How to compare noisy patches? Patch similarity beyond Gaussian noise

Charles-Alban Deledalle · Loïc Denis · Florence Tupin

February 17, 2012

Abstract Many tasks in computer vision require to match image parts. While higher-level methods consider image features such as edges or robust descriptors, low-level approaches (so-called image-based) compare groups of pixels (patches) and provide dense matching. Patch similarity is a key ingredient to many techniques for image registration, stereo-vision, change detection or denoising. Recent progress in natural image modeling also makes intensive use of patch comparison.

A fundamental difficulty when comparing two patches from "real" data is to decide whether the differences should be ascribed to noise or intrinsic dissimilarity. Gaussian noise assumption leads to the classical definition of patch similarity based on the squared differences of intensities. For the case where noise departs from the Gaussian distribution, several similarity criteria have been proposed in the literature of image processing, detection theory and machine learning.

By expressing patch (dis)similarity as a detection test under a given noise model, we introduce these criteria with a new one and discuss their properties. We then assess their performance for different tasks: patch discrimination, image denoising, stereo-matching and motion-tracking under gamma and Poisson noises. The proposed criterion based on the generalized likelihood ratio is shown to be both easy to derive and powerful in these diverse applications.

Keywords Patch similarity, Likelihood ratio, Detection, Matching

1 Introduction

Patches are small image parts that capture both texture and local structure information. Though being crude low-level features compared to higher level descriptors, they have led to very powerful approaches in a wide range of computer vision tasks and image processing models.

Charles-Alban Deledalle
Institut Telecom ; Telecom ParisTech ; CNRS LTCI
46, rue Barrault 75634 Paris cedex 13, France
E-mail: charles-alban.deledalle@telecom-paristech.fr

Loïc Denis
Université de Lyon, F-42023, Saint-Etienne, France,
CNRS, UMR5516, Laboratoire Hubert Curien, F-42000, Saint-Etienne, France,
Université de Saint-Etienne, Jean Monnet, F-42000, Saint-Etienne, France.
E-mail: loic.denis@univ-st-etienne.fr

Florence Tupin
Institut Telecom ; Telecom ParisTech ; CNRS LTCI
46, rue Barrault 75634 Paris cedex 13, France
E-mail: florence.tupin@telecom-paristech.fr

To classify textures, Varma and Zisserman (2003) have shown that patch-based classifiers lead to better performance than higher-level features computed using filter banks. State-of-the-art methods for texture synthesis (i.e., generation of a larger image from a given texture image) or inpainting (i.e., filling missing information in images) heavily rely on the concept of patch (e.g. Efros and Freeman, 2001; Liang et al., 2001; Kwatra et al., 2003; Criminisi et al., 2004). Image editing based on user-defined constraints is also performed through a decomposition into image patches (Cho et al., 2009).

The notion of patches is central to statistical models of natural images in early computational vision (Hyvärinen et al., 2009). The non-local means (NL-Means) (Buades et al., 2005b) introduced the concept of non-locality central to most of recent denoising techniques (Katkovnik et al., 2010). State-of-the art denoising techniques all rely on patches, either for dictionary learning (Elad and Aharon, 2006), for collaborative denoising of blocks of similar patches (Dabov et al., 2007) or for non-local sparse models (Mairal et al., 2009). Regularization with non-local patch-based weights have shown to improve on classical regularization involving only local neighborhoods (Gilboa and Osher, 2008; Peyré et al., 2008; Zhang et al., 2010). Pairs of low-resolution and high-resolution patches can be combined to design a super-resolution method (Freeman et al., 2002).

Estimated patch-similarity (or patch-dissimilarity) is at the heart of numerous image processing methods, e.g., region-based methods for image registration (Zitova and Flusser, 2003), matching in stereo-vision (Scharstein and Szeliski, 2002) or block selection for denoising (Buades et al., 2005b). Similarity between pixel values has been defined in many different ways in the literature, depending on the vision problem at hand, the noise model and the prior knowledge. While the shape and size of patches should adapt to the multi-scale and anisotropic behaviour of natural images (Dabov et al., 2008; Deledalle et al., 2011b), the choice of the similarity criterion is rather a problem related to the nature of noise. When comparing noisy patches, adaptation to noise distribution is essential for robust similarity evaluation.

We focus in the following on how to compare noisy values, and how similarity criteria can be derived from a given noise distribution. The comparison of noise-free patches (design of a suitable metric in noise-free patches space) and the similarity between a noise-free and a noisy version of a patch (template matching) are out of the scope of this paper.

There have been few attempts to define a methodology for the derivation of patch-similarity criteria adapted to given noise distributions. In the context of image block matching, Alter et al. (2006) were among the first to address this problem. They have shown that their criterion, based on maximum likelihood estimation, improves over the classical Euclidean-distance. This criterion has later been refined by Matsushita and Lin (2007) to avoid the maximum likelihood estimation step and to better take into account the shape of the likelihood distributions. This corresponds also to the approach considered in our previous work on patch-based denoising with non-Gaussian noise, for multiplicative noise (Deledalle et al., 2009b), impulsive noise (Deledalle et al., 2009a) and multi-dimensional complex data with circular complex Gaussian likelihood (Deledalle et al., 2011a).

Matsushita's approach has, however, several limitations: the criterion is hard to derive in closed-form, it requires defining a *prior* model and its performance depends heavily on the choice of the representation domain of the observations. The latter limitation has recently been pointed out by Teuber and Lang (Teuber and Lang, 2011) who showed that the criterion we proposed for multiplicative noise in (Deledalle et al., 2009b) leads to different expressions whether it is derived for squared data or log-transformed data. Depending on the transformation choice, such criteria can lead to the following paradox: two different values can be more similar than two identical values. It appears that this result has been known since 1995 in the community of pattern recognition and information theory. Indeed, Matsushita's criterion can be traced back to the *stochastic equivalence predicate* introduced by Yianilos (1995) on metric learning where the above paradox is referred to as *the self-recognition paradox*.

At the end of the 90s, Minka (2000) exhibited an equivalence between *canonical distance measures*, developed in (Baxter, 1995; Baxter and Bartlett, 1998), and the work of Yianilos, thanks to a Bayesian formulation based on *prior* distributions. He referred to his criterion as the evidence ratio and linked it to mutual information (Minka, 1998). Concurrently, in the context of machine learning, Seeger (2002) introduced the mutual information kernel as an inner product in a high dimensional space. As he stated himself, his kernel can be also interpreted as a Bayesian extension of Yianilos' criterion. Compared to (Yianilos, 1995; Alter et al., 2006; Matsushita and Lin, 2007), their methodology provides criteria with unchanged expression whatever the representation of the observations, and, as we show in section 3, Seeger's criterion solves *the self-recognition paradox*. A common limitation to all these approaches is the introduction of a *prior* model on the distribution of the underlying noise-free values.

Recently, we have introduced another criterion (Deledalle et al., 2011c) used in the case of Poisson noise in (Deledalle et al., 2010) which can be viewed as a combination or unification of (Minka, 2000; Seeger, 2002; Alter et al., 2006). Independently, Chen et al. (2011) proposed a similar definition for complex Wishart distributions. This methodology is *prior*-less, independent of the representation of the observations and solves *the self-recognition paradox* under reasonable assumptions. In this paper, we show that it corresponds to the generalized likelihood ratio (GLR) test.

