol{\Omega}$. The analysis regularization framework [20], corresponds to the following choice of prior

$$p(\boldsymbol{z}) \propto \exp\left(-\rho^{\nu} \|\boldsymbol{\Omega}\boldsymbol{z}\|_{\nu}^{\nu}\right) \tag{9}$$

where $\rho > 0$ and $\nu > 0$. Again, choosing $\nu \leqslant 1$ enforces correlations to be mostly zeros – *co-sparsity* – and a few large non-zero correlations. A difficulty with such an approach is to cope with the non-injectivity of $\boldsymbol{\Omega}$.

GMM priors. Rather than modeling clean patches as spread on a union of non-orthogonal sub-spaces, an alternative is to consider a union of ellipsoids (called clusters). To this end, the authors of [61, 62] suggested using a zero-mean Gaussian mixture model (GMM) prior[1], that, for any patch $\boldsymbol{z} \in \mathbb{R}^P$, is given by

$$p(\boldsymbol{z}) = \sum_{k=1}^{K} w_k \mathcal{N}_P(\boldsymbol{z}; 0_P, \boldsymbol{\Sigma}_k) \tag{10}$$

where K is the number of components, $w_k > 0$ are weights such that $\sum_k w_k = 1$, and $\mathcal{N}_P(0_P, \boldsymbol{\Sigma}_k)$ denotes the multi-variate Gaussian distribution with zero-mean and covariance

[1] To enforce the zero-mean assumption, patches are first centered on zero, then denoised using eq. (6), and, finally, their initial means are added back. In fact, one can show that it corresponds to modeling $p(\boldsymbol{z} - \bar{\boldsymbol{z}})$ with a GMM where $\bar{z}_j = \frac{1}{P}\sum_i z_i$ for all $1 \leqslant j \leqslant P$.

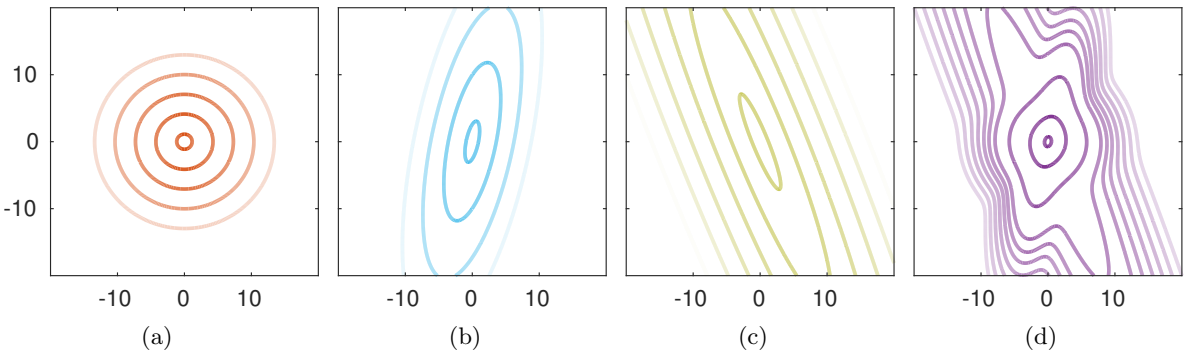


Figure 3. *Illustrations of the iso-lines of (a-c) three two-dimensional zero-mean Gaussian distributions and (d) their mixture with weights* $1/2$*,* $1/3$ *and* $1/6$ *respectively.*

$\boldsymbol{\Sigma}_k \in \mathbb{R}^{P\times P}$. An illustration of a GMM and its components is given in Figure 3. The hyper-parameters w_k and $\boldsymbol{\Sigma}_k$ can be learned using the Expectation Maximization algorithm [13] on a large dataset of clean patches [62]. As will be discussed in detail in Section 3, such a prior in the optimization problem (6) will require (i) looking for the cluster $k^\star$ that best explains the given patch $\boldsymbol{z}$, (ii) performing whitening by projecting $\boldsymbol{z}$ over the main directions of that cluster (given by the eigenvectors of $\boldsymbol{\Sigma}_{k^\star}$), and (iii) applying a linear shrinkage on the coefficients with respect to the spread of the cluster (encoded by the eigenvalues). As a function of $\boldsymbol{z}$, the resulting approach is a piece-wise linear estimator (PLE) [61]. Last but not least, when $K = 1$, this approach is equivalent to the whitening approach with $\rho\boldsymbol{W} = \boldsymbol{\Sigma}_1^{-1/2}$ and $\nu = 2$.

An advantage of the GMM prior over sparse priors is that all parameters of the prior model (*i.e.*, $\boldsymbol{\Sigma}_k$ and w_k) are learned offline. As a consequence, provided that the training dataset is representative of the image to be restored, no regularization parameters are required to be tuned during runtime[2]. By contrast, synthesis (resp. analysis) sparse priors often consider dictionaries $\boldsymbol{D}$ (resp. $\boldsymbol{\Omega}$) with atoms of unit or prescribed norm. Consequently, the scaling of the prior controlled by ρ cannot be fixed once for all images, but will depend on the specific underlying image and the inverse problem that is being solved (*i.e.*, $\mathcal{A}$ and σ). For this reason, the parameter ρ must be tuned by the practitioner by trial and error, cross-validation or other dedicated techniques such as [24, 15, 25, 49, 12].

Note that the overview given here is mainly to provide a background for our method and is far from being exhaustive. Beyond the four aforementioned priors, many other image and patch priors have been proposed in the literature, including Gaussian scale mixtures [47], Fields-of-Experts [51], Total Generalized Variation [4] and Student mixture models [55] to just cite a few that we have omitted.

In this paper, we propose to fill the gap between sparse regularization techniques and GMM priors. To this end we suggest using a mixture of generalized Gaussian distributions

[2]Though all regularization parameters are learned, the solver for eq. (2) will have some free internal parameters, controlling the convergence, that require some tuning.

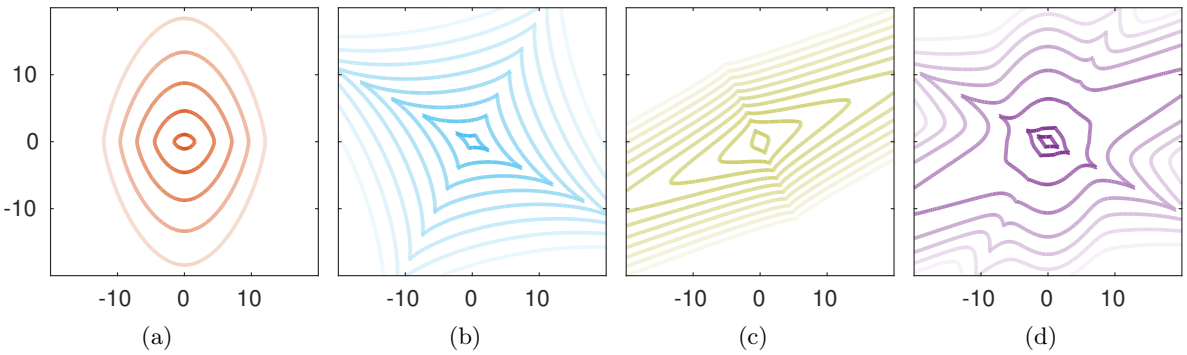


Figure 4. *Illustrations of the iso-lines of (a-c) three two-dimensional zero-mean generalized Gaussian distributions with* $0 < \nu < 2$ *and (d) their mixture with weights* $1/2$, $1/3$ *and* $1/6$ *respectively.*

that will enable image patches to be spread over clusters that are bell shaped in some directions while being peaky in others, see Figure 4. While the use of GMM priors leads to piece-wise linear estimator, our GGMM prior will lead to a piecewise non-linear shrinkage estimator.

3. Generalized Gaussian Mixture Models. In this paper, we aim to learn K orthogonal transforms such that each of them can map a subset (cluster) of clean patches into independent zero-mean coefficients. Instead of assuming the coefficients to be identically distributed, we consider that both the scale and the shape of their distributions may vary from one coordinate to another (within the same transform). Our motivation to assume such a highly flexible model is based on the observation illustrated in Figure 5. Given one of such transform and its corresponding cluster of patches, we have displayed the histogram of the patch coefficients for six different coordinates. It can be clearly observed that the shape of the distribution varies depending on the coordinate. Some of them are peaky with heavy tails, and, therefore, would not be faithfully captured by a Gaussian distribution, as done in EPLL [62]. By contrast, some others have a bell shape, and so would not be captured properly by a peaky and heavy tailed distribution, as done by analysis sparse models [20]. This shows that one cannot simultaneously decorrelate and sparsify a cluster of clean patches for all coordinates. Since some of the coordinates reveal sparsity while some others reveal Gaussianity, we propose to use a more flexible model that can capture such variations. We propose using a multi-variate zero-mean generalized Gaussian mixture model (GGMM)

$$p(\boldsymbol{z}) = \sum_{k=1}^{K} w_k \mathcal{G}(\boldsymbol{z}; 0_P, \boldsymbol{\Sigma}^{(k)}, \boldsymbol{\nu}^{(k)}) \tag{11}$$

where K is the number of components and $w_k > 0$ are weights such that $\sum_k w_k = 1$. The notation $\mathcal{G}(0_P, \boldsymbol{\Sigma}, \boldsymbol{\nu})$ denotes the P-dimensional generalized Gaussian distribution (GGD) with zero-mean, covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{P \times P}$ (symmetric positive definite) and shape parameter $\boldsymbol{\nu} \in \mathbb{R}^P$, whose expression is

$$\mathcal{G}(\boldsymbol{z}; 0_P, \boldsymbol{\Sigma}, \boldsymbol{\nu}) = \frac{\mathcal{K}}{2|\boldsymbol{\Sigma}_{\boldsymbol{\nu}}|^{1/2}} \exp\left[-\|\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{-1/2}\boldsymbol{z}\|_{\boldsymbol{\nu}}^{\boldsymbol{\nu}}\right] \quad \text{with} \quad \|\boldsymbol{x}\|_{\boldsymbol{\nu}}^{\boldsymbol{\nu}} = \sum_{j=1}^{P} |x_j|^{\nu_j}, \tag{12}$$

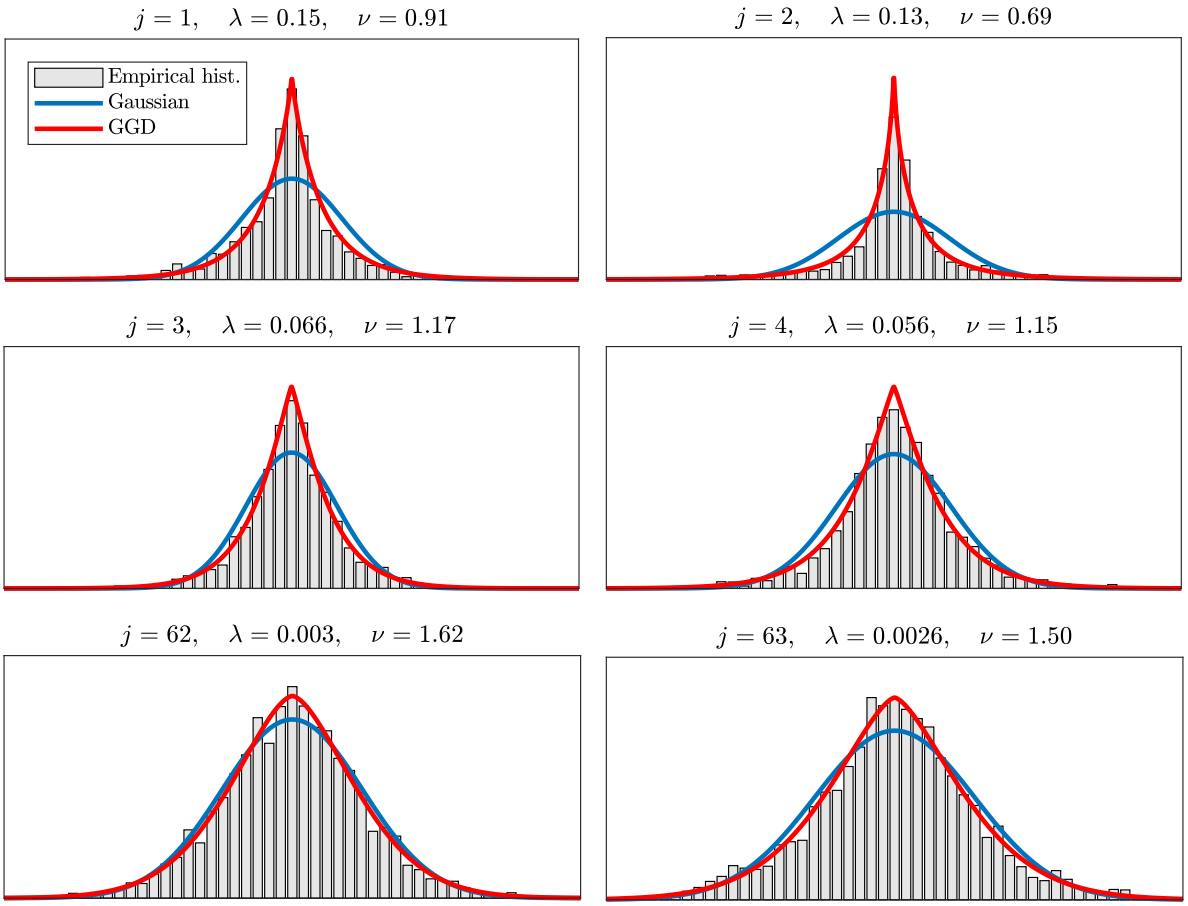


Figure 5. *Histograms of the projection of* $200,000$ *clean patches on 6 eigenvectors* $j = 1, 2, 3, 4, 62$ *and* 63 *of the covariance matrix of one component* k *of the mixture (with weight* $w_k = 1.3\%$*). The contribution of each clean patch in the histograms is given by its membership values onto this component* k *(as obtained during the E-Step of EM). For each histogram, a generalized Gaussian distribution was adjusted by estimating the parameters* λ *and* ν *by moment estimation (as obtained during the M-Step of our modified EM). For comparisons, we have also provided illustrations of the best fit obtained with a Gaussian distribution.*

$$\text{(13)} \qquad \text{where} \quad \mathcal{K} = \prod_{j=1}^{P} \frac{\nu_j}{\Gamma(1/\nu_j)} \quad \text{and} \quad \boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{1/2} = \boldsymbol{\Sigma}^{1/2} \begin{pmatrix} \sqrt{\frac{\Gamma(1/\nu_1)}{\Gamma(3/\nu_1)}} & & \\ & \ddots & \\ & & \sqrt{\frac{\Gamma(1/\nu_P)}{\Gamma(3/\nu_P)}} \end{pmatrix}.$$

Denoting the eigen decomposition of matrix $\boldsymbol{\Sigma}$ by $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^t$ such that $\boldsymbol{U} \in \mathbb{R}^{P \times P}$ is unitary and $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1^k, \lambda_2^k, \ldots, \lambda_P^k)^2$ is diagonal with positive diagonal elements $(\lambda_j^k)^2$, $\boldsymbol{\Sigma}^{1/2}$ in the above expression is defined as $\boldsymbol{\Sigma}^{1/2} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}$ and $\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^t$ is its inverse.

When $\boldsymbol{\nu}$ is a constant vector with all entries equal to $\nu_j = 2$, $\mathcal{G}(0_P, \boldsymbol{\Sigma}, \boldsymbol{\nu})$ is the multi-variate Gaussian distribution $\mathcal{N}(0_P, \boldsymbol{\Sigma})$ (as used in EPLL [62]). When all $\nu_j = 1$, it is the multi-variate Laplacian distribution and the subsequent GGMM is a Laplacian Mixture Model (LMM). When all $\nu_j < 1$, it is the multi-variate hyper-Laplacian distribution and the

subsequent GGMM is a hyper-Laplacian Mixture Model (HLMM). Choosing $K = 1$ with a constant vector $\boldsymbol{\nu}$ corresponds to ℓ_ν regularization after whitening with $\rho \boldsymbol{W} = \boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{-1/2}$. But as motivated earlier, unlike classical multivariate GGD models [3, 45, 42], we allow for the entries of $\boldsymbol{\nu}$ to vary from one coordinate j to another. To the best of our knowledge, the proposed work is the first one to consider this fully flexible model.

Proposition 1. *The multi-variate zero-mean GGD can be written as*

$$\mathcal{G}(\boldsymbol{z}; 0_P, \boldsymbol{\Sigma}, \boldsymbol{\nu}) = \prod_{j=1}^{P} \mathcal{G}(x_j; 0, \lambda_j, \nu_j) \quad \textit{with} \quad \boldsymbol{x} = \boldsymbol{U}^t \boldsymbol{z} \tag{14}$$

$$\textit{where} \quad \mathcal{G}(x; 0, \lambda, \nu) = \frac{\kappa}{2\lambda_\nu} \exp\left[-\left(\frac{|x|}{\lambda_\nu} \right)^\nu \right]$$

$$\textit{where} \quad \kappa = \frac{\nu}{\Gamma(1/\nu)} \quad \textit{and} \quad \lambda_\nu = \lambda \sqrt{\frac{\Gamma(1/\nu)}{\Gamma(3/\nu)}} \,,$$

where $x \mapsto \mathcal{G}(x; 0, \lambda, \nu)$ is a real, even, unimodal, bounded and continuous probability density function. It is also differentiable everywhere except for $x = 0$ when $\nu \leqslant 1$.