Main contributions We address the problem of defining patch similarity under non-Gaussian noise.

We first propose to express formally patch dissimilarity as a statistical test. In the light of this test, we describe several similarity criteria proposed in the literature and discuss their theoretical grounding. The definition of patch dissimilarity as a statistical test provides a new point of view on these criteria driven by many years of research on detection theory.

We consider the properties that a satisfying similarity criterion should fulfill and discuss which properties each criterion fulfills. This provides arguments in favour of well-behaved criteria.

We then turn to a task-based evaluation of the criteria. We compare the ability of each criterion to discriminate patches from a dictionary learnt on a natural image. The performance of each criterion is assessed for non-local denoising under Poisson and gamma noises. We illustrate the use of non-quadratic matching costs in stereo matching when the stereo pair is corrupted by non-Gaussian noise. In a motion-tracking problem for glacier monitoring, we show the superiority of a similarity criterion designed for the multiplicative speckle noise that occurs in synthetic aperture radar (SAR) images.

We advocate that the proposed formulation based on GLR offers a flexible yet powerful way to generalize patch similarity to non-Gaussian noises. Beyond dissimilarity detection, task-specific weighting of the similarity criterion is required to reach optimal performance. For low to moderate noise levels, quadratic difference computed on stabilized-variance data proves preferable to unweighted use of other criteria.

Outline Section 2 proposes a definition of patch (dis)similarity and describes several criteria. Some desirable properties of similarity criteria necessary for comparing patch similarities are then discussed in Sect. 3. Task-based evaluation of the criteria is performed in Sect. 4. We discuss the importance of adapting patch similarity to noise models in Sect. 5 and draw some conclusions from our comparisons of similarity criteria.

2 Patch similarity criteria

In this section, we propose to express the similarity between noisy patches based on the detection of dissimilarity. Noisy patches are modeled in a probabilistic way in order to take into account the noise statistics. The notations are given as well as the fundamental concepts of detection theory. Seven criteria, extracted from the fields of image processing, detection theory and machine learning, are studied. Their concepts, origins and motivations are given. Their theoretical performance and limitations to solve our detection problem are then discussed.

By $\boldsymbol{x}$ we denote a patch, i.e., a collection of N observations (pixel values). At each pixel, the observation may be $D-$dimensional (e.g., $D = 1$ for gray-level images, $D = 3$ for RGB color images), so that $\boldsymbol{x}$ is a $N \times D$ vector obtained by stacking the observations of each pixel of the patch. We do not specify here a shape for the patch but consider that the values in vector $\boldsymbol{x}$ are ordered so that when two patches $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are compared, values with identical index are in (spatial) correspondence.

We assume that noise can be modeled by a given distribution so that a noisy patch $\boldsymbol{x}$ is a realization of an $N \times D$-dimensional random variable $\boldsymbol{X}$ characterized by the probability density function (pdf) $p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta})$ (written $p(\boldsymbol{x}|\boldsymbol{\theta})$ in the following, for the sake of notational simplicity). The vector of parameters $\boldsymbol{\theta}$ of that pdf is referred to in the following as the noise-free patch[1].

For example, a patch $\boldsymbol{x}$ damaged by additive white Gaussian noise with standard deviation σ can be modeled by:

$$\boldsymbol{x} = \boldsymbol{\theta} + \sigma \boldsymbol{n} \tag{1}$$

[1] the vector of parameters $\boldsymbol{\theta}$ may have a different number of dimensions than noisy patches $\boldsymbol{x}$

where $\boldsymbol{\theta}$ is the noise-free patch and $\boldsymbol{n}$ is the realization of a zero-mean normalized Gaussian random vector with independent elements. It is straightforward to see that $\boldsymbol{X}|\boldsymbol{\theta}$ follows a Gaussian distribution with mean $\boldsymbol{\theta}$ and standard deviation σ. While such decompositions exist also for some other distributions (e.g., gamma distribution involves a multiplicative decomposition), there is not necessarily a decomposition of $\boldsymbol{x}$ in terms of $\boldsymbol{\theta}$ and an independent noise component (this is for example the case with Poisson noise). In general, when noise departs from additive Gaussian noise, the link between $\boldsymbol{X}$ and $\boldsymbol{\theta}$ is then described by its likelihood function $p(\boldsymbol{x}|\boldsymbol{\theta})$.

Detecting dissimilarity: a pair of (noisy) patches $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ is considered similar (i.e., in-match) when $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are realizations of independent random variables $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ following the same parametric distribution of common parameter $\boldsymbol{\theta}_{12}$ (i.e., the underlying noise-free patch). The evaluation of the similarity between noisy patches can then be rephrased as the following hypothesis test (i.e., a parameter test):

$$\mathscr{H}_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 \equiv \boldsymbol{\theta}_{12} \qquad \text{(null hypothesis)}, \tag{2}$$

$$\mathscr{H}_1 : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 \qquad \text{(alternative hypothesis)}. \tag{3}$$

A similarity criterion $\mathscr{C}_{\boldsymbol{X}_1,\boldsymbol{X}_2}$ (written $\mathscr{C}$ in short) defines a mapping from a pair of noisy patches $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ to a real value. The larger the value of $\mathscr{C}(\boldsymbol{x}_1, \boldsymbol{x}_2)$, the more similar $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are considered to be. For a given similarity criterion $\mathscr{C}$, the *probability of false alarm* (to decide $\mathscr{H}_1$ under $\mathscr{H}_0$) and the *probability of detection* (to decide $\mathscr{H}_1$ under $\mathscr{H}_1$) are defined as:

$$P_{FA}^{\mathscr{C}}(\tau) = \mathbb{P}(\mathscr{C}(\boldsymbol{X}_1, \boldsymbol{X}_2) < \tau; \boldsymbol{\theta}_{12}, \mathscr{H}_0), \tag{4}$$

$$P_{D}^{\mathscr{C}}(\tau) = \mathbb{P}(\mathscr{C}(\boldsymbol{X}_1, \boldsymbol{X}_2) < \tau; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathscr{H}_1). \tag{5}$$

where τ is a real threshold value. Note that the inequality symbols are reversed compared to usual definitions since we consider detecting dissimilarity based on similarity measure $\mathscr{C}$.

According to Neyman-Pearson theorem, the optimal criterion, i.e., the criterion which maximizes P_D for any given P_{FA}, is the likelihood ratio test (see Kay, 1998):

$$\mathcal{L}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{p(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{\theta}_{12}, \mathscr{H}_0)}{p(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathscr{H}_1)}. \tag{6}$$

The application of the likelihood ratio test requires the knowledge of the parameters $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_{12}$ (the noise-free patches) which are, in practice, unavailable. The problem is thus a *composite hypothesis problem*.

Kendall and Stuart (1979) showed that there is no uniformly most powerful (UMP) detector for such *composite hypothesis problem*, i.e, any criteria $\mathscr{C}$ can be defeated by another criteria $\mathscr{C}'$ at a specific false alarm rate:

$$\forall \mathscr{C},\ \exists \mathscr{C}',\ \tau,\ \tau' \text{ such that}$$
$$P_{FA}^{\mathscr{C}}(\tau) = P_{FA}^{\mathscr{C}'}(\tau') \text{ and } P_{D}^{\mathscr{C}}(\tau) < P_{D}^{\mathscr{C}'}(\tau')\ . \tag{7}$$

The research of a universal similarity criterion is then futile. We review in the following seven similarity criteria in the light of dissimilarity detection. We then turn to task-based evaluation of the criteria on natural images.

2.1 Euclidean distance and Gaussian kernel

The usual way to measure the similarity between two noisy patches is to consider their squared Euclidean distance:

$$\mathcal{D}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2. \tag{8}$$

$\mathcal{D}$ is minimal when the two patches $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are identical. It is common to use an exponential kernel of bandwidth $h > 0$, leading to the following similarity criterion:

$$\mathcal{G}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\left(-\tfrac{1}{h}\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2\right), \tag{9}$$

or if noise is correlated with covariance matrix $\boldsymbol{\Gamma}$, by substituting $\mathcal{D}$ with the Mahalanobis distance:

$$\mathcal{G}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\left[-\tfrac{1}{h}(\boldsymbol{x}_1 - \boldsymbol{x}_2)^t \boldsymbol{\Gamma}^{-1}(\boldsymbol{x}_1 - \boldsymbol{x}_2)\right]. \tag{10}$$

Under the assumption of Gaussian noise, all the similarity criteria we consider in the following boil down to this same expression. There is then more than one way to justify or interpret the expression of the similarity criterion $\mathcal{G}$ in that case. For this reason and its link with Gaussian kernels, $\mathcal{G}$ will be referred as the Gaussian kernel.

Under Gaussian noise assumption, the distribution[2] of $\mathcal{G}$ can be used to choose a threshold τ with a given P_{FA} value. It is a *constant false alarm rate detector* (CFAR), which means that a constant P_{FA} can be maintained with a given τ independently of the underlying noise-free patches.

The performance of this criterion however drops when noise departs from a Gaussian distribution. While parameter h in equation (9) could be set globally from the noise variance, difficulties arise when the variance is signal-dependent, and therefore varies between and inside patches. A classical approach to extend the applicability of Euclidean distance to some non-Gaussian noise distributions is to apply a transformation to the noisy patches. The transformation is chosen so that the transformed patches follow a (close to) Gaussian distribution with constant variance (hence their name: variance-stabilization transforms). This leads for instance to the homomorphic approach which maps multiplicative noise to additive noise with stationary variance (see Jain, 1989). This is also the principle of Anscombe transform and its variants used for Poisson noise. These approaches are popular and frequently used, e.g., for density estimation (Brown et al., 2010) or for patch selection (i.e., block-matching) in many denoising algorithms (e.g. Mäkitalo et al., 2010; Boulanger et al., 2010; Mäkitalo and Foi, 2011).

Given an invertible application $\boldsymbol{s}$ which stabilizes the variance for a specific noise pdf, the similarity is computed from the transformed patches:

$$\mathcal{S}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathcal{G}(\boldsymbol{s}(\boldsymbol{x}_1), \boldsymbol{s}(\boldsymbol{x}_2)). \tag{11}$$

An important limitation lies in the non-linear distortion of noise-free patches introduced by $\boldsymbol{s}$. For instance, in the homomorphic approach, the logarithm transforms the contrast of noise-free patches; the performance is affected accordingly. A more fundamental limit is the nonexistence of a variance stabilizing transform $\boldsymbol{s}$ for some distributions.

2.2 Likelihood ratio extensions

Motivated by optimality guarantees of the likelihood ratio test $\mathcal{L}$ given in equation (6), similarity criteria can be defined from statistical tests designed for *composite hypothesis problems*.

The Bayesian likelihood ratio $\mathcal{L}_B$ considers noise-free patches as realizations of random vectors with known *prior* pdf:

$$\mathcal{L}_B(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{p(\boldsymbol{x}_1, \boldsymbol{x}_2; \mathscr{H}_0)}{p(\boldsymbol{x}_1, \boldsymbol{x}_2; \mathscr{H}_1)} = \frac{\int p(\boldsymbol{x}_1|\boldsymbol{\theta}_{12}=\boldsymbol{t}) p(\boldsymbol{x}_2|\boldsymbol{\theta}_{12}=\boldsymbol{t}) p(\boldsymbol{\theta}_{12}=\boldsymbol{t})\, \mathrm{d}\boldsymbol{t}}{\int p(\boldsymbol{x}_1|\boldsymbol{\theta}_1=\boldsymbol{t}_1) p(\boldsymbol{\theta}_1=\boldsymbol{t}_1) \mathrm{d}\boldsymbol{t}_1 \int p(\boldsymbol{x}_2|\boldsymbol{\theta}_2=\boldsymbol{t}_2) p(\boldsymbol{\theta}_2=\boldsymbol{t}_2) \mathrm{d}\boldsymbol{t}_2}. \tag{12}$$

This criterion has been used in the context of classification: Minka (2000) exhibits a relationship between $\mathcal{L}_B$ and the *canonical distance measure* minimizing errors in nearest neighborhood classifiers. He also relates $\mathcal{L}_B$ to mutual information: the more additional knowledge is brought by $\boldsymbol{x}_2$ compared to the observation of $\boldsymbol{x}_1$ alone, the more dissimilar the underlying parameters are (Minka, 1998).

Despite its theoretical performance, this approach suffers from two drawbacks in practice. First, it requires computation of integrals which, depending on the distributions, may not be known in closed form and therefore are time-consuming to evaluate numerically. Second, it requires knowledge of the *prior* pdf. In the absence of a statistical model of noise-free patches, a *non-informative prior* can be used. Jeffreys' *prior* is independent upon the choice of the noise-free patch representation (e.g., testing that two gamma random values share identical standard deviations $\theta_{12} = \sigma$ or identical variances $\theta_{12} = \sigma^2$ leads to the same expression of $\mathcal{L}_B$ when Jeffreys' *prior* are used).

[2] $\log(\mathcal{G})$ follows a Chi square distribution

Rather than modeling noise-free patches as random variables, the generalized likelihood ratio $\mathcal{L}_G$ (GLR) replaces $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_{12}$ in equation (6) by their maximum likelihood estimates (MLE) under each hypothesis:

$$\begin{aligned}\mathcal{L}_G(\boldsymbol{x}_1, \boldsymbol{x}_2) &= \frac{\sup_{\boldsymbol{t}} p(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{\theta}_{12} = \boldsymbol{t}, \mathscr{H}_0)}{\sup_{\boldsymbol{t}_1, \boldsymbol{t}_2} p(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{\theta}_1 = \boldsymbol{t}_1, \boldsymbol{\theta}_2 = \boldsymbol{t}_2, \mathscr{H}_1)} \\ &= \frac{p(\boldsymbol{x}_1; \boldsymbol{\theta}_1 = \hat{\boldsymbol{t}}_{12}) p(\boldsymbol{x}_2; \boldsymbol{\theta}_2 = \hat{\boldsymbol{t}}_{12})}{p(\boldsymbol{x}_1; \boldsymbol{\theta}_1 = \hat{\boldsymbol{t}}_1) p(\boldsymbol{x}_2; \boldsymbol{\theta}_2 = \hat{\boldsymbol{t}}_2)}.\end{aligned} \tag{13}$$

For low levels of noise, the MLE is very close to the true value and $\mathcal{L}_G$ approaches $\mathcal{L}$. As a consequence, the distribution of $\mathcal{L}_G$ is asymptotically known for low noise levels. It results that P_{FA} values associated to any given threshold τ are known: $\mathcal{L}_G$ is asymptotically CFAR (asymptotically to vanishing levels of noise). $\mathcal{L}_G$ is also asymptotically UMP among all invariant tests (see Sect. 3 and Lehmann, 1959).