The proof follows directly by injecting the eigen decomposition of $\boldsymbol{\Sigma}$ in (12) and basic properties of $x \mapsto |x|^\nu$. Proposition 1 shows that, for each of the K clusters, eq. (12), indeed, models a prior that is separable in a coordinate system obtained by applying the whitening transform $\boldsymbol{U}^t$. Not only is the prior separable for each coordinate j, but the shape (ν_j) and scale (λ_j) of the distribution may vary. As observed in Figure 5, the proposed GGMM models the underlying distributions of a cluster of clean-patches much better than GGM.

Patch denoising under GGMM prior. We now explain why solving (6) is non-trivial when using a GGMM patch prior. In this case, for a noisy patch $\tilde{\boldsymbol{z}}$ with variance σ^2, equation (6) becomes

$$\hat{\boldsymbol{z}} \leftarrow \operatorname*{argmin}_{\boldsymbol{z} \in \mathbb{R}^P} \frac{1}{2\sigma^2} \|\tilde{\boldsymbol{z}} - \boldsymbol{z}\|^2 - \log \left[\sum_{k=1}^{K} w_k \mathcal{G}(\boldsymbol{z}; 0_P, \boldsymbol{\Sigma}^{(k)}, \boldsymbol{\nu}^{(k)}) \right] . \tag{15}$$

Due to the multi-modality of the GGMM prior, this optimization problem is highly non-convex. To circumvent this issue, we follow the strategy used by EPLL [62] in the specific case of Gaussian mixture model prior. The idea is to restrict the sum involved in the logarithm in eq. (15) to only one component $k^\star$. If we consider the best $k^\star$ to be given (the strategy to select the best $k^\star$ will be discussed in the next section), then eq. (15) is approximated by the following simpler problem

$$\hat{\boldsymbol{z}} \leftarrow \operatorname*{argmin}_{\boldsymbol{z} \in \mathbb{R}^P} \left\{ \frac{\|\tilde{\boldsymbol{z}} - \boldsymbol{z}\|^2}{2\sigma^2} - \log \mathcal{G}(\boldsymbol{z}; 0_P, \boldsymbol{\Sigma}^{(k^\star)}, \boldsymbol{\nu}^{(k^\star)}) = \frac{\|\tilde{\boldsymbol{z}} - \boldsymbol{z}\|^2}{2\sigma^2} + \|\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{(k^\star)^{-1/2}} \boldsymbol{z}\|_\nu^\nu \right\} . \tag{16}$$

The main advantage of this simplified version is that, by virtue of Proposition 1, the underlying optimization becomes tractable and can be separated into P one-dimensional optimization problems, as:

$$\hat{\boldsymbol{z}} = \boldsymbol{U} \hat{\boldsymbol{x}} \quad \text{where} \quad \hat{x}_j = s_{\sigma, \lambda_j}^{\nu_j}(\tilde{x}_j) \quad \text{with} \quad \tilde{\boldsymbol{x}} = \boldsymbol{U}^t \tilde{\boldsymbol{z}} \tag{17}$$

$$\text{and} \quad s^{\nu}_{\sigma,\lambda}(x) \in \operatorname*{argmin}_{t\in\mathbb{R}} \frac{1}{2\sigma^2}(t-x)^2 + \frac{|t|^{\nu}}{\lambda_{\nu}^{\nu}} \quad \text{where} \quad \lambda_{\nu} = \lambda\sqrt{\frac{\Gamma(1/\nu)}{\Gamma(3/\nu)}} \; . \tag{18}$$

While the problem is not necessarily convex, its solution $s^{\nu}_{\sigma,\lambda}$ is always uniquely defined almost everywhere (see, Section 5). We call this almost everywhere real function $s^{\nu}_{\sigma,\lambda} : \mathbb{R} \to \mathbb{R}$ shrinkage function. When $\nu = 2$, it is a linear function that is often referred to as Wiener shrinkage. When $\nu \neq 2$, as we will discuss in Section 5, it is a non-linear shrinkage function that can be computed in closed form for some cases or with some approximations.

Now, we address the question of finding a strategy for choosing a relevant component $k^{\star}$ to replace the mixture distribution inside the logarithm. The optimal component $k^{\star}$ can be obtained by maximizing the posterior as

$$k^{\star} \in \operatorname*{argmax}_{1\leqslant k\leqslant K} p(k \mid \tilde{\boldsymbol{z}}) = \operatorname*{argmax}_{1\leqslant k\leqslant K} w_k p(\tilde{\boldsymbol{z}} \mid k) = \operatorname*{argmin}_{1\leqslant k\leqslant K} \; -\log w_k - \log p(\tilde{\boldsymbol{z}} \mid k) \tag{19}$$

where the weights of the GGMM corresponds to the prior probability $w_k = p(k)$. We next use the fact that the patch $\tilde{\boldsymbol{z}}$ (conditioned on k) can be expressed as $\tilde{\boldsymbol{z}} = \boldsymbol{z} + \boldsymbol{n}$ where $\boldsymbol{z}$ and $\boldsymbol{n}$ are two independent random variables from distributions $\mathcal{G}(0_P, \boldsymbol{\Sigma}^{(k)}, \boldsymbol{\nu}^{(k)})$ and $\mathcal{N}(0_P, \sigma^2\mathbf{Id}_P)$ respectively. It follows that the distribution of $\tilde{\boldsymbol{z}}$ is the convolution of these latter two, and then

$$-\log p(\tilde{\boldsymbol{z}} \mid k) = -\log \int_{\mathbb{R}^P} \mathcal{G}(\tilde{\boldsymbol{z}} - \boldsymbol{z}; 0_P, \boldsymbol{\Sigma}^{(k)}, \boldsymbol{\nu}^{(k)}) \cdot \mathcal{N}(\boldsymbol{z}; 0_P, \sigma^2\mathbf{Id}_P) \, \mathrm{d}\boldsymbol{z} \; . \tag{20}$$

We next use Proposition 1 to separate this integration problem into P one-dimensional integration problems as

$$-\log p(\tilde{\boldsymbol{z}} \mid k) = \sum_{j=1}^{P} f^{\nu_j^{(k)}}_{\sigma,\lambda_j^{(k)}}(\tilde{x}_j) \quad \text{with} \quad \tilde{\boldsymbol{x}} = \boldsymbol{U}^t\tilde{\boldsymbol{z}} \; , \tag{21}$$

$$\text{where} \quad f^{\nu}_{\sigma,\lambda}(x) = -\log \int_{\mathbb{R}} \mathcal{G}(x - t; 0, \lambda, \nu) \cdot \mathcal{N}(t; 0, \sigma^2) \, \mathrm{d}t \; . \tag{22}$$

We call the real function $f^{\nu}_{\sigma,\lambda} : \mathbb{R} \to \mathbb{R}$ the discrepancy function which measures the goodness of fit of a GGD to the noisy value x. When $\nu = 2$, this function is quadratic with x. For $\nu \neq 2$, as we will discuss in Section 4, it is a non-quadratic function, that can be efficiently approximated based on an in-depth analysis of its asymptotic behavior.

The next two sections are dedicated to the analysis and approximations of the discrepancy function $f^{\nu}_{\sigma,\lambda}$ and the shrinkage function $s^{\nu}_{\sigma,\lambda}$, respectively.

4. Discrepancy function: analysis and approximations. From its definition given in eq. (22), the discrepancy function reads for $\nu > 0$, $\sigma > 0$ and $\lambda > 0$, as

$$f^{\nu}_{\sigma,\lambda}(x) = -\log \frac{1}{\sqrt{2\pi}\sigma} \frac{\nu}{2\lambda_{\nu}\Gamma(1/\nu)} - \log \int_{-\infty}^{\infty} \exp\left(-\frac{(x-t)^2}{2\sigma^2}\right) \exp\left[-\left(\frac{|t|}{\lambda_{\nu}}\right)^{\nu}\right] \mathrm{d}t \; . \tag{23}$$

It corresponds to the negative logarithm of the distribution of the sum of a zero-mean generalized Gaussian and a zero-mean Gaussian random variables. When $\nu = 2$, the generalized

Gaussian random variable becomes Gaussian, and the resulting distribution is also Gaussian with zero-mean and variance $\sigma^2 + \lambda^2$, and then

$$f^2_{\sigma,\lambda}(x) = \frac{1}{2}\left[\log 2\pi + \log(\sigma^2+\lambda^2) + \frac{x^2}{\sigma^2+\lambda^2}\right] . \tag{24}$$

Remark 1. *For $\nu = 2$, a direct consequence of (24) is that $-\log p(\tilde{\boldsymbol{z}} \mid k)$ is as an affine function of the Mahanalobis distance between $\tilde{\boldsymbol{z}}$ and 0_P for the covariance matrix $\boldsymbol{\Sigma}^{(k)}+\sigma^2\mathbf{Id}_P$:*

$$-\log p(\tilde{\boldsymbol{z}} \mid k) = \frac{1}{2}\left[P\log 2\pi + \log|\boldsymbol{\Sigma}^{(k)} + \sigma^2\mathbf{Id}_P| + \tilde{\boldsymbol{z}}^t(\boldsymbol{\Sigma}^{(k)} + \sigma^2\mathbf{Id}_P)^{-1}\tilde{\boldsymbol{z}}\right] . \tag{25}$$

When $\nu = 1$, the generalized Gaussian random variable becomes Laplacian, and the distribution resulting from the convolution also has a closed form which leads to the following discrepancy function

$$f^1_{\sigma,\lambda}(x) = \log(2\sqrt{2}\lambda) - \frac{\sigma^2}{\lambda^2} - \log\left[e^{\frac{\sqrt{2}x}{\lambda}}\operatorname{erfc}\left(\frac{x}{\sqrt{2}\sigma}+\frac{\sigma}{\lambda}\right) + e^{-\frac{\sqrt{2}x}{\lambda}}\operatorname{erfc}\left(-\frac{x}{\sqrt{2}\sigma}+\frac{\sigma}{\lambda}\right)\right] , \tag{26}$$

refer to Appendix A for derivation.

To the best of our knowledge, there are no simple expressions for other values of ν. One solution proposed by [56] is to express this in terms of the bi-variate Fox-H function [37]. This, rather cumbersome expression, is computationally demanding. In practice, this special function requires numerical integration techniques over complex lines [46], and is thus difficult to numerically evaluate it efficiently. Since, in our application, we need to evaluate this function a large number of times, we cannot utilize this solution.

In [56], the authors have also proposed to approximate this non-trivial distribution by another GGD. For fixed values of σ, λ and ν, they proposed three different numerical techniques to estimate its parameters λ' and ν' that best approximate either the kurtosis, the tail or the cumulative distribution function. Based on their approach, the discrepancy function $f^\nu_{\sigma,\lambda}(x)$ would thus be a power function of the form $x^{\nu'}$.

In this paper, we show that $f^\nu_{\sigma,\lambda}$ does, indeed, asymptotically behave as a power function for small and large values of x, but the exponent can be quite different for these two asymptotics. We believe that these different behaviors are important to be preserved in our application context. For this reason, $f^\nu_{\sigma,\lambda}$ cannot be modeled as a power function through a GGD distribution. Instead, we provide an alternative solution that is able to capture the correct behavior for both of these asymptotics, and that also permits fast computation.

4.1. Theoretical analysis. In this section, we perform a thorough theoretical analysis of the discrepancy function, in order to approximate it accurately. Let us first introduce some basic properties regarding the discrepancy function.

Proposition 2. *Let $\nu > 0$, $\sigma > 0$, $\lambda > 0$ and $f^\nu_{\sigma,\lambda}$ as defined in eq. (22). The following relations hold true*

$$f^\nu_{\sigma,\lambda}(x) = \log\sigma + f^\nu_{1,\lambda/\sigma}(x/\sigma) , \tag{reduction}$$

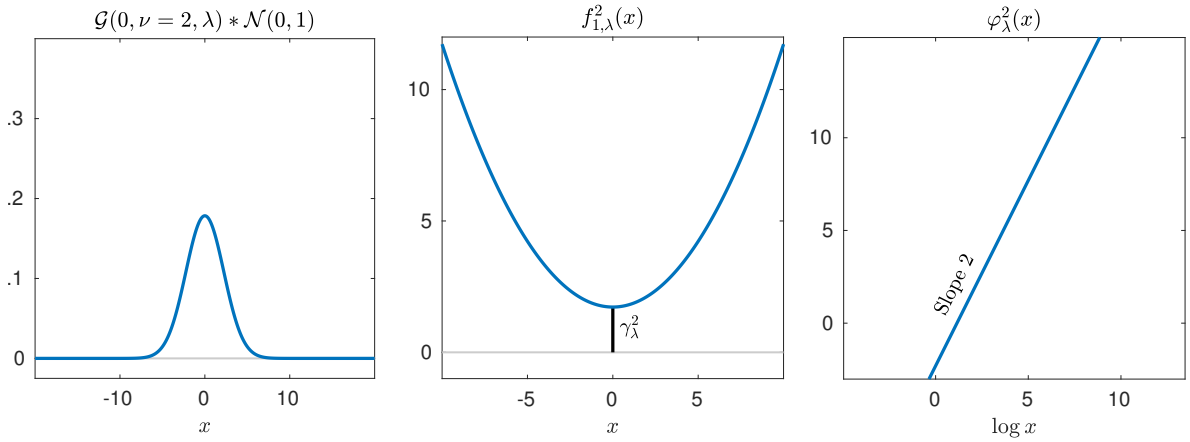


Figure 6. *From left to right: the convolution of a Gaussian distribution with standard deviation $\lambda = 2$ with a Gaussian distribution with standard deviation $\sigma = 1$, the corresponding discrepancy function and log-discrepancy function.*

$$\text{(even)} \qquad f^\nu_{\sigma,\lambda}(x) = f^\nu_{\sigma,\lambda}(-x) \ ,$$

$$\text{(unimodality)} \qquad |x| \geqslant |y| \Leftrightarrow f^\nu_{\sigma,\lambda}(|x|) \geqslant f^\nu_{\sigma,\lambda}(|y|) \ ,$$

$$\text{(lower bound at 0)} \qquad \min_{x\in\mathbb{R}} f^\nu_{\sigma,\lambda}(x) = f^\nu_{\sigma,\lambda}(0) > -\infty \ .$$

The proofs can be found in Appendix B. Based on Proposition 2, we can now express the discrepancy function $f^\nu_{\sigma,\lambda}(x) : \mathbb{R} \to \mathbb{R}$ in terms of a constant γ^ν_λ and another function $\varphi^\nu_\lambda : \mathbb{R}^*_+ \to \mathbb{R}$, both of which can be parameterized by only two parameters $\lambda > 0$ and $\nu > 0$, as

$$f^\nu_{\sigma,\lambda}(x) = \log\sigma + \gamma^\nu_{\lambda/\sigma} + \begin{cases} e^{\varphi^\nu_{\lambda/\sigma}(|x/\sigma|)} & \text{if} \quad x \neq 0 \ , \\ 0 & \text{otherwise} \ , \end{cases} \tag{27}$$

$$\text{where} \quad \varphi^\nu_\lambda(x) = \log\left[f^\nu_{1,\lambda}(x) - \gamma^\nu_\lambda\right] \quad \text{and} \quad \gamma^\nu_\lambda = f^\nu_{1,\lambda}(0) \ . \tag{28}$$

We call φ^ν_λ the log-discrepancy function.

At this point, let us consider an instructive toy example for the case when $\nu = 2$. In this case, from eq. (24), we can deduce that the log-discrepancy function is a log-linear function (*i.e.*, a linear function of $\log x$)

$$\varphi^2_\lambda(x) = \alpha \log x + \beta \ , \tag{29}$$

$$\text{and} \quad \gamma^2_\lambda = \tfrac{1}{2}\left[\log 2\pi + \log(1+\lambda^2)\right] \ , \tag{30}$$

$$\text{where} \quad \alpha = 2 \quad \text{and} \quad \beta = -\log 2 - \log(1+\lambda^2) \ . \tag{31}$$

Here, the slope $\alpha = 2$ reveals the quadratic behavior of the discrepancy function. Figure 6 gives an illustration of the resulting convolution (a Gaussian distribution), the discrepancy function (a quadratic function) and the log-discrepancy (a linear function with slope 2). Note that quadratic metrics are well-known to be non-robust to outliers, which is in complete agreement with the fact that Gaussian priors have thin tails.