Compared to the Bayesian likelihood ratio $\mathcal{L}_B$, the generalized likelihood ratio $\mathcal{L}_G$ is easier to implement, since it requires only the computation of the MLE (generally known in closed-form, or estimated in few iterations), and does not require (nor rely on) the definition of a *prior* model.

The main drawback of $\mathcal{L}_G$ lies in the lack of theoretical guarantees on how it behaves in low signal-to-noise ratio (SNR) conditions (i.e., for too small patches according to the noise level). It is known that, for low SNR and specific applications, $\mathcal{L}_G$ can be defeated by other invariant detectors (Kim and Hero III, 2001). This limitation is due to its dependency on MLE which behaves poorly for low SNR (e.g., the $\mathcal{L}_G$ that two random Gaussian vectors share an identical covariance matrix $\boldsymbol{\theta}_{12}$ is undefined since MLE of $\boldsymbol{\theta}_1$ from $\boldsymbol{x_1}$ only would not be positive definite).

2.3 Joint likelihood criteria

Other criteria use the joint likelihood of observations under $\mathscr{H}_0$ to evaluate similarities between noisy data. This leads to the Bayesian joint likelihood criteria (Yianilos, 1995; Seeger, 2002; Matsushita and Lin, 2007; Deledalle et al., 2009b; Teuber and Lang, 2011):

$$\begin{aligned}\mathcal{Q}_B(\boldsymbol{x}_1, \boldsymbol{x}_2) &= p(\boldsymbol{x}_1, \boldsymbol{x}_2; \mathscr{H}_0) \\ &= \int p(\boldsymbol{x}_1 | \boldsymbol{\theta}_1 = \boldsymbol{t}) p(\boldsymbol{x}_2 | \boldsymbol{\theta}_2 = \boldsymbol{t}) p(\boldsymbol{\theta}_{12} = \boldsymbol{t}) \, \mathrm{d}\boldsymbol{t}\end{aligned} \tag{14}$$

or, following the simplification of GLR, the maximum joint likelihood (Alter et al., 2006):

$$\begin{aligned}\mathcal{Q}_G(\boldsymbol{x}_1, \boldsymbol{x}_2) &= \sup_{\boldsymbol{t}} p(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{\theta}_{12} = \boldsymbol{t}, \mathscr{H}_0) \\ &= p(\boldsymbol{x}_1; \boldsymbol{\theta}_1 = \hat{\boldsymbol{t}}_{12}) p(\boldsymbol{x}_2; \boldsymbol{\theta}_2 = \hat{\boldsymbol{t}}_{12}) \, .\end{aligned} \tag{15}$$

Such criteria have been designed to measure the likelihood of sharing a common parameter. However, the likelihood provides relative information compared to the likelihoods of other hypotheses. The evaluation of the joint likelihood under $\mathscr{H}_0$ cannot provide information if it is not confronted against the alternative hypothesis $\mathscr{H}_1$. This leads to non-invariance issues and to the violation of the maximal self-similarity property (Property 2, section 3) as pointed out recently (Teuber and Lang, 2011). Yianilos (1995) already referred to this problem as *the self-recognition paradox*: "there are queries which do not recognize themselves, i.e., even if the query is in the database, some other element may be preferred.". This issue is further discussed in appendix A.

However, $\mathcal{Q}_B$ offers a useful property: it corresponds to an inner product in the space of functions $\boldsymbol{\theta} \mapsto \mathbb{R}$, the feature of $\boldsymbol{x}$ being $(p(\boldsymbol{x}|\boldsymbol{\theta} = \boldsymbol{t}))_{\boldsymbol{t}}$ (Seeger, 2002). The "mutual information" kernel follows this interpretation.

2.4 Mutual information kernel

Given the Bayesian joint criterion $\mathcal{Q}_B(\boldsymbol{x}_1, \boldsymbol{x}_2)$, Seeger (2002) defines a covariance kernel related to the sample mutual information between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$:

$$\mathcal{K}_B(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{\mathcal{Q}_B(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\sqrt{\mathcal{Q}_B(\boldsymbol{x}_1, \boldsymbol{x}_1) \mathcal{Q}_B(\boldsymbol{x}_2, \boldsymbol{x}_2)}}. \tag{16}$$

	Max. self sim.	Eq. self sim.	Id. of indiscernible	Invariance	Asym. CFAR	Asym. UMPI
$\mathcal{Q}_B$	×	×	×	×	×	×
$\mathcal{Q}_G$	×	×	×	×	×	×
$\mathcal{L}_B$	×	×	×	√	×	×
$\mathcal{L}_G$	√	√	√(†)	√	√	√
$\mathcal{K}_B$	√	√	√(‡)	√	×	×
$\mathcal{G}$	√	√	√	×	×	×
$\mathcal{S}$	√(⋆)	√(⋆)	√(⋆)	√(⋆)	√(⋆)	×

Table 1 Properties of the different studied criteria. Legend: (√) the criterion holds, (×) the criterion does not hold. Holds only if the observations are statistically identifiable (†) through their MLE or (‡) through their likelihood (such assumptions are frequently true). (⋆) Holds only for an exact variance stabilizing transform $s(\cdot)$ (such an assumption is usually wrong). The proofs of all these properties are available in Appendix C.

Since $\mathcal{Q}_B$ can be seen as an inner product in the feature space, $\mathcal{K}_B$ corresponds to a cosine in the feature space $\mathcal{K}_B(\boldsymbol{x}_1, \boldsymbol{x}_2) = \dfrac{\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle}{\|\boldsymbol{x}_1\| \|\boldsymbol{x}_2\|}$. Seeger shows that it is a kernel covariance matrix and coins it the mutual information kernel. Algorithms can be adapted to the noise pdf using the so-called *kernel tricks*, i.e., by considering higher dimensional space while never mapping the data in practice. This leads for instance to non-linear support vector machines or non-linear principal component analysis. Note that a *prior*-less extension using MLE would lead to the generalized likelihood ratio $\mathcal{L}_G$. Compared to $\mathcal{L}_G$, the main limitation of the mutual information kernel is its dependency on the *prior* pdf and the lack of asymptotic results.

Among criteria involving probability densities, $\mathcal{L}_B$, $\mathcal{L}_G$ and $\mathcal{K}_B$ are dimensionless thanks to their definition as ratios of likelihoods (in terms of dimensional analysis), which is not the case for $\mathcal{Q}_B$ and $\mathcal{Q}_G$. We show in section 3 that similarity criteria that are not dimensionless lack some important properties. For this reason, we will refer to $\mathcal{L}_B$, $\mathcal{L}_G$ and $\mathcal{K}_B$ as normalized criteria and $\mathcal{Q}_B$ and $\mathcal{Q}_G$ as unnormalized criteria.

3 Desirable properties for similarity criteria

Beyond the theoretical grounding of each of the criteria described in the previous section, there are some desirable properties that are necessary to compare together given similarity criteria.

It is natural to require that the similarity between two patches does not depend on the order in which the patches are compared:

Property 1 (Symmetry) *The similarity between patch $\boldsymbol{x}_1$ and patch $\boldsymbol{x}_2$ is equal to the similarity between patch $\boldsymbol{x}_2$ and patch $\boldsymbol{x}_1$:*

$$\mathscr{C}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathscr{C}(\boldsymbol{x}_2, \boldsymbol{x}_1).$$

All previously considered criteria are symmetrical.

For some criteria, it may occur that a distinct pair $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ is more similar than the pair formed by repeating observation $\boldsymbol{x}_1$: $(\boldsymbol{x}_1, \boldsymbol{x}_1)$. This phenomenon is called *the self-recognition paradox* (Yianilos, 1995). It is desirable to ask for maximal self-similarity:

Property 2 (Maximal self-similarity) *No distinct pair $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ can be more similar than the observed patch $\boldsymbol{x}_1$ is similar to itself:*

$$\forall \boldsymbol{x}_1, \boldsymbol{x}_2,\ \mathscr{C}(\boldsymbol{x}_1, \boldsymbol{x}_2) \leq \mathscr{C}(\boldsymbol{x}_1, \boldsymbol{x}_1).$$

Joint likelihood criteria do not verify property 2. For a proof, consider a noise distribution with a variance depending on the signal level, like gamma distribution that models speckle noise. For the pixel-based comparison, we have (table 2 with $L = 1$): $\mathcal{Q}_B(x_1, x_2) = (x_1 + x_2)^{-2}$. Choose observation x_1 to be $x_1 = 2x_2$. Since $\mathcal{Q}_B(x_1, x_2) = (3\,x_2)^{-2}$ is larger than $\mathcal{Q}_B(x_1, x_1) = (4\,x_2)^{-2}$, property 2 is violated.

Most criteria with a normalization like the generalized likelihood ratio $\mathcal{L}_G$ and mutual information kernel $\mathcal{K}_B$ fulfill property 2 (see table 1).

Property 2 does not guarantee that a pair $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ of distinct patches is always less similar than a pair formed by the repetition of a third observation $(\boldsymbol{x}_3, \boldsymbol{x}_3)$. A supplementary property is needed:

Property 3 (Equal self-similarities) *Two pairs of identical patches always have equal similarity:*

$$\forall \boldsymbol{x}_1, \boldsymbol{x}_2,\ \mathscr{C}(\boldsymbol{x}_1, \boldsymbol{x}_1) = \mathscr{C}(\boldsymbol{x}_2, \boldsymbol{x}_2).$$

Criteria $\mathcal{L}_G$ and $\mathcal{K}_B$ verify both property 2 and 3 and their self-similarities are always equal to one (see table 1).

Additionally, one may ask that the criterion is maximal only in case of strict patch equality, and for every comparison between identical patches:

Property 4 (Identity of the indiscernibles) *The similarity reaches its maximum if and only if the compared patches are identical:*

$$\forall \boldsymbol{x}_1, \boldsymbol{x}_2,\ \mathscr{C}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \max_{\boldsymbol{x}} \mathscr{C}(\boldsymbol{x}, \boldsymbol{x}) \quad \textit{iif} \quad \boldsymbol{x}_1 = \boldsymbol{x}_2.$$

For likelihood-based criteria, it is clear that property 4 cannot be verified if two different observations lead to the same likelihoods. We need then to require that the observations be statistically identifiable through their likelihood:

$$\forall \boldsymbol{x}_1, \boldsymbol{x}_2,\ \ \boldsymbol{x}_1 \neq \boldsymbol{x}_2 \ \Rightarrow\ \exists \boldsymbol{\theta}, p(\boldsymbol{x}_1|\boldsymbol{\theta}) \neq p(\boldsymbol{x}_2|\boldsymbol{\theta}). \tag{17}$$

Provided that observations are statistically identifiable through their likelihood, property 4 is fulfilled by the mutual information kernel $\mathcal{K}_B$. For $\mathcal{L}_G$ we require that the observations are statistically identifiable through their MLE, i.e., that the likelihood has a unique maximum and:

$$\forall \boldsymbol{x}_1, \boldsymbol{x}_2,\ \ \boldsymbol{x}_1 \neq \boldsymbol{x}_2 \ \Rightarrow\ \operatorname*{argmax}_{\theta} p(\boldsymbol{x}_1|\boldsymbol{\theta}) \neq \operatorname*{argmax}_{\theta} p(\boldsymbol{x}_2|\boldsymbol{\theta}). \tag{18}$$

The statistical answer of a similarity criterion should not depend on the choice of a specific noisy patch representation:

Property 5 (Invariance) *Let $\boldsymbol{g}$ be an invertible and differentiable function mapping random vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ to random vectors $\boldsymbol{X}_1' = \boldsymbol{g}(\boldsymbol{X}_1)$ and $\boldsymbol{X}_2' = \boldsymbol{g}(\boldsymbol{X}_2)$. Let $\mathscr{C}_{\boldsymbol{X}_1,\boldsymbol{X}_2}$ and $\mathscr{C}_{\boldsymbol{X}_1',\boldsymbol{X}_2'}$ be, respectively, the similarity criteria derived from the family of parametric distributions followed by $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ (resp. $\boldsymbol{X}_1'$ and $\boldsymbol{X}_2'$). An invariant similarity criterion leads to the same similarity whether it is evaluated with $\mathscr{C}_{\boldsymbol{X}_1,\boldsymbol{X}_2}$ on $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ or with $\mathscr{C}_{\boldsymbol{X}_1',\boldsymbol{X}_2'}$ on $(\boldsymbol{g}(\boldsymbol{x}_1), \boldsymbol{g}(\boldsymbol{x}_2))$:*

$$\forall \boldsymbol{x}_1, \boldsymbol{x}_2,\ \mathscr{C}_{\boldsymbol{X}_1,\boldsymbol{X}_2}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathscr{C}_{\boldsymbol{X}_1',\boldsymbol{X}_2'}(\boldsymbol{g}(\boldsymbol{x}_1), \boldsymbol{g}(\boldsymbol{x}_2)).$$

Due to their unnormalization, joint likelihood criteria $\mathcal{Q}_B$ and $\mathcal{Q}_G$ typically do not have the invariance property. Transforming the patches by, for example, taking their squared value leads to modified probability densities with different dimensions (i.e., different units). The change of variables leads to a similarity criterion with a different scaling from the original one. Normalized criteria, defined as a ratio of probability densities, are the only ones to fulfil property 5.