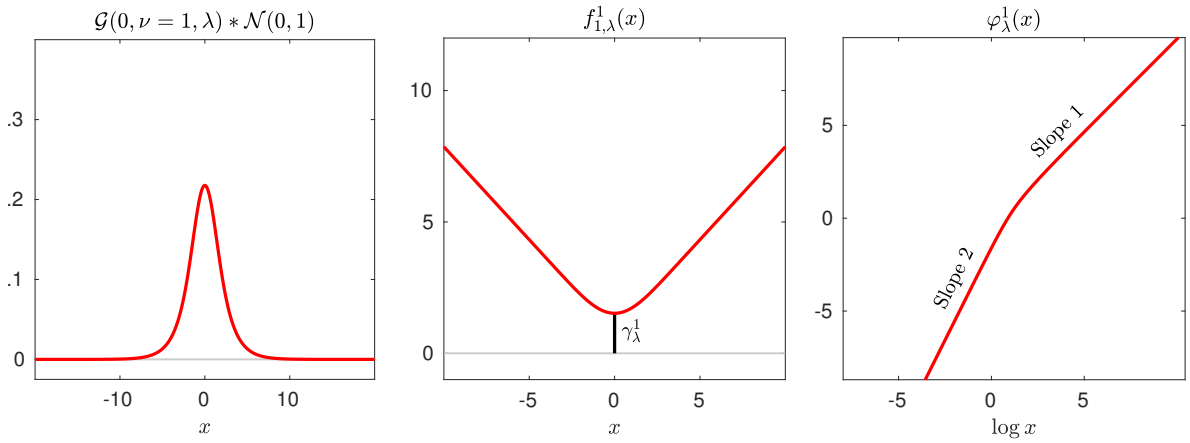


Figure 7. *From left to right: the convolution of a Laplacian distribution with standard deviation $\lambda = 2$ with a Gaussian distribution with standard deviation $\sigma = 1$, the corresponding discrepancy function and log-discrepancy function.*

Another example is the case of $\nu = 1$. From eq. (26), the log-discrepancy is given by

$$\varphi^1_\lambda(x) = \log\left[\log\left[2\operatorname{erfc}\left(\tfrac{1}{\lambda}\right)\right] - \log\left[e^{\frac{\sqrt{2}x}{\lambda}}\operatorname{erfc}\left(\tfrac{x}{\sqrt{2}} + \tfrac{1}{\lambda}\right) + e^{-\frac{\sqrt{2}x}{\lambda}}\operatorname{erfc}\left(-\tfrac{x}{\sqrt{2}} + \tfrac{1}{\lambda}\right)\right]\right] , \tag{32}$$

$$\text{and} \quad \gamma^1_\lambda = \frac{1}{2}\log 2 + \log\lambda - \frac{1}{\lambda^2} - \log\left[\operatorname{erfc}\left(\tfrac{1}{\lambda}\right)\right] . \tag{33}$$

Unlike for $\nu = 2$, this function is not log-linear and thus $f^1_{\sigma,\lambda}$ is not a power function. Nevertheless, as shown by the next two theorems, it is also asymptotically log-linear for small and large values of x.

Theorem 1. *The function φ^1_λ is asymptotically log-linear in the vicinity of* 0

$$\varphi^1_\lambda(x) \underset{0}{\sim} \alpha_1 \log x + \beta_1 , \tag{34}$$

$$\textit{where} \quad \alpha_1 = 2 \quad \textit{and} \quad \beta_1 = -\log\lambda + \log\left[\frac{1}{\sqrt{\pi}}\frac{\exp\left(-\frac{1}{\lambda^2}\right)}{\operatorname{erfc}\left(\frac{1}{\lambda}\right)} - \frac{1}{\lambda}\right] . \tag{35}$$

The proof can be found in Appendix C.

Theorem 2. *The function φ^1_λ is asymptotically log-linear in the vicinity of $+\infty$*

$$\varphi^1_\lambda(x) \underset{\infty}{\sim} \alpha_2 \log x + \beta_2 , \tag{36}$$

$$\textit{where} \quad \alpha_2 = 1 \quad \textit{and} \quad \beta_2 = \frac{1}{2}\log 2 - \log\lambda . \tag{37}$$

The proof can be found in Appendix D.

Theorem 1 and Theorem 2 show that φ^1_λ has two different asymptotics that can be approximated by a log-linear function. Interestingly, the exponent $\alpha_1 = 2$ in the vicinity of 0 shows that the Gaussian distribution involved in the convolution prevails over the Laplacian distribution and thus, the behavior of $f^1_{\sigma,\lambda}$ is quadratic. Similarly, the exponent $\alpha_2 = 1$ in the

vicinity of $+\infty$ shows that the Laplacian distribution involved in the convolution prevails over the Gaussian distribution and the behavior of $f^1_{\sigma,\lambda}$ is then linear. These results are supported by Figure 7 which illustrates the resulting convolution, the discrepancy function (eq. (26)) and the log-discrepancy function (eq. (32)). Furthermore, the discrepancy function $f^1_{\sigma,\lambda}$ shares a similar behavior with the well-known Huber loss function [28], known to be more robust to outliers. This is again in complete agreement with the fact that Laplacian priors have heavier tails.

In the case $\frac{2}{3} < \nu < 2$, even though φ^ν_λ has no simple closed form expression, the similar conclusions can be made as a result of the next two theorems.

Theorem 3. *Let $\nu > 0$. The function φ^ν_λ is asymptotically log-linear in the vicinity of 0*

$$\varphi^\nu_\lambda(x) \underset{0}{\sim} \alpha_1 \log x + \beta_1 \ ,$$

$$\textit{where} \quad \alpha_1 = 2 \quad \textit{and} \quad \beta_1 = -\log 2 + \log\left(1 - \frac{\displaystyle\int_{-\infty}^{\infty} t^2 e^{-\frac{t^2}{2}} \exp\left[-\left(\frac{|t|}{\lambda_\nu}\right)^\nu\right] \mathrm{d}t}{\displaystyle\int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} \exp\left[-\left(\frac{|t|}{\lambda_\nu}\right)^\nu\right] \mathrm{d}t}\right) \ .$$

The proofs can be found in Appendix E.

Theorem 4. *Let $\frac{2}{3} < \nu < 2$, then φ^ν_λ is asymptotically log-linear in the vicinity of $+\infty$*

$$\varphi^\nu_\lambda(x) \underset{\infty}{\sim} \alpha_2 \log x + \beta_2 \ ,$$

$$\textit{where} \quad \alpha_2 = \nu \quad \textit{and} \quad \beta_2 = -\nu \log \lambda - \frac{\nu}{2} \log \frac{\Gamma(1/\nu)}{\Gamma(3/\nu)} \ .$$

The proofs rely on a result of Berman (1992) [2] and is detailed in Appendix F.

Remark 2. *For $\nu > 2$, an asymptotic log-linear behavior with $\alpha_2 = 2$ and $\beta_2 = -\log 2$ can be obtained using exactly the same sketch of proof as the one of Theorem 4.*

Remark 3. *For $\nu = 2$, we have φ^2_λ is linear, $\beta_1 = -\log 2 - \log\left(1 + \lambda^2\right)$ and $\beta_2 = -\log 2 - \log \lambda^2$, which shows that Theorem 4 cannot hold true for $\nu = 2$.*

Remark 4. *For $\nu = 1$, Theorem 1 and Theorem 2 coincide with Theorem 3 and Theorem 4.*

Remark 5. *For $0 < \nu \leqslant \frac{2}{3}$, though we did not succeed in proving it, our numerical simulations also revealed a log-linear asymptotic behavior for $x \to \infty$ in perfect agreement with the expression of α_2 and β_2 given in Theorem 4.*

Again, the exponent $\alpha_1 = 2$ in the vicinity of 0 shows that the Gaussian distribution involved in the convolution prevails over the generalized Gaussian distribution and the behavior of $f^\nu_{\sigma,\lambda}$ is then quadratic. Similarly, the exponent $\alpha_2 = \nu$ in the vicinity of $+\infty$ shows that the generalized Gaussian distribution involved in the convolution prevails over the Gaussian distribution and the behavior of $f^\nu_{\sigma,\lambda}$ is then a power function of the form x^ν. These results are supported by Figure 8 that illustrates the resulting convolution, the discrepancy function and the log-discrepancy function for $\nu = 1.5$ and $\nu = .3$. Moreover, the discrepancy function $f^\nu_{\sigma,\lambda}$ with $\nu \leqslant 1$ shares a similar behavior with well-known robust M-estimator loss functions

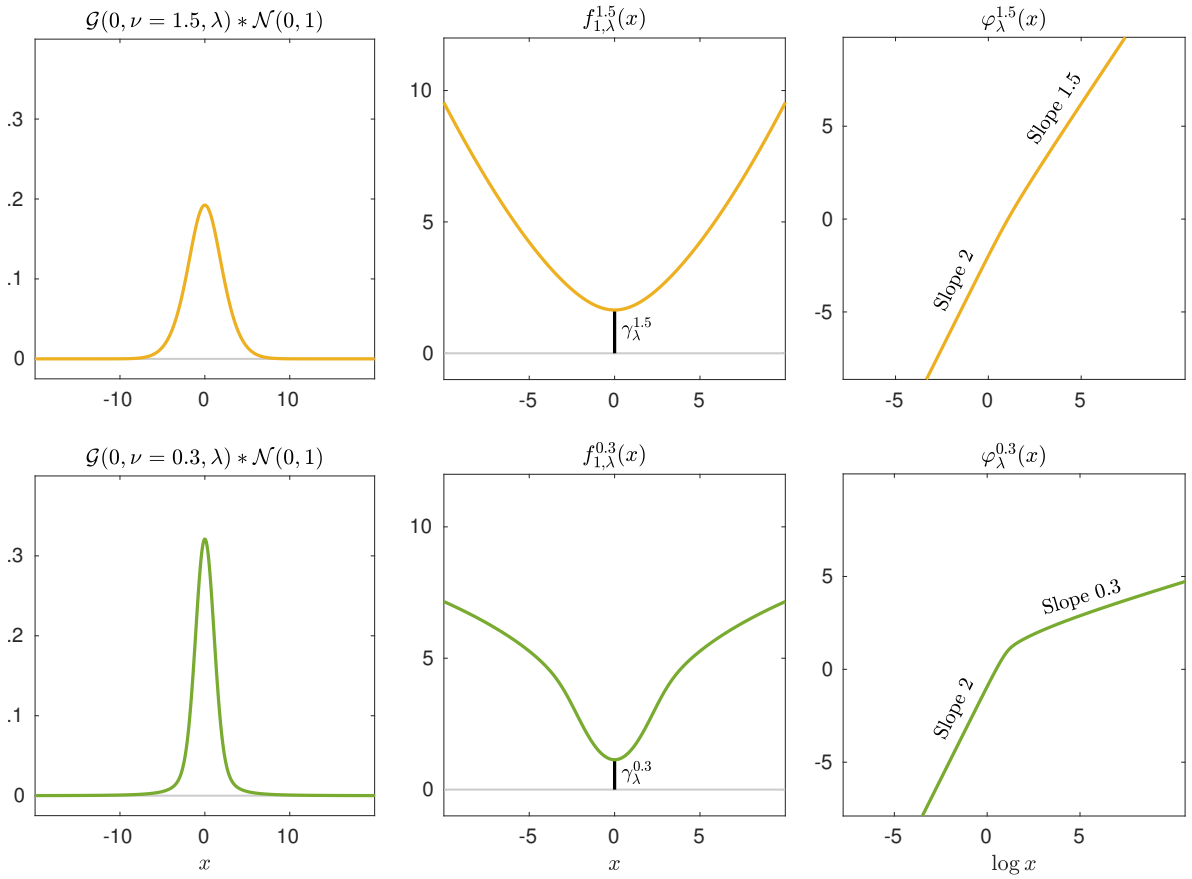


Figure 8. *From left to right: the convolution of a generalized Gaussian distribution with standard deviation $\lambda = 2$ with a Gaussian distribution with standard deviation $\sigma = 1$, the corresponding discrepancy function and log-discrepancy function. From top to bottom: the GGD has a shape parameter $\nu = 1.5$ and .3, respectively.*

[27]. In particular, the asymptotic case for $\nu \to 0$ resembles the Tukey's bisquare loss, known to be insensitive to outliers. This is again in complete agreement with GGD priors having larger tails as ν goes to 0.

Figure 9 shows the evolution of the log-discrepancy function for various values of ν in the context of three different signal-to-noise ratio λ/σ (SNR). One can observe that as the SNR decreases (resp., increases), the left (resp., right) asymptotic behavior starts dominating over the right (resp., left) asymptotes. In other words, for $\nu < 2$, the intersection of the two asymptotes goes to $+\infty$ (resp., $-\infty$). Last but not least, for $0 < \nu \leqslant 2$, the log-discrepancy function φ^{ν}_{λ} is always concave and since $\alpha_2 \leqslant \alpha_1$ it is thus upper-bounded by its left and right asymptotes.

From Theorem 3, Theorem 4 and Remark 5, we can now build two asymptotic log-linear approximations for φ^{ν}_{λ}, with $0 < \nu < 2$, and subsequently an asymptotic power approximation for $f^{\nu}_{\sigma,\lambda}$ by using the relation (27). Next, we explain the approximation process of the in-between behavior, as well as its efficient evaluation.

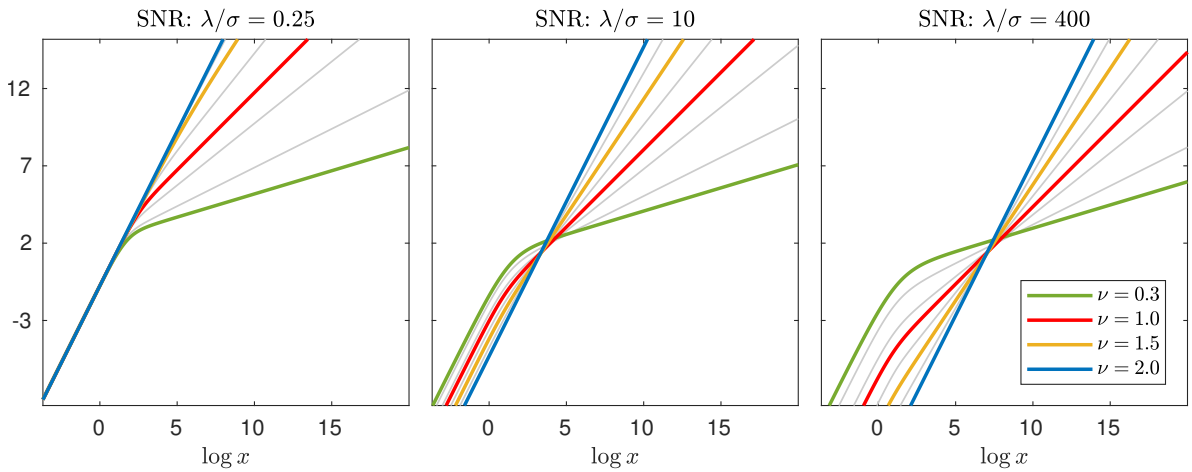


Figure 9. *Illustrations of the log-discrepancy function for various $0 < \nu < 2$ and SNR λ/σ.*

4.2. Numerical approximation. We now describe the proposed approximation of the discrepancy function $f^{\nu}_{1,\lambda}$ through approximating $\hat{\varphi}^{\nu}_{\lambda}$ of the log-discrepancy function as

$$\hat{f}^{\nu}_{1,\lambda}(x) = \gamma^{\nu}_{\lambda} + \exp \hat{\varphi}^{\nu}_{\lambda}(x) \quad \text{where} \quad \gamma^{\nu}_{\lambda} = f^{\nu}_{1,\lambda}(0) \ . \tag{38}$$

Based on our previous theoretical analysis, a solution preserving the asymptotic, increasing and concave behaviors of $\varphi^{\nu}_{1,\lambda}$ can be defined by making use of the following approximations

$$\hat{\varphi}^{\nu}_{\lambda}(x) = \alpha_1 \log|x| + \beta_1 - \texttt{rec}(\alpha_1 \log|x| + \beta_1 - \alpha_2 \log|x| - \beta_2) \ , \tag{39}$$

where `rec` is a so-called rectifier function that is positive, increasing, convex and satisfies

$$\lim_{x \to -\infty} \texttt{rec}(x) = 0 \quad \text{and} \quad \texttt{rec}(x) \underset{x \to \infty}{\sim} x \ . \tag{40}$$

In this paper, we consider the two following rectifying functions

$$\texttt{relu}(x) = \max(0, x) \quad \text{and} \quad \texttt{softplus}(x) = h \log\left[1 + \exp\left(\frac{x}{h}\right)\right], \ h > 0 \ , \tag{41}$$

as coined respectively in [40] and [18]. Using the function `relu` (Rectified linear unit) leads to an approximation $\hat{\varphi}^{\nu}_{\lambda}$ that is exactly equal to the asymptotes of φ^{ν}_{λ} with a singularity at their crossing point. In this paper, we will instead use the function `softplus` as it allows the approximation of φ^{ν}_{λ} to converge smoothly to the asymptotes without singularity. Its behavior is controlled by the parameter $h > 0$. The smaller the value of h is, the faster the convergence speed to the asymptotes.

The parameter h should be chosen such that the approximation error between $\hat{\varphi}^{\nu}_{\lambda}(x)$ and $\varphi^{\nu}_{\lambda}(x)$ is as small as possible. This can be done numerically by first evaluating $\varphi^{\nu}_{\lambda}(x)$ with integration techniques for a large range of values x, and then selecting the parameter h by least square. Of course, the optimal value for h depends on the parameter λ and ν.