Deciding for dissimilarity is done by thresholding the similarity: patches $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are considered dissimilar if $\mathscr{C}(\boldsymbol{x}_1, \boldsymbol{x}_2) < \tau$. The associated probability of false alarm $P_{FA}^{\mathscr{C}}$ is the probability that $\mathscr{C}(\boldsymbol{x}_1, \boldsymbol{x}_2) < \tau$ although $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2\ (= \boldsymbol{\theta}_{12})$, i.e., that the detected dissimilarity is only due to noise.

Property 6 (Constant false alarm rate) *For all threshold τ, the probability of false alarm $P_{FA}^{\mathscr{C}}$ of similarity criterion $\mathscr{C}$ is independent on the noise-free patch $\boldsymbol{\theta}_{12}$:*

$$\forall \tau,\ P_{FA}^{\mathscr{C}}(\tau) \textit{ does not depend on } \boldsymbol{\theta}_{12}.$$

name	pdf	$\mathcal{Q}_B$	$\mathcal{Q}_G$	$\mathcal{L}_B$	$\mathcal{L}_G$	$\mathcal{K}_B$	$\mathcal{S}$	$\mathcal{G}$
Gaussian	$\frac{e^{-\frac{(x-\theta)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}$	$e^{-(x_1-x_2)^2}$						
Gamma	$\frac{L^L x^{L-1} e^{-\frac{Lx}{\theta}}}{\Gamma(L)\theta^L}$	$\frac{1}{x_1 x_2}\left(\frac{x_1 x_2}{(x_1+x_2)^2}\right)^L$		$\frac{x_1 x_2}{(x_1+x_2)^2}$			$e^{-\left(\log\frac{x_1}{x_2}\right)^2}$	
Poisson	$\frac{\theta^x e^{-\theta}}{x!}$	$\frac{\Gamma'(x_1+x_2)}{2^{x_1+x_2}x_1!x_2!}$	$\frac{(x_1+x_2)^{x_1+x_2}}{(2e)^{x_1+x_2}x_1!x_2!}$	$\frac{\Gamma'(x_1+x_2)}{2^{x_1+x_2}\Gamma'(x_1)\Gamma'(x_2)}$	$\frac{(x_1+x_2)^{x_1+x_2}}{2^{x_1+x_2}x_1^{x_1}x_2^{x_2}}$	$\frac{\Gamma'(x_1+x_2)}{\sqrt{\Gamma'(2x_1)\Gamma'(2x_2)}}$	$e^{-\left(\sqrt{x_1+a}-\sqrt{x_2+a}\right)^2}$	

Table 2 Instances of the seven criteria for Gaussian, gamma and Poisson noise (parameters σ and L are fixed and known). All Bayesian criteria are obtained with Jeffreys' *priors* (resp. $1/\sigma$, $\sqrt{L}/\theta$, $\sqrt{1/\theta}$). All constant terms which do not affect the detection performance are omitted. For clarity reason, we define $\Gamma'(x) = \Gamma(x+0.5)$ and the Anscombe constant $a = 3/8$.

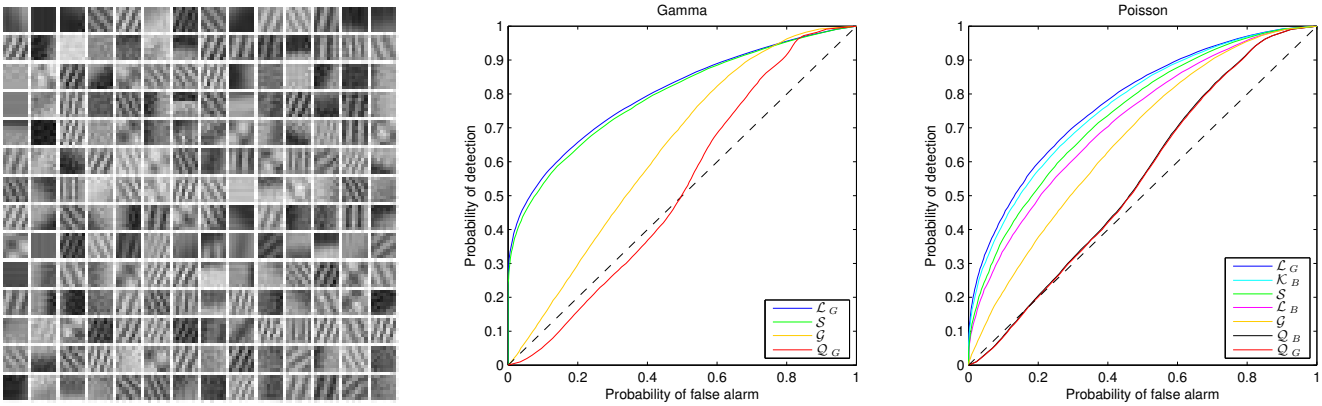


Fig. 1 (left) Patch dictionary. (center) ROC curve obtained under gamma noise and (right) ROC curve obtained under Poisson noise. In both experiments, the SNR over the whole dictionary is about 1 dB.

The Gaussian kernel $\mathcal{G}$ is an example of a criterion which does not guarantee property 6. For instance in case of two Poisson noisy values x_1 and x_2, $\mathbb{E}\left[\|x_1 - x_2\|_2^2 \mid \mathscr{H}_0\right] = 2\theta_{12}$, hence, the distribution of $P_{FA}^{\mathcal{G}}$ clearly depends on θ_{12}. Due to the efficiency of MLE with respect to the noise level, $\mathcal{L}_G$ is asymptotically CFAR (see Kay, 1998).

Based on the properties presented so-far, a proper similarity criterion can be selected. However, it is also important to compare the relative performance of similarity criteria. While we mentioned in Sect. 2 that there is no UMP detector for the considered *composite hypothesis problem*, the optimality can be studied on a subset of similarity criteria.

Property 7 (Uniformly Most Powerful Invariant) *A similarity criterion $\mathscr{C}$ is said to be the uniformly most powerful invariant (UMPI) if it is an invariant criterion (property 5) and its probability of detection is larger than that of all other invariant criteria $\mathscr{C}'$ for any given false-alarm rate:*

$$\forall \tau,\ \tau' \quad P_{FA}^{\mathscr{C}}(\tau) = P_{FA}^{\mathscr{C}'}(\tau') \Rightarrow P_D^{\mathscr{C}}(\tau) \geq P_D^{\mathscr{C}'}(\tau')\ . \tag{19}$$

Asymptotically to the noise level, $\mathcal{L}_G$ is UMPI (see Lehmann, 1959). All other invariant criteria are then asymptotically defeated by $\mathcal{L}_G$.

Table 1 summarizes the properties of each of the seven considered criteria. The unnormalized criteria $\mathcal{Q}_B$ and $\mathcal{Q}_G$ fulfil none of the properties while the generalized likelihood ratio $\mathcal{L}_G$ fulfils all of them. Note that some properties require that observations are statistically identifiable. Such assumptions are generally true, except, e.g., for multi-modal distributions or when two different observations lead to equal likelihood function (e.g., a Gaussian distribution with zero mean and unknown variance leads to the same likelihood function for the observation of x or $-x$). Finally, note that $\mathcal{S}$ verifies most of these properties when the function $\boldsymbol{s}$ exists, which is generally not the case, e.g., there is no exact variance stabilization for the Poisson distribution or the Cauchy distribution.