Figure 10 gives an illustration of our approximations of the log-discrepancy and the corresponding distribution obtained with `relu` and `softplus`. On this figure the underlying

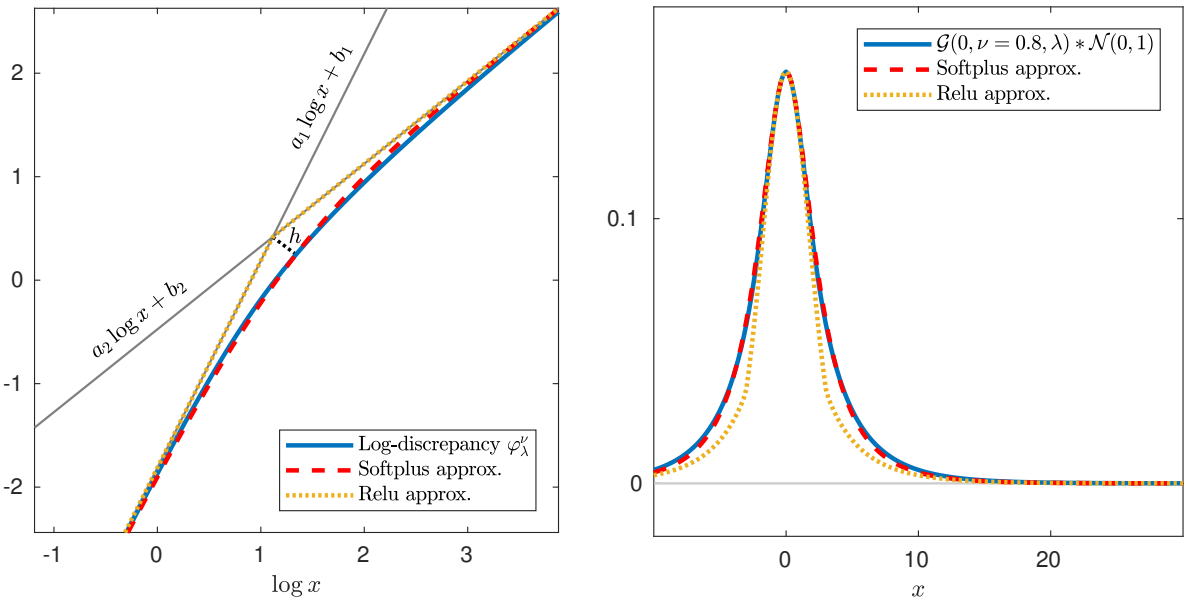


Figure 10. *Illustrations of our approximations of φ_λ^ν and the corresponding underlying posterior distribution $\mathcal{N}(0, \nu, \lambda) * \mathcal{G}(0, 1)$ (where $\nu = .8$ and $\lambda = 4$). The blue curves have been obtained by evaluating the convolution using numerical integration techniques for all x. The dashed curves are obtained using the proposed* `relu`*- and* `softplus`*-based approximations that have closed-form expressions.*

functions have been obtained by numerical integration for a large range of value of x. One can observe that using `softplus` provides a better approximation than `relu`.

Our approximation for $\hat{f}_{1,\lambda}^\nu(x)$ is parameterized by six scalar values: γ_ν^λ, α_1, β_1, α_2, β_2 and h that depend only on the original parameters λ and ν. From our previous analysis, we have that $\alpha_1 = 2$ and $\alpha_2 = \nu$. The other parameters are non-linear functions of λ and ν. The parameters γ_ν^λ, β_1 and β_2 require either performing numerical integration or evaluating the special function Γ. As discussed, the parameter h requires numerical integration for various x and then optimization. For these reasons, these values cannot be computed during runtime. Instead, we pre-compute these four parameters offline for $10,000$ different combinations of λ and ν values in the intervals $[10^{-3}, 10^3]$ and $[0.3, 2]$, respectively (the choice for this range will be motivated in Section 6). The resulting values are then stored in four corresponding look-up-tables. During runtime, these parameters are retrieved online by bi-linear extrapolation and interpolation. The four look-up-tables are given in Figure 11. We will see in Section 6 that using the approximation $\hat{f}_{1,\lambda}^\nu$ results in substantial acceleration without significant loss of performance as compared to computing $f_{1,\lambda}^\nu$ directly by numerical integration during runtime.

5. Shrinkage functions: analysis and approximations. Recall that from its definition given in eq. (18), the shrinkage function is defined for $\nu > 0$, $\sigma > 0$ and $\lambda > 0$, as

$$s_{\sigma,\lambda}^\nu(x) \in \operatorname*{argmin}_{t \in \mathbb{R}} \frac{(x-t)^2}{2\sigma^2} + \lambda_\nu^{-\nu} |t|^\nu \ . \tag{42}$$

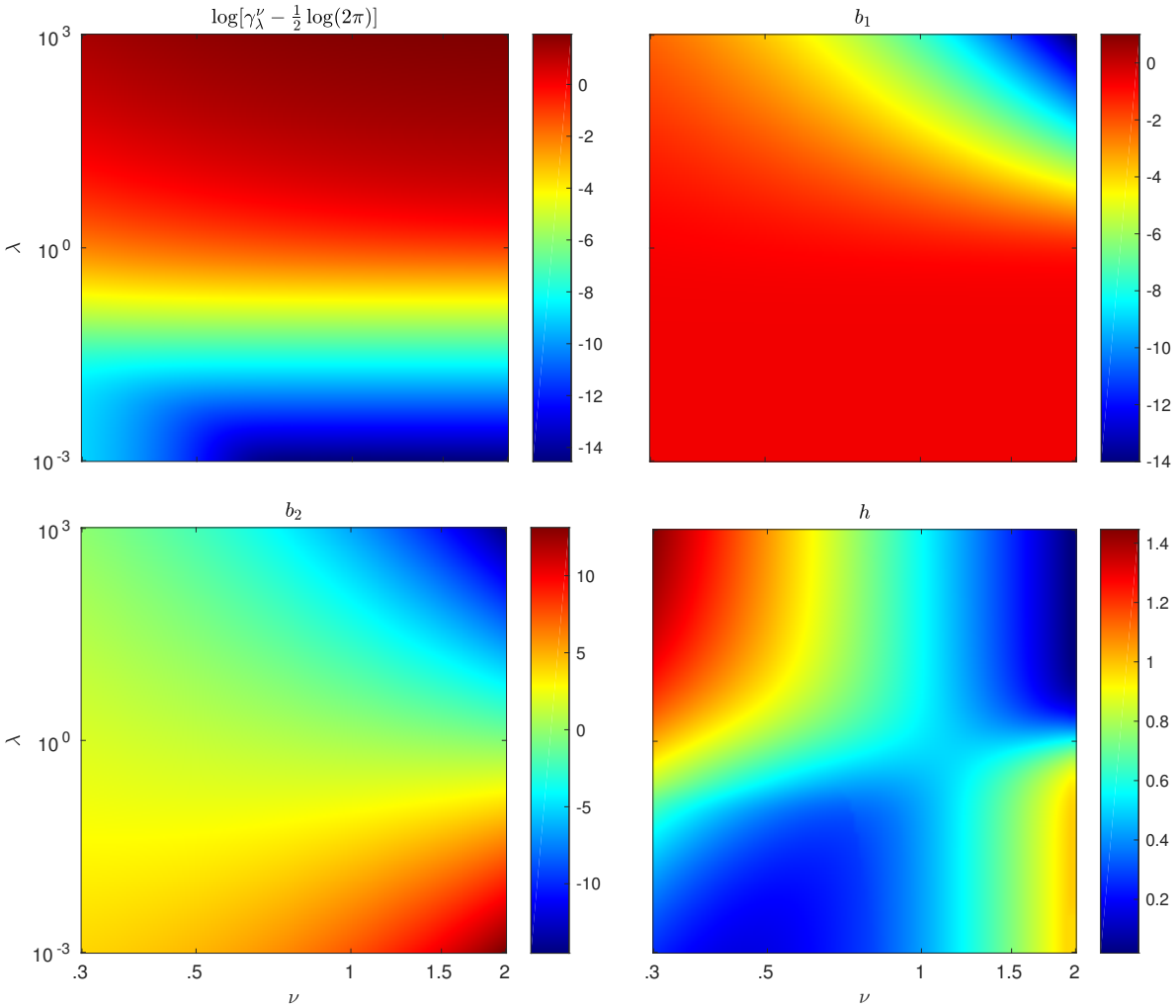


Figure 11. *Look-up-tables used to store the values of the parameters γ_λ^ν, b_1, b_2 and h for various $.3 \leqslant \nu \leqslant 2$ and $10^{-3} \leqslant \lambda \leqslant 10^3$. A regular grid of 100 values has been used for ν and a logarithmic grid of 100 values has been used for λ. This leads to a total of $10,000$ combinations for each of the four look-up-tables.*

5.1. Theoretical analysis. Except for some particular values of ν (see, Subsection 5.2), Problem (42) does not have explicit solutions. Nevertheless, as shown in [39], Problem (42) admits two (not necessarily distinct) solutions. One of them is implicitly characterized as

$$s_{\sigma,\lambda}^\nu(x) = \begin{cases} 0 & \text{if} \quad 0 < \nu \leqslant 1 \quad \text{and} \quad |x| \leqslant \tau_\lambda^\nu \ , \\ t^\star & \text{otherwise} \ , \end{cases} \tag{43}$$

$$\text{where} \quad t^\star = x - \operatorname{sign}(t^\star)\nu\sigma^2\lambda_\nu^{-\nu}|t^\star|^{\nu-1} \ ,$$

$$\text{and} \quad \tau_\lambda^\nu = \begin{cases} (2-\nu)(2-2\nu)^{-\frac{1-\nu}{2-\nu}}(\sigma^2\lambda_\nu^{-\nu})^{\frac{1}{2-\nu}} & \text{if} \quad \nu < 1 \ , \\ \sigma^2\lambda^{-1} & \text{otherwise } (\nu = 1) \ . \end{cases}$$

The other one is obtained by changing $|x| \leqslant \tau_\lambda^\nu$ to $|x| < \tau_\lambda^\nu$ in (43), and so they coincide for almost every $(x, \lambda, \sigma, \nu)$. As discussed in [39], for $\nu > 1$, $s_{\sigma,\lambda}^\nu(x)$ is differentiable, and for

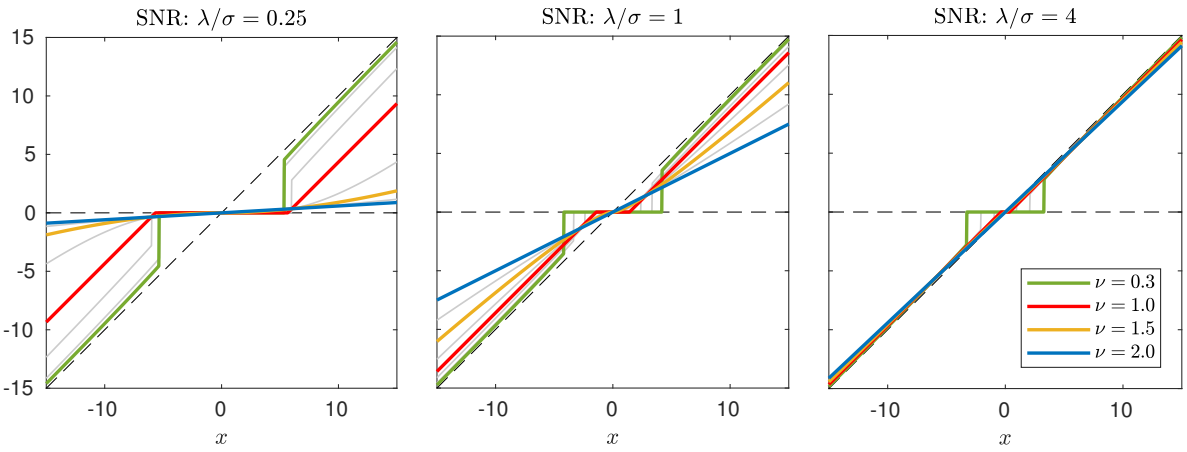


Figure 12. *Illustrations of the shrinkage function for various* $0 < \nu < 2$ *and SNR* λ/σ.

$\nu \leqslant 1$, the shrinkage exhibits a threshold τ_λ^ν that produces sparse solutions. Proposition 3 summarizes a few important properties.

Proposition 3. *Let* $\nu > 0$, $\sigma > 0$, $\lambda > 0$ *and* $s^\nu_{\sigma,\lambda}$ *as defined in eq.* (18). *The following relations hold true*

(reduction) $$s^\nu_{\sigma,\lambda}(x) = \sigma s^\nu_{1,\frac{\lambda}{\sigma}}\left(\frac{x}{\sigma}\right) ,$$

(odd) $$s^\nu_{\sigma,\lambda}(x) = -s^\nu_{\sigma,\lambda}(-x) ,$$

(shrinkage) $$s^\nu_{\sigma,\lambda}(x) \in \begin{cases} [0,x] & \text{if} \quad x \geqslant 0 \\ [x,0] & \text{otherwise} \end{cases} ,$$

(increasing with x) $$x_1 \geqslant x_2 \Leftrightarrow s^\nu_{\sigma,\lambda}(x_1) \geqslant s^\nu_{\sigma,\lambda}(x_2) ,$$

(increasing with λ) $$\lambda_1 \geqslant \lambda_2 \Leftrightarrow s^\nu_{\sigma,\lambda_1}(x) \geqslant s^\nu_{\sigma,\lambda_2}(x) ,$$

(kill low SNR) $$\lim_{\frac{\lambda}{\sigma}\to 0} s^\nu_{\sigma,\lambda}(x) = 0 ,$$

(keep high SNR) $$\lim_{\frac{\lambda}{\sigma}\to +\infty} s^\nu_{\sigma,\lambda}(x) = x .$$

The proofs can be found in Appendix G. These properties show that $s^\nu_{\sigma,\lambda}$ is indeed a shrinkage function (non-expansive). It shrinks the input coefficient x according to the model ν and the modeled signal to noise ratio $\frac{\lambda}{\sigma}$ (SNR). When x is small in comparison to the SNR, it is likely that its noise component dominates the underlying signal, and is, therefore, shrunk towards 0. Similarly, when x is large, it will likely be preserved. This is even more likely when ν is small, since in this case large coefficients are favored by the *prior*. Illustrations of shrinkage functions for various SNR and ν are given in Figure 12.

5.2. Numerical approximations. The shrinkage function $s^\nu_{\sigma,\lambda}$, implicitly defined in (43) does not have a closed form expression in general. Nevertheless, for fixed values of x, σ, λ, ν,

Table 1
Shrinkage function under generalized Gaussian priors

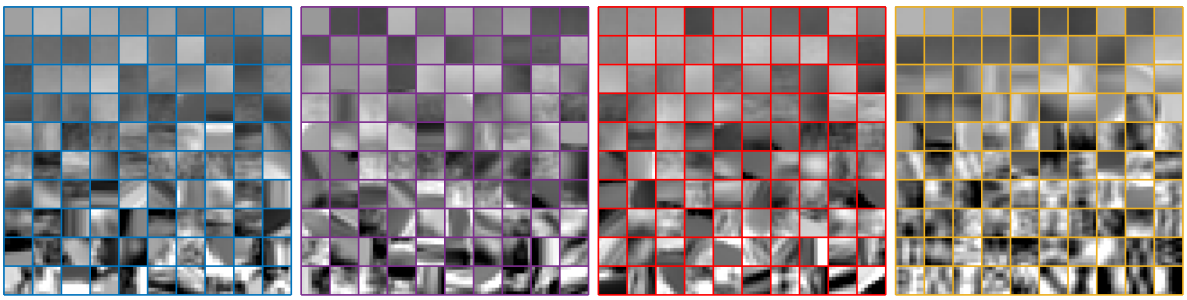
ν	Shrinkage $s^{\nu}_{\sigma,\lambda}(x)$	Remark
< 1	$\begin{cases} x - \gamma x^{\nu-1} + O(x^{2(\nu-1)}) & \text{if} \quad \lvert x\rvert \geqslant \tau^{\nu}_{\lambda} \\ 0 & \text{otherwise} \end{cases}$	$\approx$ Hard-thresholding [39]
1	$\operatorname{sign}(x)\max\left(\lvert x\rvert - \dfrac{\sqrt{2}\sigma^2}{\lambda}, 0\right)$	Soft-thresholding [16]
4/3	$x + \gamma\left(\sqrt[3]{\dfrac{\zeta - x}{2}} - \sqrt[3]{\dfrac{\zeta + x}{2}}\right)$	[7]
3/2	$\operatorname{sign}(x)\dfrac{\left(\sqrt{\gamma^2 + 4\lvert x\rvert} - \gamma\right)^2}{4}$	[7]
2	$\dfrac{\lambda^2}{\lambda^2 + \sigma^2} \cdot x$	Wiener (LMMSE)

with $\gamma = \nu\sigma^2\lambda_{\nu}^{-\nu}$ and $\zeta = \sqrt{x^2 + 4\left(\dfrac{\gamma}{3}\right)^3}$.

$s^{\nu}_{\sigma,\lambda}(x)$ can be estimated using iterative solvers such as Newton descent or Halleys root-finding method. These approaches converge quite fast and, in practice, reaches a satisfying solution within ten iterations. However, since in our application of interest we need to evaluate this function a large number times, we will follow a different path in order to reduce computation time (even though we have implemented this strategy).

As discussed earlier, $s^{\nu}_{\sigma,\lambda}$ is known in closed form for some values of ν, more precisely: $\nu = \{1, 4/3, 3/2, 2\}$ (as well as $\nu = 3$ but this is out of the scope of this study), see for instance [7]. When $\nu = 2$, we retrieve the linear minimum mean square estimator (known in signal processing as Wiener filtering) and related to Tikhonov regularization and ridge regression. This shrinkage is linear and the slope of the shrinkage goes from 0 to 1 as the SNR increases (see Figure 12). When $\nu = 1$, the shrinkage is the well-known soft-thresholding [16], corresponding to the maximum a posteriori estimator under a Laplacian prior. When $\nu < 1$, the authors of [39] have shown that (i) the shrinkage admits a threshold with a closed-form expression (given in eq. (43)), and (ii) the shrinkage is approximately equal to hard-thresholding with an error term that vanishes when $|x| \to \infty$. All these expressions are summarized in Table 1.

In order to keep our algorithm as fast as possible, we propose to use the approximation of the shrinkage given for $\nu < 1$ in [39]. Otherwise, we pick one of the four shrinkage functions corresponding to $\nu = \{1, 4/3, 3/2, 2\}$ by nearest neighbor on the actual value of $\nu \in [1, 2]$. Though this approximation may seem coarse compared to the one based on iterative solvers,



(a) GMM ($\nu = 2$) (b) GGMM ($.3 \leqslant \nu \leqslant 2$) (c) LMM ($\nu = 1$) (d) HLMM ($\nu = .5$)

Figure 13. *Set of 100 patches, sorted by the norm of their gradient, and generated to be independently distributed according to (from left to right) a GMM, GGMM, LMM and HLMM. For ease of visualization, only the top eigendirections corresponding to* 80% *of the variance have been chosen. Near-constant patches with variance smaller than* $\frac{2}{P}\|\mathbf{\Sigma}_k\|_F^2$ *have also been discarded.*

we did not observe any significant loss of quality in our numerical experiments (see Section 6). Nonetheless, this alternative leads to 10 times speed-up while evaluating shrinkage.

6. Experimental evaluation. In this section we explain the methodology used to learn and validate the GGMM model, and present numerical experiments to compare the performance of the proposed GGMM model over existing GMM-based image denoising algorithms. To demonstrate the superiority of allowing for a flexible GGMM model, we also present results using GGMM models with fixed shape parameters, $\nu = 1$ (Laplacian mixture model) and $\nu = 0.5$ (Hyper-Laplacian mixture model).

Learning. For consistency purposes, we keep the training data and the number of mixture components in the models the same as that used in the original EPLL algorithm [62]. Specifically, we train our models on 2 million clean patches extracted from Berkeley Segmentation Dataset (BSDS) [35]. We learn $K = 200$ zero-mean generalized Gaussian mixture components from patches of size 8×8. Parameter estimation is carried out using the Expectation-Maximization (EM) algorithm [13], that is known to monotonically increase the likelihood and converge to a local minimum. For applying EM to GGMM learning, we leverage standard strategies used for parameter estimation for GGD and/or GGMM that are reported in previous works [33, 3, 38, 45, 31]. We opted for a warm-start training by initializing our GGMM model with the GMM model from [62] and with initial values of shape parameters as 2. The EM algorithm is run until convergence or for a preset number of iterations (*e.g.*, 100 iterations). We noticed that a shape parameter $\nu_j^{(k)} < .3$, leads to numerical issues and $\nu_j^{(k)} > 2$ leads to a local minima with several degenerate components. For this reason, we impose the constraint that the learned shape parameters satisfy $\nu_j^{(k)} \in [.3, 2]$. This observation is consistent with earlier works that have attempted to learn GGMM shape parameter from data [50]. Lastly, for learning Laplacian mixture model (LMM) and hyper-Laplacian mixture model (HLMM), we use the same procedure described above but force all shape parameters to be equal to 1 or 0.5, respectively.

Model validation. As discussed in Section 3, Figure 5 illustrates the validity of our model choices with histograms of different dimensions of a single patch cluster. It clearly shows the

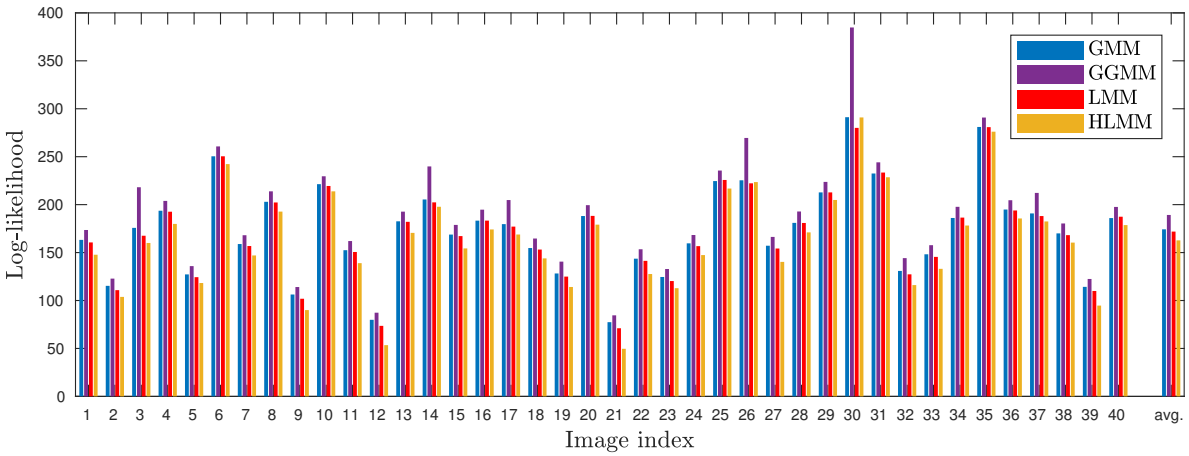


Figure 14. *Average log-likelihood of all non-overlapping patches (with subtracted mean) of each of the* 40 *images of our validation subset of the testing BSDS dataset for the GMM, GGMM, LMM and HLMM. The total average over the* 40 *images is shown in the last column.*

importance of allowing the shape and scale parameter to vary across dimensions for capturing underlying patch distributions.

Since GGMM (and obviously, GMM) falls into the class of *generative* models, one can also assess the *expressivity* of a model by analyzing the variability of generated patches and its ability to generate relevant image features (edges, texture elements etc.). This can be tested by selecting a component k of the GGMM (or GMM) with probability w_k and sampling patches from the GGD (or GD) as described in [41]. Figure 13 presents a collage of 100 patches independently generated by this procedure using GMM, GGMM, LMM and HLMM. As observed, patches generated from GGMM show greater balance between strong/faint edges, constant patches and subtle textures than the models that use constant shape parameters such as GGM, LMM and HLMM.

The superiority of our GGMM model over GMM, LMM or HLMM models can also be illustrated by comparing the log-likelihood (LL) achieved by these models over a set of clean patches from natural images. Note that, to maintain objectivity, the models have to be tested on data that is different than the dataset used during training. To this end, we compute the LL of the four above-mentioned models on all non-overlapping patches of 40 randomly selected images extracted from BSDS testing set [35], which is a different set than the training images used in the EM algorithm (parameter estimation/model learning). One can observe that not only GGMM is a better fit than GMM, LMM and HLMM on average for the 40 images, but it is also a better fit on each single image.

Denoising evaluation. Following the verification of the model, we provide a thorough evaluation of our GGMM prior in denoising task by comparing its performance against EPLL that uses a GMM prior [62] and with our LMM and HLMM models explained above. For the ease of comparison, we utilize the pipeline that was prescribed for the original EPLL [62] algorithm. Specifically, we subject the noisy images to 5 iterations of the denoising procedure with the estimate from one iteration fed in as the input for the next one. The iterative procedure is initialized using first estimate $\hat{\boldsymbol{x}} = \boldsymbol{y}$ and with the parameter β set to $\frac{1}{\sigma^2}\{1, 4, 8, 16, 32\}$ for

Table 2

Image denoising performance comparison of EPLL algorithm with different priors. PSNR and SSIM values are obtained on the BSDS test set (average over 60 images), and on six standard images corrupted with 3 different levels of noise. BM3D algorithm results are also included for reference purposes.

σ	Algo.	BSDS	barbara	cameraman	hill	house	lena	mandrill	Avg.
		PSNR							
5	BM3D	37.33	38.30	38.33	36.03	39.79	38.70	35.24	37.37
	GMM	37.26	37.58	38.13	**35.92**	38.83	38.50	35.21	37.27
	GGMM	**37.33**	37.72	**38.18**	**35.92**	**38.91**	**38.52**	**35.22**	**37.34**
	LMM	37.31	**37.83**	38.15	35.87	38.90	38.49	35.17	37.32
	HLMM	36.85	37.42	37.70	35.40	38.33	38.08	34.75	36.86
20	BM3D	29.38	31.81	30.39	28.59	33.79	33.02	26.62	29.50
	GMM	29.35	29.74	30.10	28.45	32.77	32.37	26.62	29.41
	GGMM	**29.43**	30.08	**30.22**	**28.50**	33.01	32.59	**26.68**	**29.50**
	LMM	29.29	**30.21**	30.01	28.38	**33.23**	**32.72**	26.46	29.37
	HLMM	28.48	29.34	29.00	27.71	32.54	32.11	25.47	28.56
60	BM3D	24.81	26.26	25.39	24.49	28.73	28.16	21.71	24.90
	GMM	24.54	23.78	25.10	24.24	27.33	27.19	**21.56**	24.57
	GGMM	**24.63**	**23.96**	**25.27**	**24.28**	27.65	27.49	21.48	**24.67**
	LMM	24.55	23.88	25.10	24.27	**27.73**	**27.55**	21.33	24.59
	HLMM	23.95	23.17	23.80	23.86	26.97	26.98	20.65	23.98
		SSIM							
5	BM3D	.9619	.9643	.9617	.9507	.9567	.9435	.9585	.9614
	GMM	.9627	.9617	**.9608**	**.9511**	**.9474**	**.9435**	**.9598**	**.9619**
	GGMM	**.9628**	**.9618**	.9604	.9503	.9461	.9424	.9592	**.9619**
	LMM	.9620	.9615	.9599	.9488	.9443	.9408	.9576	.9611
	HLMM	.9562	.9571	.9570	.9408	.9359	.9352	.9501	.9553
20	BM3D	.8236	.9050	.8687	.7791	.8729	.8769	.7962	.8260
	GMM	**.8313**	.8691	.8681	**.7813**	.8582	.8641	**.8048**	**.8322**
	GGMM	.8296	**.8728**	**.8691**	.7777	.8614	.8679	.8010	.8307
	LMM	.8176	.8729	.8599	.7639	**.8638**	**.8687**	.7849	.8192
	HLMM	.7884	.8502	.8426	.7325	.8577	.8594	.7383	.7907
60	BM3D	.6373	.7562	.7572	.5802	.7917	.7753	.5012	.6424
	GMM	**.6205**	.6455	.7172	**.5675**	.7386	.7281	**.5014**	**.6232**
	GGMM	.6176	**.6512**	**.7318**	.5586	.7564	.7416	.4769	.6208
	LMM	.6120	.6475	.7299	.5577	**.7596**	**.7443**	.4614	.6155
	HLMM	.5775	.6117	.7025	.5268	.7487	.7338	.3843	.5812

each iteration, respectively. To reduce the computation time of *all* EPLL-based algorithms, we utilize the random patch overlap procedure introduced by [44]. That is, instead of extracting all patches at each iteration, a randomly selected but different subset of overlapping patches consisting of only 3% of all possible patches is processed in each iteration. For the

sake of reproducibility, we made our MATLAB/MEX-C implementation available online at https://bitbucket.org/cdeledalle/ggmm-epll.

The evaluation is carried out on classical images such as *Barbara*, *Cameraman*, *Hill*, *House*, *Lena* and *Mandrill*, and on 60 images taken from BSDS testing set [35] (the original BSDS test data contains 100 images, the other 40 were used for model validation experiments). All image have been corrupted independently by additive white Gaussian noise with standard deviation $\sigma = 5, 20$ and 60 (with pixel values between $[0, 255]$). The EPLL algorithm using mixture of Gaussian, generalized Gaussian, Laplacian and hyper-Laplacian priors are indicated as GMM, GGMM, LMM and HLMM in Table 2. BM3D algorithm [9] also included for reference purposes. To stay with the focus of this paper, *i.e.*, on the effect of image priors on EPLL-based algorithms, BM3D will be excluded from our performance comparison discussions. The denoising performance of the algorithms are measured in terms of Peak Signal to Noise Ratio (PSNR) and Structural SIMilarity (SSIM) [58]. As can be observed in Table 2, in general, GGMM prior provides better PSNR performance than the three other priors. In terms of SSIM, GGMM prior is comparable to GMM. The differences in denoising performance can also be verified visually in Figure 15, Figure 16 and Figure 17. The denoised images obtained using GGMM prior show much fewer edge artifacts as compared to GMM-EPLL results. On the other hand, GGMM prior is also able to better preserve textures than LMM and HLMM.

Prior fitness for image denoising. In this work, we have considered non-blind image denoising. That is, the noise standard deviation is assumed to be known. In this setting, if the restoration model is accurate, one should ideally achieve optimal restoration performance when using the true degradation. To verify this, we conducted a denoising task with image corrupted with noise with standard deviation $\sigma = 20$. We used GMM, LMM and GGMM priors in the restoration framework with assumed σ values ranging from 15 to 30. Figure 18 shows the evolution of average restoration performance over 40 images from BSDS testing set (kept aside for validation, as mentioned above) with varying noise variances. GGMM prior attains its best performance when the noise variance used in the restoration model matches with the ground truth $\sigma = 20$. In contrast to GGMM, GMM (resp., LMM) reaches its best performance at a larger (resp., lower) value of σ than the correct noise used during degradation. This is because GMM tends to under-smooth clean patches (resp., over-smooth) so that a larger (resp., lower) value of σ is required to compensate the mismatch between the assumed prior and the actual distribution in the restoration model. This indicates that GGMM is a better option to model distribution of image patches than GMM or LMM.

Influence of our approximations. All previous experiments using GGMM patch priors were conducted based on the two proposed approximations introduced in Section 4 and Section 5. In Figure 19, we provide a quantitative illustration of the speed-ups provided by these approximations and their effect on denoising performance. Figure 19.(a) shows the result obtained by calculating original discrepancy function via numerical integration and the shrinkage function via Halley's root-finding method. This makes the denoising process extremely slow and takes 6 hours and 20 minutes for denoising an image of size 128×128 pixels. The approximated discrepancy function provides 4 orders of magnitude speed-up with no perceivable drop in performance (Figure 19.(b)). In addition, incorporating the shrinkage approximation provides further acceleration that allows the denoising to complete in less than 2 seconds with a very minor drop in PSNR/SSIM. Although the shrinkage approximation provides an acceler-

ation of ten-fold to the shrinkage calculation step, please note that the 10× speed-up is not reflected in the denoising process due to the major bottleneck caused by discrepancy function calculation. The approximately 15,000× speed-up realized without any perceivable drop in denoising performance underscores the efficacy of our proposed approximations.

7. Conclusions and Discussion. In this work, we suggest using a mixture of generalized Gaussians for modeling the patch distribution of clean images. We then provide a detailed study of the challenges that one encounters when using a highly flexible GGMM prior for image restoration in place of a more common GMM prior. We identify the two main bottlenecks in the restoration procedure using EPLL and GGMM namely the patch classification step and the shrinkage step. One of the main contributions of this paper, is the thorough theoretical analysis of the classification problem allowing us to introduce an asymptotically accurate approximation being computationally efficient. In order to tackle the shrinkage step, we collate and extend the existing solutions for the shrinkage function under GGMM prior.

Our numerical experiments indicate that our flexible patch GGMM is a better fit for each single image than GMM and other mixtures with constant shape parameters such as LMM or HLMM. In image denoising tasks, we have shown that using GGMM priors often outperforms GMM when used in the EPLL framework. Nevertheless, we believe the performance of GGMM prior in these scenarios falls short of its expected potential. Given GGMM is systematically a better prior than GMM, one would expect the GGMM-EPLL to outperform GMM-EPLL consistently. We postulate that even though the GGMM prior is improving the quality of the global solution, the half quadratic splitting strategy used in EPLL has no guarantee to return a better solution due to the non-convexity of the underlying problem. For this reason, we will focus our future work in designing specific optimization strategies for GGMM-EPLL leveraging the better expressivity of the model. Also, preliminary results on inverse problems indicates that degradation- and prior-specific algorithmic design should be employed to make GGMM-EPLL more competitive against GMM-EPLL (otherwise they both are on par). The development of such a formulation designed to adapt according to the properties of chosen priors and presented degradations is still under investigation.

Acknowledgments. The authors would like to thank Charles Dossal and Nicolas Papadakis for fruitful discussions.

Appendix A. Proof of equation (26).