		Noisy	$\mathcal{Q}_G$	$\mathcal{L}_G$	$\mathcal{S}$	$\mathcal{G}$	Noisy	$\mathcal{Q}_B$	$\mathcal{Q}_G$	$\mathcal{L}_B$	$\mathcal{L}_G$	$\mathcal{K}_B$	$\mathcal{S}$	$\mathcal{G}$
		Gamma					Poisson							
Strong noise levels	barbara	5.86	20.25	**20.97**	20.90	20.33	5.68	20.25	20.25	20.52	**20.68**	20.65	20.59	20.42
	boat	5.32	20.90	**21.47**	21.42	20.97	5.23	20.90	20.90	21.11	**21.21**	21.19	21.15	21.04
	bridge	6.09	18.44	**19.21**	19.16	18.49	5.83	18.36	18.36	18.65	**18.81**	18.78	18.72	18.53
	cameraman	5.54	18.56	**20.88**	20.87	7.48	5.59	18.61	18.61	19.17	**19.56**	19.49	19.37	19.01
	couple	5.98	20.93	**21.54**	21.51	20.99	5.55	20.91	20.91	21.11	**21.20**	21.18	21.15	21.04
	fingerprint	4.60	15.34	**16.30**	16.22	15.57	4.87	15.48	15.48	16.18	**16.41**	16.38	16.30	15.96
	hill	6.35	20.18	**20.68**	20.61	20.20	5.88	20.13	20.13	20.41	**20.54**	20.52	20.47	20.31
	house	4.84	20.54	**21.20**	21.13	20.64	4.94	20.48	20.49	20.81	**20.97**	20.94	20.88	20.67
	lena	5.64	22.14	**22.89**	22.83	22.23	5.44	22.14	22.15	22.44	**22.59**	22.56	22.49	22.30
	man	6.47	21.56	**22.16**	22.10	21.64	5.89	21.55	21.55	21.77	**21.89**	21.87	21.82	21.69
	mandril	5.52	20.22	**20.44**	20.41	20.27	5.31	20.23	20.23	20.34	**20.38**	20.37	20.36	20.30
	peppers	5.56	18.59	**20.44**	20.43	18.65	5.46	18.55	18.56	19.09	**19.46**	19.38	19.25	18.88
Medium noise levels	barbara	14.34	22.61	25.66	**25.67**	23.83	14.43	23.59	23.57	25.43	25.40	25.41	**25.44**	24.79
	boat	13.78	23.40	**25.50**	**25.50**	24.06	13.99	24.00	23.98	25.28	25.26	25.27	**25.29**	24.74
	bridge	14.58	20.17	**22.36**	**22.36**	21.01	14.58	21.06	21.04	22.30	22.29	22.30	**22.31**	21.84
	cameraman	13.96	23.88	**25.04**	25.01	14.93	14.33	23.63	23.57	25.01	25.02	25.02	**25.03**	24.22
	couple	14.37	23.19	**25.08**	25.06	23.68	14.31	23.54	23.52	**24.88**	24.85	24.86	**24.88**	24.29
	fingerprint	13.00	18.37	21.88	**21.89**	20.27	13.62	20.59	20.58	22.03	21.99	22.00	**22.04**	21.60
	hill	14.80	21.46	**24.24**	**24.24**	22.47	14.62	22.49	22.48	**23.98**	23.96	23.97	**23.98**	23.36
	house	13.35	22.52	26.33	**26.34**	24.36	13.73	24.36	24.34	**26.58**	26.57	26.57	**26.58**	25.76
	lena	14.09	24.61	27.71	**27.72**	25.61	14.20	25.57	25.55	**27.40**	27.37	27.38	**27.40**	26.58
	man	14.88	23.49	26.00	**26.01**	24.50	14.64	24.08	24.06	25.66	25.65	25.66	**25.67**	25.09
	mandril	14.02	21.61	**23.20**	**23.20**	22.22	14.03	22.18	22.17	23.03	23.01	23.02	**23.04**	22.68
	peppers	14.02	22.95	**25.54**	25.51	23.41	14.20	23.38	23.35	**25.45**	25.41	25.43	**25.45**	24.41

Table 3 PSNR values obtained by NL-Means denoising using different similarity criteria on 13 standard images corrupted by gamma noise and Poisson noise with (top) strong noise levels and (bottom) medium noise levels.

4 Evaluation of similarity criteria

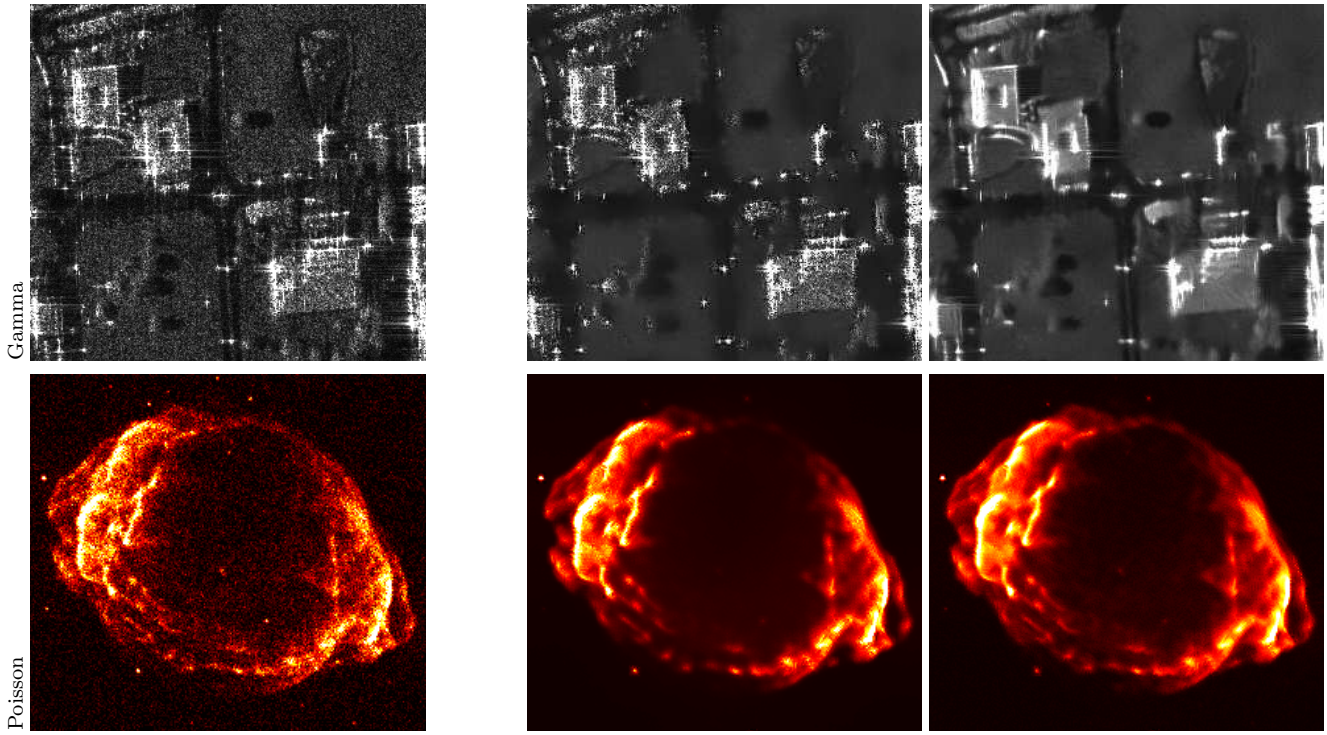
All criteria have been derived[3] in the case of gamma or Poisson noise (table 2). In practice, Bayesian criteria are more difficult to obtain due to integrations over the noise-free patch space. While all criteria are equivalent for Gaussian noise, there are four different expressions for gamma noise and they are all different for Poisson noise. The distinction seems to emerge with the "complexity" induced by the noise distribution (by considering that gamma noise is more challenging than Gaussian noise, and that Poisson noise is even more challenging).