Proof. For $\nu = 1$, using $\Gamma(1) = 1$ and $\Gamma(3) = 2$, we obtain

$$f^1_{\sigma,\lambda}(x) = \log(2\sqrt{\pi}\sigma\lambda) - \log \int_{-\infty}^{\infty} e^{-\frac{t^2}{2\sigma^2} - \frac{\sqrt{2}|x-t|}{\lambda}} \, \mathrm{d}t \;, \tag{44}$$

$$= \log(2\sqrt{\pi}\sigma\lambda) - \log \left[e^{-\frac{\sqrt{2}x}{\lambda}} \int_{-\infty}^{x} e^{-\frac{t^2}{2\sigma^2} + \frac{\sqrt{2}t}{\lambda}} \, \mathrm{d}t + e^{\frac{\sqrt{2}x}{\lambda}} \int_{x}^{\infty} e^{-\frac{t^2}{2\sigma^2} - \frac{\sqrt{2}t}{\lambda}} \, \mathrm{d}t \right] . \tag{45}$$

Since $\mathrm{erf}'(t) = \frac{2e^{-t^2}}{\sqrt{\pi}}$, it follows that for $a > 0$ and $b > 0$

$$\frac{\partial}{\partial t} \left[-\frac{\sqrt{\pi a}}{2} e^{\frac{a}{4b^2}} \mathrm{erf}\left(-\frac{t}{\sqrt{a}} - \frac{\sqrt{a}}{2b} \right) \right] = e^{-\frac{t^2}{a} - \frac{t}{b}} \;. \tag{46}$$

(a) Reference $\boldsymbol{x}$

(b) Noisy $\boldsymbol{y}$

(c) GMM ($\nu = 2$)

(d) GGMM ($.3 \leqslant \nu \leqslant 2$)

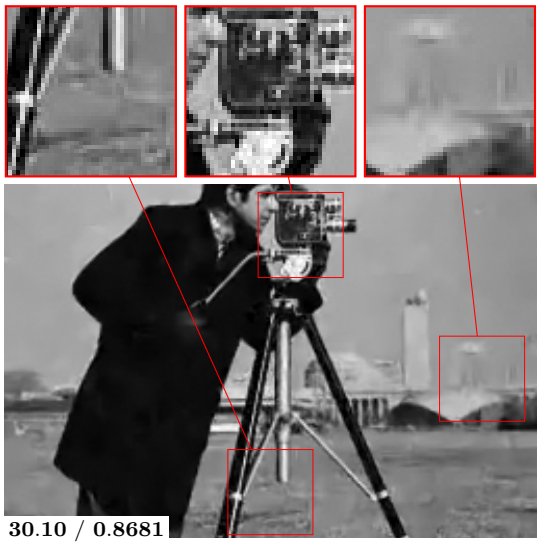
(e) LMM ($\nu = 1$)

(f) HLMM ($\nu = 0.5$)

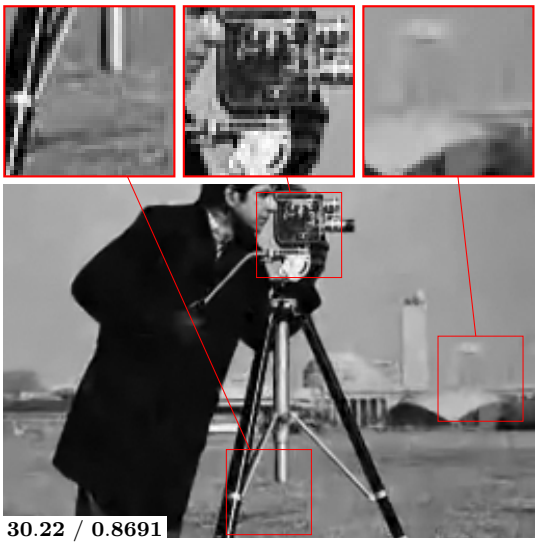
Figure 15. *(a) Close in on the image Castle from the BSDS testing dataset, (b) a noisy version degraded by additive white Gaussian noise with standard deviation $\sigma = 20$ and (c)-(f) results of EPLL under four patch priors: GMM, GGMM, LMM and HLMM, respectively. PSNR and SSIM are given in the bottom-left corner.*

(a) Reference $\boldsymbol{x}$

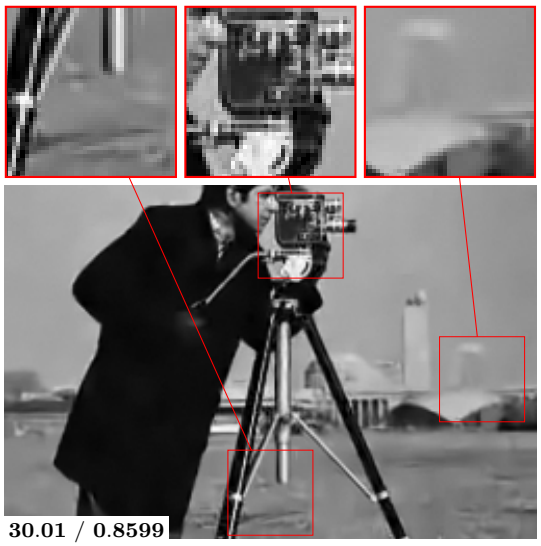
(b) Noisy $\boldsymbol{y}$



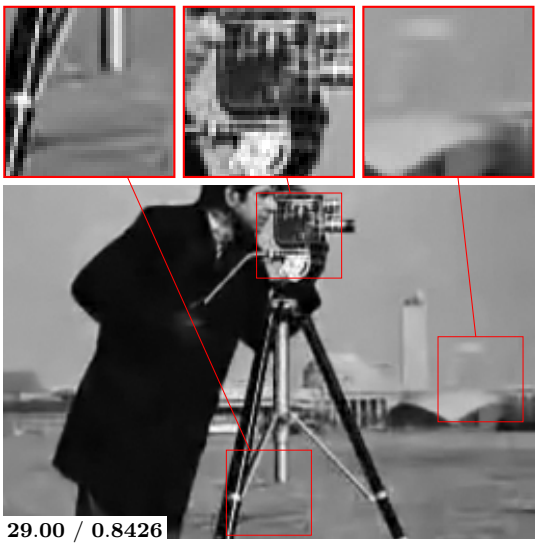
(c) GMM ($\nu = 2$)



(d) GGMM ($.3 \leqslant \nu \leqslant 2$)



(e) LMM ($\nu = 1$)



(f) HLMM ($\nu = 0.5$)

Figure 16. *(a) Close in on the standard image Cameraman. (b) a noisy version degraded by additive white Gaussian noise with standard deviation $\sigma = 20$ and (c)-(f) results of EPLL under four patch priors: GMM, GGMM, LMM and HLMM, respectively. PSNR and SSIM are given in the bottom-left corner.*

(a) Reference $\boldsymbol{x}$

(b) Noisy $\boldsymbol{y}$

(c) GMM ($\nu = 2$)

(d) GGMM ($.3 \leqslant \nu \leqslant 2$)

(e) LMM ($\nu = 1$)

(f) HLMM ($\nu = 0.5$)

Figure 17. *(a) Close in on the standard image Barbara. (b) a noisy version degraded by additive white Gaussian noise with standard deviation $\sigma = 20$ and (c)-(f) results of EPLL under four patch priors: GMM, GGMM, LMM and HLMM, respectively. PSNR and SSIM are given in the bottom-left corner.*

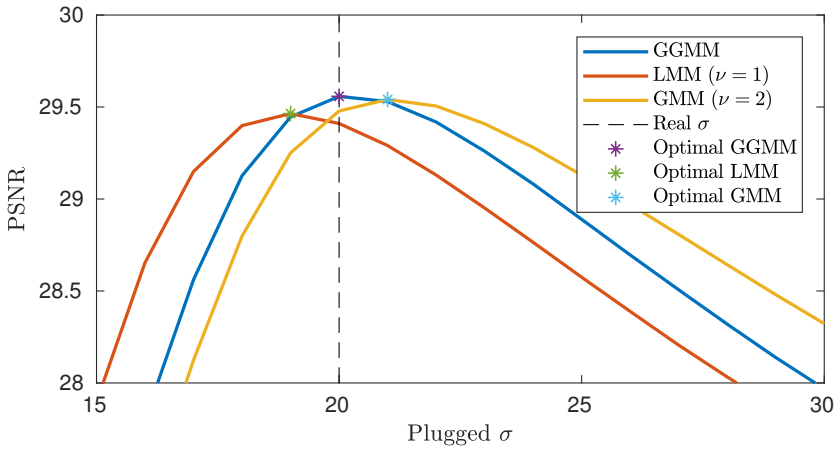


Figure 18. *Evolution of performance of EPLL with a GGMM, LMM ($\nu = 1$) and a GMM ($\nu = 2$) under misspecification of the noise standard deviation σ. Performances are measured in terms of PSNR on the BSDS dataset corrupted by a Gaussian noise with standard deviation $\sigma = 20$. For each of the three priors, EPLL has been run assuming σ was ranging from 15 to 30.*

Therefore we have with $a = 2\sigma^2$ and $b = \lambda/\sqrt{2}$

$$\int_{-\infty}^{x} e^{-\frac{t^2}{2\sigma^2}+\frac{\sqrt{2}t}{\lambda}}\,\mathrm{d}t = \frac{-\sqrt{\pi}\sigma e^{\frac{\sigma^2}{\lambda^2}}}{\sqrt{2}}\left[\operatorname{erf}\left(-\frac{t}{\sqrt{2}\sigma}+\frac{\sigma}{\lambda}\right)\right]_{-\infty}^{x} = \frac{\sqrt{\pi}\sigma e^{\frac{\sigma^2}{\lambda^2}}}{\sqrt{2}}\operatorname{erfc}\left(-\frac{x}{\sqrt{2}\sigma}+\frac{\sigma}{\lambda}\right)\,, \tag{47}$$

since $\lim_{t\to\infty} \operatorname{erf}(t) = 1$ and $\operatorname{erfc}(t) = 1 - \operatorname{erf}(t)$. Similarly, we get

$$\int_{x}^{\infty} e^{-\frac{t^2}{2\sigma^2}-\frac{\sqrt{2}t}{\lambda}}\,\mathrm{d}t = \frac{-\sqrt{\pi}\sigma e^{\frac{\sigma^2}{\lambda^2}}}{\sqrt{2}}\operatorname{erfc}\left(\frac{x}{\sqrt{2}\sigma}+\frac{\sigma}{\lambda}\right)\,. \tag{48}$$

Plugging these two last expressions in (45) and rearranging the terms conclude the proof. ∎

Appendix B. Proof of Proposition 2.

Proof. Starting from the definition of $f^{\nu}_{\sigma,\lambda}$ and using the change of variable $t \to \sigma t$, eq. (reduction) follows as

$$\begin{aligned} f^{\nu}_{\sigma,\lambda}(x) &= -\log\int_{\mathbb{R}} \sigma\frac{\kappa}{2\lambda_\nu}\exp\left[-\left(\frac{\sigma|t|}{\lambda_\nu}\right)^{\nu}\right]\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\sigma t)^2}{2\sigma^2}\right]\,\mathrm{d}t\,, \\ &= \log\sigma - \log\int_{\mathbb{R}} \frac{\kappa}{2\lambda_\nu/\sigma}\exp\left[-\left(\frac{|t|}{\lambda_\nu/\sigma}\right)^{\nu}\right]\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{(x/\sigma-t)^2}{2}\right]\,\mathrm{d}t = \log\sigma + f^{\nu}_{1,\lambda/\sigma}\left(\frac{x}{\sigma}\right)\,. \end{aligned} \tag{49}$$

Properties (even) and (unimodality) hold since the convolution of two real even unimodal distributions is even unimodal [59, 48]. Property (lower bound at 0) follows from (even), (unimodality) and the fact that the convolution of continuous and bounded real functions are continuous and bounded. ∎

(a) No approx. (6h20) (b) Approx. discrepancy (2s10) (c) Approx. disc. and shrink. (1s53)

Figure 19. *Results obtained by GGMM-EPLL on a* 128×128 *cropped image of the BSDS testing dataset damaged by additive white Gaussian noise with* $\sigma = 20$*. These are obtained respectively by (a) evaluating the classification and shrinkage problem with numerical solvers (numerical integration and Halley's root-finding method), (b) approximating the classification problem only, (c) approximating both problems. PSNR and SSIM are given in the bottom-left corner. Running time are indicated on the captions: our accelerations lead to a speed-up of* $\times 15,000$ *and* $\times 1.4$ *respectively.*

Appendix C. Proof of Theorem 1.

Lemma C.1. *Let* $a > 0$ *and* $b > 0$*. For* x *in the vicinity of* 0*, we have*

$$\frac{1}{2abx}\log\left[\frac{\operatorname{erfc}(ax+b)+e^{-4abx}\operatorname{erfc}(-ax+b)}{2\operatorname{erfc}(b)}\right] = -1+\left(ab-\frac{ae^{-b^2}}{\operatorname{erfc}(b)\sqrt{\pi}}\right)x+o(x)\ . \tag{50}$$

Proof. Since $\operatorname{erfc}'(x) = -\frac{2e^{-x^2}}{\sqrt{\pi}}$ and $\operatorname{erfc}''(x) = \frac{2xe^{-x^2}}{\sqrt{\pi}}$, using second order Taylor's expansion for x in the vicinity of 0, it follows that

$$\operatorname{erfc}(ax+b) \underset{0}{\sim} \operatorname{erfc}(b) - \frac{2ae^{-b^2}}{\sqrt{\pi}}x + \frac{2a^2be^{-b^2}}{\sqrt{\pi}}x^2\ , \tag{51}$$

$$\text{and}\quad \operatorname{erfc}(-ax+b) \underset{0}{\sim} \operatorname{erfc}(b) + \frac{2ae^{-b^2}}{\sqrt{\pi}}x + \frac{2a^2be^{-b^2}}{\sqrt{\pi}}x^2\ . \tag{52}$$

We next make the following deductions

$$e^{-4abx}\operatorname{erfc}(-ax+b)) \underset{0}{\sim} e^{-4abx}\operatorname{erfc}(b) + e^{-4abx}\frac{2ae^{-b^2}}{\sqrt{\pi}}x + e^{-4abx}\frac{2a^2be^{-b^2}}{\sqrt{\pi}}x^2\ , \tag{53}$$

$$\operatorname{erfc}(ax+b)+e^{-4abx}\operatorname{erfc}(-ax+b)) \underset{0}{\sim} (1+e^{-4abx})\left(\operatorname{erfc}(b)+\frac{2a^2be^{-b^2}}{\sqrt{\pi}}x^2\right)-(1-e^{-4abx})\frac{2ae^{-b^2}}{\sqrt{\pi}}x\ ,$$

$$\underbrace{\frac{\operatorname{erfc}(ax+b)+e^{-4abx}\operatorname{erfc}(-ax+b))}{2\operatorname{erfc} b}}_{A(x)} \underset{0}{\sim} \frac{1+e^{-4abx}}{2}\left(1+\frac{2a^2be^{-b^2}}{\operatorname{erfc}(b)\sqrt{\pi}}x^2\right)-(1-e^{-4abx})\frac{ae^{-b^2}}{\operatorname{erfc}(b)\sqrt{\pi}}x\ .$$

The left-hand side $A(x)$ of this last equation is then, in the vicinity of $x = 0$, equals to

$$A(x) = \frac{1 + e^{-4abx}}{2}\left(1 + \frac{2a^2be^{-b^2}}{\operatorname{erfc}(b)\sqrt{\pi}}x^2\right) - (1 - e^{-4abx})\frac{ae^{-b^2}}{\operatorname{erfc}(b)\sqrt{\pi}}x + o(x^2)\ . \tag{54}$$

We next use second-order Taylor's expansion for e^{-4abx}, leading to

$$\begin{aligned} A(x) &= (1 - 2abx + 4a^2b^2x^2 + o(x^2))\left(1 + \frac{2a^2be^{-b^2}}{\operatorname{erfc}(b)\sqrt{\pi}}x^2\right) \\ &\qquad - (4abx - 8a^2b^2x^2 + o(x^2))\frac{ae^{-b^2}}{\operatorname{erfc}(b)\sqrt{\pi}}x + o(x^2)\ , \end{aligned} \tag{55}$$

$$= 1 - 2abx + \left(4a^2b^2 - \frac{2a^2be^{-b^2}}{\operatorname{erfc}(b)\sqrt{\pi}}\right)x^2 + o(x^2)\ . \tag{56}$$

By using the second-order Taylor's expansion of $\log(1 + x)$, it follows that

$$\log\left[A(x)\right] = -2abx + \left(4a^2b^2 - \frac{2a^2be^{-b^2}}{\operatorname{erfc}(b)\sqrt{\pi}}\right)x^2 - 2a^2b^2x^2 + o(x^2)\ . \tag{57}$$

Dividing both sides by $2abx$ then concludes the proof,

$$\frac{1}{2abx}\log\left[\frac{\operatorname{erfc}(ax + b) + e^{-4abx}\operatorname{erfc}(-ax + b))}{2\operatorname{erfc} b}\right] = -1 + \left(ab - \frac{ae^{-b^2}}{\operatorname{erfc}(b)\sqrt{\pi}}\right)x + o(x)\ . \tag{58}$$

■

Proof of Theorem 1. We first rewrite $\varphi^1_\lambda(x)$ as

$$\varphi^1_\lambda(x) = \log\left[\log\left[2\operatorname{erfc}\left(\frac{1}{\lambda}\right)\right] - \log\left[e^{\frac{\sqrt{2}x}{\lambda}}\operatorname{erfc}\left(\frac{x}{\sqrt{2}} + \frac{1}{\lambda}\right) + e^{-\frac{\sqrt{2}x}{\lambda}}\operatorname{erfc}\left(-\frac{x}{\sqrt{2}} + \frac{1}{\lambda}\right)\right]\right]\ ,$$

$$= \log\left[\frac{\sqrt{2}x}{\lambda}\right] + \log\left[-1 - \frac{\lambda}{\sqrt{2}x}\log\left[\frac{\operatorname{erfc}\left(\frac{x}{\sqrt{2}} + \frac{1}{\lambda}\right) + e^{-\frac{2\sqrt{2}x}{\lambda}}\operatorname{erfc}\left(-\frac{x}{\sqrt{2}} + \frac{1}{\lambda}\right)}{2\operatorname{erfc}\left(\frac{1}{\lambda}\right)}\right]\right]\ . \tag{59}$$

Next, using Lemma C.1 with $a = 1/\sqrt{2}$ and $b = 1/\lambda$, it follows that

$$\varphi^1_\lambda(x) = \log\left[\frac{\sqrt{2}x}{\lambda}\right] + \log\left[-\left(\frac{1}{\sqrt{2}\lambda} - \frac{e^{-\frac{1}{\lambda^2}}}{\sqrt{2}\operatorname{erfc}(\frac{1}{\lambda})\sqrt{\pi}}\right)x + o(x)\right]\ , \tag{60}$$

$$= \log\left[\frac{x^2}{\lambda}\right] + \log\left[\frac{e^{-\frac{1}{\lambda^2}}}{\operatorname{erfc}(\frac{1}{\lambda})\sqrt{\pi}} - \frac{1}{\lambda} + o(1)\right]\ , \tag{61}$$

$$= \log\left[\frac{x^2}{\lambda}\right] + \log\left[\frac{1}{\sqrt{\pi}}\frac{e^{-\frac{1}{\lambda^2}}}{\operatorname{erfc}(\frac{1}{\lambda})} - \frac{1}{\lambda}\right] + o(1)\ , \tag{62}$$

where the last equation follows from the first-order Taylor expansion of $\log(a + x)$. ■

Appendix D. Proof of Theorem 2.

Lemma D.1. *Let $a > 0$ and $b > 0$. For x in the vicinity of $+\infty$, we have*

$$\frac{1}{2abx}\log\left[\frac{\operatorname{erfc}(ax+b)+e^{-4abx}\operatorname{erfc}(-ax+b)}{2}\right] = -2+o(1)\ . \tag{63}$$

Proof. We have $\lim_{x\to+\infty}\operatorname{erfc}(x)=0$ and $\lim_{x\to+\infty}\operatorname{erfc}(-x)=2$, it follows that

$$\operatorname{erfc}(-ax+b) \underset{+\infty}{\sim} 2\ , \tag{64}$$

$$e^{-4abx}\operatorname{erfc}(-ax+b) \underset{+\infty}{\sim} 2e^{-4abx}\ , \tag{65}$$

$$\frac{\operatorname{erfc}(ax+b)+e^{-4abx}\operatorname{erfc}(-ax+b)}{2} \underset{+\infty}{\sim} e^{-4abx}\ , \tag{66}$$

$$\log\left[\frac{\operatorname{erfc}(ax+b)+e^{-4abx}\operatorname{erfc}(-ax+b)}{2}\right] \underset{+\infty}{\sim} -4abx\ , \tag{67}$$

$$\frac{1}{2abx}\log\left[\operatorname{erfc}(ax+b)+e^{-4abx}\operatorname{erfc}(-ax+b)\right] \underset{+\infty}{\sim} -2\ , \tag{68}$$

where we have used the knowledge that $f\sim g$ implies that $\log f\sim\log g$. ∎

Proof of Theorem 2. By writing φ^1_λ as in eq. (59) and using Lemma D.1 with $a=1/\sqrt{2}$ and $b=1/\lambda$, it follows that

$$\varphi^1_\lambda(x) = \log\left[\frac{\sqrt{2}x}{\lambda}\right]+\log\left[1+\frac{\lambda}{\sqrt{2}x}\log\operatorname{erfc}\left(\frac{1}{\lambda}\right)+o(1)\right] = \log\left[\frac{\sqrt{2}x}{\lambda}\right]+o(1)\ .$$ ∎

Appendix E. Proof of Theorem 3.

Proof. We first decompose the term $\exp\left(-\frac{(x-t)^2}{2}\right)$ involved in the definition of $f^\nu_{1,\lambda}$ and rewrite e^{xt} using its power series

$$f^\nu_{1,\lambda}(x) = -\log\frac{1}{\sqrt{2\pi}}\frac{\nu}{2\lambda_\nu\Gamma(1/\nu)}-\log\int_{-\infty}^{\infty}\exp\left(-\frac{(x-t)^2}{2}\right)\exp\left[-\left(\frac{|t|}{\lambda_\nu}\right)^\nu\right]\mathrm{d}t\ , \tag{69}$$

$$= -\log\frac{1}{\sqrt{2\pi}}\frac{\nu}{2\lambda_\nu\Gamma(1/\nu)}-\log\int_{-\infty}^{\infty}e^{-\frac{x^2}{2}}e^{xt}e^{-\frac{t^2}{2}}\exp\left[-\left(\frac{|t|}{\lambda_\nu}\right)^\nu\right]\mathrm{d}t\ , \tag{70}$$

$$= -\log\frac{1}{\sqrt{2\pi}}\frac{\nu}{2\lambda_\nu\Gamma(1/\nu)}+\frac{x^2}{2}-\log\int_{-\infty}^{\infty}\sum_{k=0}^{\infty}\frac{(xt)^k}{k!}e^{-\frac{t^2}{2}}\exp\left[-\left(\frac{|t|}{\lambda_\nu}\right)^\nu\right]\mathrm{d}t\ . \tag{71}$$

For x in the vicinity of 0, we can consider $|x|\leqslant 1$, and then

$$\int_{-\infty}^{\infty}\sum_{k=0}^{\infty}\left|\frac{(xt)^k}{k!}e^{-\frac{t^2}{2}}\exp\left[-\left(\frac{|t|}{\lambda_\nu}\right)^\nu\right]\right|\mathrm{d}t \leqslant \int_{-\infty}^{\infty}\sum_{k=0}^{\infty}\frac{|t|^k}{k!}e^{-\frac{t^2}{2}}\exp\left[-\left(\frac{|t|}{\lambda_\nu}\right)^\nu\right]\mathrm{d}t\ , \tag{72}$$

$$\leqslant \int_{-\infty}^{\infty}e^{-\frac{t^2}{2}+|t|-\left(\frac{|t|}{\lambda_\nu}\right)^\nu}\mathrm{d}t\ <\infty\ . \tag{73}$$

Then, Fubini's theorem applies and we get

$$f^{\nu}_{1,\lambda}(x) = -\log \frac{1}{\sqrt{2\pi}} \frac{\nu}{2\lambda_\nu \Gamma(1/\nu)} + \frac{x^2}{2} - \log \sum_{k=0}^{\infty} \int_{-\infty}^{\infty} \frac{(xt)^k}{k!} e^{-\frac{t^2}{2}} \exp\left[-\left(\frac{|t|}{\lambda_\nu}\right)^{\nu}\right] \mathrm{d}t \;, \tag{74}$$

$$= -\log \frac{1}{\sqrt{2\pi}} \frac{\nu}{2\lambda_\nu \Gamma(1/\nu)} + \frac{x^2}{2} - \log \sum_{k=0}^{\infty} \frac{x^k}{k!} \int_{-\infty}^{\infty} t^k e^{-\frac{t^2}{2}} \exp\left[-\left(\frac{|t|}{\lambda_\nu}\right)^{\nu}\right] \mathrm{d}t \;. \tag{75}$$

By definition, we have

$$\gamma^{\nu}_{\lambda} \triangleq f^{\nu}_{1,\lambda}(0) = -\log \frac{1}{\sqrt{2\pi}} \frac{\nu}{2\lambda_\nu \Gamma(1/\nu)} - \log \int_{-\infty}^{\infty} \exp\left(-\frac{t^2}{2}\right) \exp\left[-\left(\frac{|t|}{\lambda_\nu}\right)^{\nu}\right] \mathrm{d}t \;. \tag{76}$$

Moreover, when k is odd, we have

$$\int_{-\infty}^{\infty} t^k e^{-\frac{t^2}{2}} \exp\left[-\left(\frac{|t|}{\lambda_\nu}\right)^{\nu}\right] \mathrm{d}t = 0 \;. \tag{77}$$

Using third-order Taylor expansion of $\log(1+x)$ for x in the vicinity of 0, it follows that

$$f^{\nu}_{1,\lambda}(x) = \gamma^{\nu}_{\lambda} + \frac{x^2}{2} - \log\left(1 + \frac{x^2}{2} \frac{\int_{-\infty}^{\infty} t^2 e^{-\frac{t^2}{2}} \exp\left[-\left(\frac{|t|}{\lambda_\nu}\right)^{\nu}\right] \mathrm{d}t}{\int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} \exp\left[-\left(\frac{|t|}{\lambda_\nu}\right)^{\nu}\right] \mathrm{d}t} + o(x^3)\right) \;, \tag{78}$$

$$= \gamma^{\nu}_{\lambda} + \frac{x^2}{2}\left(1 - \frac{\int_{-\infty}^{\infty} t^2 e^{-\frac{t^2}{2}} \exp\left[-\left(\frac{|t|}{\lambda_\nu}\right)^{\nu}\right] \mathrm{d}t}{\int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} \exp\left[-\left(\frac{|t|}{\lambda_\nu}\right)^{\nu}\right] \mathrm{d}t}\right) + o(x^3) \;. \tag{79}$$

Finally, using first-order Taylor's expansion for $\log(1+x)$, we conclude the proof as

$$\varphi^{\nu}_{\lambda}(x) = \log\left[\frac{x^2}{2}\left(1 - \frac{\int_{-\infty}^{\infty} t^2 e^{-\frac{t^2}{2}} \exp\left[-\left(\frac{|t|}{\lambda_\nu}\right)^{\nu}\right] \mathrm{d}t}{\int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} \exp\left[-\left(\frac{|t|}{\lambda_\nu}\right)^{\nu}\right] \mathrm{d}t} + o(x)\right)\right] \;, \tag{80}$$

$$= 2\log x - \log 2 + \log\left(1 - \frac{\int_{-\infty}^{\infty} t^2 e^{-\frac{t^2}{2}} \exp\left[-\left(\frac{|t|}{\lambda_\nu}\right)^{\nu}\right] \mathrm{d}t}{\int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} \exp\left[-\left(\frac{|t|}{\lambda_\nu}\right)^{\nu}\right] \mathrm{d}t}\right) + o(x) \;. \tag{81}$$

■

Appendix F. Proof of Theorem 4.

We first recall in a Lemma, a result extracted from Corollary 3.3 in [2].

Lemma F.1 (Berman). *Let p and q be differentiable real probability density functions. Define for x large enough $u(x) = p^{-1}(q(x))$ and define v and w as*

$$v(x) = -\frac{\partial}{\partial x} \log p(x) \quad \textit{and} \quad w(x) = -\frac{\partial}{\partial x} \log q(x) \;. \tag{82}$$

Assume v and w are positive continuous function and regularly oscillating, i.e.:

$$\lim_{\substack{x,x'\to\infty \\ x/x'\to 1}} \frac{v(x)}{v(x')} = 1 \quad \textit{and} \quad \lim_{\substack{x,x'\to\infty \\ x/x'\to 1}} \frac{w(x)}{w(x')} = 1 \ . \tag{83}$$

Suppose that we have

$$\lim_{x\to\infty} \frac{w(x)}{v(x)} = 0 \quad \textit{and} \quad \lim_{x\to\infty} u(x)w(x) = +\infty \ , \tag{84}$$

then, for $x \to \infty$, we have

$$\log \int_{-\infty}^{+\infty} p(x-t)q(t)\ \mathrm{d}t \sim \log q(x) \ . \tag{85}$$

Proof of Theorem 4. Using the definition of the discrepancy function,

$$f^{\nu}_{1,\lambda}(x) = -\log \int_{\mathbb{R}} \mathcal{G}(t;0,\lambda,\nu)\cdot \mathcal{N}(x-t;0,1)\ \mathrm{d}t = -\log \int_{-\infty}^{+\infty} p(x-t)q(t)\ \mathrm{d}t \ . \tag{86}$$

with $q(x) = \mathcal{G}(x;0,\lambda,\nu)$ and $p(x) = \mathcal{N}(x;0,1)$. We have

$$v(x) = -\frac{\partial}{\partial x}\log p(x) = 2x \quad \text{and} \quad w(x) = -\frac{\partial}{\partial x}\log q(x) = \frac{\nu x^{\nu-1}}{\lambda_\nu^\nu} \ . \tag{87}$$

Remark that v and w are positive continuous, and

$$\lim_{\substack{x,x'\to\infty \\ x/x'\to 1}} \frac{v(x)}{v(x')} = \lim_{\substack{x,x'\to\infty \\ x/x'\to 1}} \frac{x}{x'} = 1 \quad \text{and} \quad \lim_{\substack{x,x'\to\infty \\ x/x'\to 1}} \frac{w(x)}{w(x')} = \lim_{\substack{x,x'\to\infty \\ x/x'\to 1}} \left(\frac{x}{x'}\right)^{\nu-1} = 1 \ . \tag{88}$$

For $x > 0$ large enough and $y > 0$ small enough

$$y = p(x) \Leftrightarrow y = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right) \Leftrightarrow x = \sqrt{-\log(2\pi) - 2\log y} \ . \tag{89}$$

Then, from Proposition 1, we have

$$u(x)w(x) = \frac{\nu x^{\nu-1}}{\lambda_\nu^\nu}\sqrt{-\log(2\pi) - 2\log q(x)} \tag{90}$$

$$= \frac{\nu x^{\nu-1}}{\lambda_\nu^\nu}\sqrt{-\log(2\pi) - 2\log\left[\frac{\kappa}{2\lambda_\nu}\right] + 2\left(\frac{x}{\lambda_\nu}\right)^{\nu}} \sim \frac{\nu\sqrt{2}x^{\frac{3}{2}\nu-1}}{\lambda_\nu^{\frac{3}{2}\nu}} \ , \tag{91}$$

and thus, as $\nu > \frac{2}{3}$, $\lim_{x\to\infty} u(x)w(x) = \infty$. Moreover, for $\nu < 2$, we have

$$\lim_{x\to\infty} \frac{w(x)}{v(x)} = \lim_{x\to\infty} \frac{\nu}{2\lambda_\nu} x^{\nu-2} = 0 \ . \tag{92}$$

It follows that Lemma F.1 applies, and then for large x

$$f^{\nu}_{1,\lambda}(x) \sim -\log q(x) \sim \left(\frac{x}{\lambda_\nu}\right)^{\nu} . \tag{93}$$

Using that $\lambda_\nu = \lambda\sqrt{\frac{\Gamma(1/\nu)}{\Gamma(3/\nu)}}$, we conclude the proof since

$$\varphi^{\nu}_{\lambda}(x) = \log\left[f^{\nu}_{1,\lambda}(x) - \gamma^{\nu}_{\lambda}\right] \sim \nu\log x - \nu\log\lambda_\nu . \tag{94}$$

■

Appendix G. Proof of Proposition 3.

Proof. Starting from the definition of $s^{\nu}_{\sigma,\lambda}$ and using the change of variable $t \to \sigma t$, eq. (reduction) follows as

$$s^{\nu}_{\sigma,\lambda}(x) = \operatorname*{argmin}_{t\in\mathbb{R}} \frac{(x-t)^2}{2\sigma^2} + \lambda_\nu^{-\nu}|t|^\nu = \sigma\operatorname*{argmin}_{t\in\mathbb{R}} \frac{(x-\sigma t)^2}{2\sigma^2} + \lambda_\nu^{-\nu}|\sigma t|^\nu , \tag{95}$$

$$= \sigma\operatorname*{argmin}_{t\in\mathbb{R}} \frac{(x/\sigma-t)^2}{2} + (\lambda_\nu/\sigma)^{-\nu}|t|^\nu = \sigma s^{\nu}_{1,\lambda/\sigma}(x/\sigma) . \tag{96}$$

For eq. (odd), we use the change of variable $t \to -t$

$$s^{\nu}_{\sigma,\lambda}(-x) = \operatorname*{argmin}_{t\in\mathbb{R}} \frac{(-x-t)^2}{2\sigma^2} + \lambda_\nu^{-\nu}|t|^\nu = -\operatorname*{argmin}_{t\in\mathbb{R}} \frac{(-x+t)^2}{2\sigma^2} + \lambda_\nu^{-\nu}|t|^\nu , \tag{97}$$

$$= -\operatorname*{argmin}_{t\in\mathbb{R}} \frac{(x-t)^2}{2\sigma^2} + \lambda_\nu^{-\nu}|t|^\nu = -s^{\nu}_{\sigma,\lambda}(x) . \tag{98}$$

We now prove eq. (shrinkage). Let $t = s^{\nu}_{\sigma,\lambda}(x)$, and since t minimizes the objective, then

$$\lambda_\nu^{-\nu}|t|^\nu \leqslant \frac{(x-t)^2}{2\sigma^2} + \lambda_\nu^{-\nu}|t|^\nu \leqslant \frac{(x-x)^2}{2\sigma^2} + \lambda_\nu^{-\nu}|x|^\nu = \lambda_\nu^{-\nu}|x|^\nu , \tag{99}$$

which implies that $|t| \leqslant |x|$. Let $x > 0$ and assume $t = s^{\nu}_{\sigma,\lambda}(x) < 0$. Since t minimizes the objective, then

$$\frac{(x-t)^2}{2\sigma^2} + \lambda_\nu^{-\nu}|t|^\nu \leqslant \frac{(x+t)^2}{2\sigma^2} + \lambda_\nu^{-\nu}|t|^\nu \tag{100}$$

which implies that $-xt \leqslant xt$ and leads to a contradiction. Then for $x > 0$, $s^{\nu}_{\sigma,\lambda}(x) \in [0, x]$, which concludes the proof since $s^{\nu}_{\sigma,\lambda}$ is odd.

We now prove (increasing with x). Let $x_1 > x_2$ and define $t_1 = s^{\nu}_{\sigma,\lambda}(x_1)$ and $t_2 = s^{\nu}_{\sigma,\lambda}(x_2)$. Since t_1 and t_2 minimize their respective objectives, the following two statements hold

$$\frac{(x_1-t_1)^2}{2\sigma^2} + \lambda_\nu^{-\nu}|t_1|^\nu \leqslant \frac{(x_1-t_2)^2}{2\sigma^2} + \lambda_\nu^{-\nu}|t_2|^\nu , \tag{101}$$

$$\text{and} \quad \frac{(x_2-t_2)^2}{2\sigma^2} + \lambda_\nu^{-\nu}|t_2|^\nu \leqslant \frac{(x_2-t_1)^2}{2\sigma^2} + \lambda_\nu^{-\nu}|t_1|^\nu . \tag{102}$$

Summing both inequalities lead to

$$(x_1 - t_1)^2 + (x_2 - t_2)^2 \leqslant (x_1 - t_2)^2 + (x_2 - t_1)^2 \ , \tag{103}$$

$$\Rightarrow \quad -2x_1t_1 - 2x_2t_2 \leqslant -2x_1t_2 - 2x_2t_1 \ , \tag{104}$$

$$\Rightarrow \quad t_1(x_1 - x_2) \geqslant t_2(x_1 - x_2) \quad \Rightarrow \quad t_1 \geqslant t_2 \quad (\text{since } x_1 > x_2) \ . \tag{105}$$

We now prove (increasing with λ). Let $\lambda_1 > \lambda_2$ and define $t_1 = s^{\nu}_{\sigma,\lambda_1}(x)$ and $t_2 = s^{\nu}_{\sigma,\lambda_2}(x)$. Since t_1 and t_2 minimize their respective objectives, the following expressions hold

$$\frac{(x - t_1)^2}{2\sigma^2} + \lambda^{-\nu}_{\nu,1}|t_1|^{\nu} \leqslant \frac{(x - t_2)^2}{2\sigma^2} + \lambda^{-\nu}_{\nu,1}|t_2|^{\nu} \ , \tag{106}$$

$$\text{and} \quad \frac{(x - t_2)^2}{2\sigma^2} + \lambda^{-\nu}_{\nu,2}|t_2|^{\nu} \leqslant \frac{(x - t_1)^2}{2\sigma^2} + \lambda^{-\nu}_{\nu,2}|t_1|^{\nu} \ . \tag{107}$$

Again, summing both inequalities lead to

$$\lambda^{-\nu}_{\nu,1}|t_1|^{\nu} + \lambda^{-\nu}_{\nu,2}|t_2|^{\nu} \leqslant \lambda^{-\nu}_{\nu,1}|t_2|^{\nu} + \lambda^{-\nu}_{\nu,2}|t_1|^{\nu} \ , \tag{108}$$

$$\Rightarrow \quad (\lambda^{-\nu}_{\nu,1} - \lambda^{-\nu}_{\nu,2})|t_1|^{\nu} \leqslant (\lambda^{-\nu}_{\nu,1} - \lambda^{-\nu}_{\nu,2})|t_2|^{\nu} \ , \tag{109}$$

$$\Rightarrow \quad |t_1|^{\nu} \geqslant |t_2|^{\nu} \quad (\text{since } \lambda_1 > \lambda_2 \text{ and } \nu > 0) \ . \tag{110}$$

We now prove (keep high SNR). Consider $x > 0$. Since $\lambda \mapsto s^{\nu}_{\sigma,\lambda}(x)$ is a monotonic function and $s^{\nu}_{\sigma,\lambda}(x) \in [0, x]$ for all λ, it converges for $\lambda \to \infty$ to a value $\omega \in [0, x]$. Assume $0 < \omega < x$ and let $0 < \varepsilon < \max(\omega, x - \omega)$. By definition of the limit, for λ big enough

$$0 < \omega - \varepsilon < t \triangleq s^{\nu}_{\sigma,\lambda}(x) < \omega + \varepsilon \ . \tag{111}$$

It follows that $x - t > x - (w + \varepsilon) > 0$, and then

$$\frac{(x - (\omega + \varepsilon))^2}{2\sigma^2} + \lambda^{-\nu}_{\nu}|\omega - \varepsilon|^{\nu} < \frac{(x - t)^2}{2\sigma^2} + \lambda^{-\nu}_{\nu}|t|^{\nu} \ . \tag{112}$$

Moreover, since $\omega + \varepsilon \neq x$, we have for λ big enough

$$\lambda^{-\nu}_{\nu}|x|^{\nu} < \frac{(x - (\omega + \varepsilon))^2}{2\sigma^2} + \lambda^{-\nu}_{\nu}|\omega - \varepsilon|^{\nu} \ . \tag{113}$$

Combining the two last inequalities shows that

$$\frac{(x - x)^2}{2\sigma^2} + \lambda^{-\nu}_{\nu}|x|^{\nu} < \frac{(x - t)^2}{2\sigma^2} + \lambda^{-\nu}_{\nu}|t|^{\nu} \ , \tag{114}$$

which is in contradiction with the fact that t minimizes the objective. As a consequence, $\omega = x$, which concludes the proof since $s^{\nu}_{\sigma,\lambda}(x)$ is odd and satisfies (reduction).

We now prove (kill low SNR). Consider $x > 0$. Since $\lambda \mapsto s^{\nu}_{\sigma,1/\lambda}(x)$ is a monotonic function and $s^{\nu}_{\sigma,\lambda}(x) \in [0, x]$ for all λ, it converges for $\lambda \to 0^+$ to a value $\omega \in [0, x]$. Assume $0 < \omega < x$ and let $0 < \varepsilon < \max(\omega, x - \omega)$. Again, we have for λ small enough

$$\frac{(x - (\omega + \varepsilon))^2}{2\sigma^2} + \lambda^{-\nu}_{\nu}|\omega - \varepsilon|^{\nu} < \frac{(x - t)^2}{2\sigma^2} + \lambda^{-\nu}_{\nu}|t|^{\nu} \ . \tag{115}$$

Moreover, since $\omega \neq \varepsilon$, we have for λ small enough

$$\frac{x^2}{2\sigma^2} < \frac{(x-(\omega+\varepsilon))^2}{2\sigma^2} + \lambda_\nu^{-\nu}|\omega-\varepsilon|^\nu \ . \tag{116}$$

Combining the two last inequalities shows that

$$\frac{(x-0)^2}{2\sigma^2} + \lambda_\nu^{-\nu}|0|^\nu < \frac{(x-t)^2}{2\sigma^2} + \lambda_\nu^{-\nu}|t|^\nu \ , \tag{117}$$

which is in contradiction with the fact that t minimizes the objective. As a consequence, $\omega = 0$, which concludes the proof since $s^\nu_{\sigma,\lambda}(x)$ is odd and satisfies (reduction). ∎

REFERENCES

[1] M. Aharon, M. Elad, and A. Bruckstein, *k-SVD: An algorithm for designing overcomplete dictionaries for sparse representation*, IEEE Transactions on signal processing, 54 (2006), pp. 4311–4322.

[2] S. M. Berman, *The tail of the convolution of densities and its application to a model of HIV-latency time*, The Annals of Applied Probability, 2 (1992), pp. 481–502.

[3] L. Boubchir and J. Fadili, *Multivariate statistical modeling of images with the curvelet transform*, in ISSPA, 2005, pp. 747–750.

[4] K. Bredies, K. Kunisch, and T. Pock, *Total generalized variation*, SIAM Journal on Imaging Sciences, 3 (2010), pp. 492–526.

[5] A. Buades, B. Coll, and J.-M. Morel, *A non-local algorithm for image denoising*, in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 2, IEEE, 2005, pp. 60–65.

[6] S. G. Chang, B. Yu, and M. Vetterli, *Adaptive wavelet thresholding for image denoising and compression*, IEEE transactions on image processing, 9 (2000), pp. 1532–1546.

[7] C. Chaux, P. L. Combettes, J.-C. Pesquet, and V. R. Wajs, *A variational formulation for frame-based inverse problems*, Inverse Problems, 23 (2007), p. 1495.

[8] R. R. Coifman and D. L. Donoho, *Translation-invariant de-noising*, Wavelets and statistics, 103 (1995), pp. 125–150.

[9] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, *Image denoising by sparse 3-d transform-domain collaborative filtering*, IEEE Transactions on image processing, 16 (2007), pp. 2080–2095.

[10] G. Davis, S. Mallat, and Z. Zhang, *Adaptive time-frequency decompositions with matching pursuit*, Optical Engineering, 33 (1994).

[11] C.-A. Deledalle, J. Salmon, A. S. Dalalyan, et al., *Image denoising with patch based pca: local versus global.*, in BMVC, vol. 81, 2011, pp. 425–455.

[12] C.-A. Deledalle, S. Vaiter, J. Fadili, and G. Peyré, *Stein unbiased gradient estimator of the risk (sugar) for multiple parameter selection*, SIAM Journal on Imaging Sciences, 7 (2014), pp. 2448–2487.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the royal statistical society. Series B (methodological), (1977), pp. 1–38.

[14] M. N. Do and M. Vetterli, *Wavelet-based texture retrieval using generalized gaussian density and kullback-leibler distance*, IEEE transactions on image processing, 11 (2002), pp. 146–158.

[15] D. L. Donoho and I. M. Johnstone, *Adapting to unknown smoothness via wavelet shrinkage*, Journal of the american statistical association, 90 (1995), pp. 1200–1224.

[16] D. L. Donoho and J. M. Johnstone, *Ideal spatial adaptation by wavelet shrinkage*, biometrika, 81 (1994), pp. 425–455.

[17] R. J. Duffin and A. C. Schaeffer, *A class of nonharmonic fourier series*, Transactions of the American Mathematical Society, 72 (1952), pp. 341–366.

[18] C. Dugas, Y. Bengio, F. Bélisle, C. Nadeau, and R. Garcia, *Incorporating second-order functional knowledge for better option pricing*, in Advances in neural information processing systems, 2001, pp. 472–478.

[19] M. Elad and M. Aharon, *Image denoising via sparse and redundant representations over learned dictionaries*, IEEE Transactions on Image processing, 15 (2006), pp. 3736–3745.
[20] M. Elad, P. Milanfar, and R. Rubinstein, *Analysis versus synthesis in signal priors*, Inverse problems, 23 (2007), p. 947.
[21] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman, *Removing camera shake from a single photograph*, in ACM transactions on graphics (TOG), vol. 25, ACM, 2006, pp. 787–794.
[22] D. Geman and C. Yang, *Nonlinear image recovery with half-quadratic regularization*, IEEE Transactions on Image Processing, 4 (1995), pp. 932–946.
[23] S. Geman and D. Geman, *Stochastic relaxation, gibbs distributions, and the bayesian restoration of images*, IEEE Transactions on pattern analysis and machine intelligence, (1984), pp. 721–741.
[24] G. H. Golub, M. Heath, and G. Wahba, *Generalized cross-validation as a method for choosing a good ridge parameter*, Technometrics, 21 (1979), pp. 215–223.
[25] P. C. Hansen, *The L-curve and its use in the numerical treatment of inverse problems*, IMM, Department of Mathematical Modelling, Technical Universityof Denmark, 1999.
[26] A. Houdard, C. Bouveyron, and J. Delon, *High-Dimensional Mixture Models For Unsupervised Image Denoising (HDMI)*. Preprint hal-01544249, Aug. 2017, https://hal.archives-ouvertes.fr/hal-01544249.
[27] P. J. Huber, *Robust statistics*, in International Encyclopedia of Statistical Science, Springer, 2011, pp. 1248–1251.
[28] P. J. Huber et al., *Robust estimation of a location parameter*, The Annals of Mathematical Statistics, 35 (1964), pp. 73–101.
[29] H. Ji, S. Huang, Z. Shen, and Y. Xu, *Robust video restoration by joint sparse and low rank matrix approximation*, SIAM Journal on Imaging Sciences, 4 (2011), pp. 1122–1142.
[30] D. Krishnan and R. Fergus, *Fast image deconvolution using hyper-laplacian priors*, in Advances in Neural Information Processing Systems, 2009, pp. 1033–1041.
[31] R. Krupiński, *Approximated fast estimator for the shape parameter of generalized gaussian distribution for a small sample size*, Bulletin of the Polish Academy of Sciences Technical Sciences, 63 (2015), pp. 405–411.
[32] M. Lebrun, A. Buades, and J.-M. Morel, *A nonlocal bayesian image denoising algorithm*, SIAM Journal on Imaging Sciences, 6 (2013), pp. 1665–1688.
[33] S. Mallat, *A theory for multiresolution signal decomposition: the wavelet representation*, IEEE transactions on pattern analysis and machine intelligence, 11 (1989), pp. 674–693.
[34] S. Mallat, *A wavelet tour of signal processing: the sparse way*, Academic press, 2008.
[35] D. Martin, C. Fowlkes, D. Tal, and J. Malik, *A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics*, in Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on, vol. 2, IEEE, 2001, pp. 416–423.
[36] T. Michaeli and M. Irani, *Blind deblurring using internal patch recurrence*, in European Conference on Computer Vision, Springer, 2014, pp. 783–798.
[37] P. Mittal and K. Gupta, *An integral involving generalized function of two variables*, in Proceedings of the Indian academy of sciences-section A, vol. 75, Springer, 1972, pp. 117–123.
[38] O. M. M. Mohamed and M. Jaïdane-Saïdane, *Generalized gaussian mixture model*, in Signal Processing Conference, 2009 17th European, IEEE, 2009, pp. 2273–2277.
[39] P. Moulin and J. Liu, *Analysis of multiresolution image denoising schemes using generalized gaussian and complexity priors*, IEEE Transactions on Information Theory, 45 (1999), pp. 909–919.
[40] V. Nair and G. E. Hinton, *Rectified linear units improve restricted boltzmann machines*, in Proceedings of the 27th international conference on machine learning (ICML-10), 2010, pp. 807–814.
[41] M. Nardon and P. Pianca, *Simulation techniques for generalized gaussian densities*, Journal of Statistical Computation and Simulation, 79 (2009), pp. 1317–1329.
[42] M. Niknejad, J. M. Bioucas-Dias, and M. A. Figueiredo, *Class-specific image denoising using importance sampling*, in IEEE International Conference on Image Processing, IEEE, 2017.
[43] V. Papyan and M. Elad, *Multi-scale patch-based image restoration*, IEEE Transactions on image processing, 25 (2016), pp. 249–261.
[44] S. Parameswaran, C.-A. Deledalle, L. Denis, and T. Q. Nguyen, *Accelerating gmm-based patch*

priors for image restoration: Three ingredients for a 100× speed-up, arXiv preprint arXiv:1710.08124, (2017).

[45] F. Pascal, L. Bombrun, J.-Y. Tourneret, and Y. Berthoumieu, *Parameter estimation for multivariate generalized gaussian distributions*, IEEE Transactions on Signal Processing, 61 (2013), pp. 5960–5971.

[46] K. P. Peppas, *A new formula for the average bit error probability of dual-hop amplify-and-forward relaying systems over generalized shadowed fading channels*, IEEE Wireless Communications Letters, 1 (2012), pp. 85–88.

[47] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, *Image denoising using scale mixtures of gaussians in the wavelet domain*, IEEE Transactions on Image processing, 12 (2003), pp. 1338–1351.

[48] S. Purkayastha, *Simple proofs of two results on convolutions of unimodal distributions*, Statistics & probability letters, 39 (1998), pp. 97–100.

[49] S. Ramani, T. Blu, and M. Unser, *Monte-carlo sure: A black-box optimization of regularization parameters for general denoising algorithms*, IEEE Transactions on Image Processing, 17 (2008), pp. 1540–1554.

[50] A. A. Roenko, V. V. Lukin, I. Djurovic, and M. Simeunovic, *Estimation of parameters for generalized gaussian distribution*, in Communications, Control and Signal Processing (ISCCSP), 2014 6th International Symposium on, IEEE, 2014, pp. 376–379.

[51] S. Roth and M. J. Black, *Fields of experts: A framework for learning image priors*, in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 2, IEEE, 2005, pp. 860–867.

[52] R. Rubinstein, T. Peleg, and M. Elad, *Analysis K-SVD: A dictionary-learning algorithm for the analysis sparse model*, IEEE Transactions on Signal Processing, 61 (2013), pp. 661–677.

[53] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Physica D: Nonlinear Phenomena, 60 (1992), pp. 259–268.

[54] M. J. Shensa, *The discrete wavelet transform: wedding the a trous and mallat algorithms*, IEEE Transactions on signal processing, 40 (1992), pp. 2464–2482.

[55] G. Singh and J. Singh, *Adapted non-gaussian mixture model for image denoising*, Journal of Basic and Applied Engineering Research, 4 (2017).

[56] H. Soury and M.-S. Alouini, *New results on the sum of two generalized gaussian random variables*, in Signal and Information Processing (GlobalSIP), 2015 IEEE Global Conference on, IEEE, 2015, pp. 1017–1021.

[57] J. Sulam and M. Elad, *Expected patch log likelihood with a sparse prior*, Energy Minimization Methods in Computer Vision and Pattern Recognition, Lecture Notes in Computer Science, (2015), pp. 99–111.

[58] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, *Image quality assessment: from error visibility to structural similarity*, IEEE transactions on image processing, 13 (2004), pp. 600–612.

[59] A. Wintner, *Asymptotic distributions and infinite convolutions. edwards brothers*, Ann Arbor, MI. In Bickel, PJ, (1938).

[60] J. Yang, Y. Zhang, and W. Yin, *An efficient tvl1 algorithm for deblurring multichannel images corrupted by impulsive noise*, SIAM Journal on Scientific Computing, 31 (2009), pp. 2842–2865.

[61] G. Yu, G. Sapiro, and S. Mallat, *Solving inverse problems with piecewise linear estimators: From gaussian mixture models to structured sparsity*, IEEE Transactions on Image Processing, 21 (2012), pp. 2481–2499.

[62] D. Zoran and Y. Weiss, *From learning models of natural image patches to whole image restoration*, in International Conference on Computer Vision, IEEE, November 2011, pp. 479–486, https://doi.org/10.1109/ICCV.2011.6126278.