4.1 Performance for patch discrimination

We evaluate the relative performance of the seven aforementioned criteria on a dictionary composed of 196 noise-free patches of size $N = 8 \times 8$. The noise-free patches have been obtained using the k-means on patches extracted from the classical 512×512 *Barbara* image. The noisy patches are noisy realizations of the noise-free patches under gamma or Poisson noise with an overall SNR of about 1 dB. All criteria are evaluated for all pairs of noisy patches. The process is repeated 200 times with independent noise realizations.

Numerically, the performance of the similarity criteria is given in term of their *receiver operating characteristic* (ROC) curve, i.e., the curve of P_D with respect to P_{FA}. Results are given in Figure 1. For small P_{FA}, the generalized likelihood ratio (GLR) is the most powerful followed by the mutual information kernel, the Bayesian likelihood ratio and the variance stabilization criteria. Other criteria behave poorly for such a low SNR. Such behaviors agree with the theoretical predictions. The poor performance of the joint likelihood based criteria (worse than a detector that would not make use of the data) can arise from their non-invariance and the induced self-similarity paradox. The low performance of $\mathcal{G}$ is certainly due to its non-adaptivity to either the target noise or the target noise variance. The variance stabilization criteria are always defeated by GLR, due to the distortions of the noise-free patches as well as the consideration of the noise variance only, instead of the full noise pdf. The worse performance of Bayesian criteria compared to criteria that use MLE may be due to the low quality of the *prior* pdf (non-informative Jeffreys' prior have been used).

[3] the complete derivations are available in Appendix B



(a) Noisy image (b) Gaussian kernel $\mathcal{G}$ (c) GLR $\mathcal{L}_G$

Fig. 2 Results of NL-Means on (a) noisy images using (b) the Gaussian kernel ($\mathcal{G}$) and (c) the generalized likelihood ratio ($\mathcal{L}_G$). The images are (top) a SAR image of two buildings suffering from gamma noise (©CNES) and (bottom) an X-ray image of a supernova explosion in the Milky Way of the supernova remnant G1.9+0.3 suffering from Poisson noise (with a colormap varying smoothly from black through shades of red, orange, and yellow, to white) (image courtesy to Chandra X-ray Observatory – data identifier: `ADS/Sa.CXO#Contrib/ ChandraDeepField`).

4.2 Application to image denoising

Patch correspondence is at the heart of most recent image denoising approaches since the introduction of the NL-Means (Buades et al., 2005a). It has led to the elaboration of powerful denoising filters, such as the so-called BM3D algorithm (Dabov et al., 2007) or the non-local sparse model proposed in (Mairal et al., 2009). Most attempts to adapt such approaches for non-Gaussian noise relies on variance stabilization (e.g. Mäkitalo et al., 2010; Boulanger et al., 2010; Mäkitalo and Foi, 2011). Few authors try to extend the NL-Means by directly considering non-Gaussian noise distributions (Kervrann et al., 2007; Deledalle et al., 2009b).

While local filters lead to biases and resolution loss, non-local techniques are known to efficiently reduce noise and preserve structures. Instead of combining neighboring pixels, the non-local means average similar pixels. Let $x(p)$ and $\boldsymbol{x}(p)$ be respectively the observed noisy value and the observed noisy patch at pixel $p \in \Omega$ and $\theta(p)$ and $\boldsymbol{\theta}(p)$ its underlying noise-free value and noise-free patch. The NL-Means define the estimate $\hat{\theta}(p)$ as a weighted average:

$$\hat{\theta}(p) = \frac{\sum_q \mathscr{C}(\boldsymbol{x}(p), \boldsymbol{x}(q))^{1/h} x(q)}{\sum_q \mathscr{C}(\boldsymbol{x}(p), \boldsymbol{x}(q))^{1/h}} \tag{20}$$

where q is a pixel index located in a search window centered on p, and $h > 0$ is a filtering parameter. The similarity criterion $\mathscr{C}(\boldsymbol{x}(p), \boldsymbol{x}(q))$, through the power function $(.)^{1/h}$, plays the role of a data-driven weight depending on the similarity between two patches centered around pixels of indices p and q respectively. While patch-similarity is originally defined by the Gaussian kernel $\mathcal{G}$, we suggest comparing the denoising performance of the NL-Means when the similarity criterion is substituted by one of the seven aforementioned criteria.

We evaluate first the denoising performance of NL-Means obtained using each of the 7 similarity criteria on 13 standard images synthetically damaged by gamma or Poisson noise. The NL-Means is used with a 21×21

search window and 7×7 patches. The filtering parameter h as well as the central weight $\mathscr{C}(\boldsymbol{x}(p), \boldsymbol{x}(p))$ should be selected from the statistics of the similarity criterion $\mathscr{C}$ under $\mathscr{H}_0$ (Kervrann and Boulanger, 2008; Salmon, 2010). Unfortunately, such solutions cannot be investigated here since some of the studied criteria are not CFAR: the statistics vary locally with respect to $\theta(p)$. The central weight should rather be replaced with the maximum of the weights in the search window, following the solution proposed in (Buades et al., 2009). Here, since our motivation is to compare patch similarity criteria, we have decided to use the true noise-free image $\boldsymbol{\theta}$ to select the best value of h for each criteria. In practice, we apply a gradient descent on h to optimize the mean square error $\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2^2$. This allows us to compare similarity criteria in the most favorable case when each denoiser reaches its optimal performance.

Denoising performance is given in terms of the peak signal to noise ratio (PSNR) defined by:

$$PSNR(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = 10 \log_{10} \frac{255^2}{\frac{1}{|\Omega|}\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2^2}. \tag{21}$$

Table 3 displays the obtained PSNR values. Two levels of noise are considered, the first one, very strong, leads to a noisy image with a PSNR around 5dB, and the second one, medium, provides a PSNR around 14dB. For strong noise levels, the generalized likelihood ratio $\mathcal{L}_G$ outperforms all other similarity criteria while for medium noise levels, the criterion based on variance stabilizations works generally better. In medium/low levels of noise, the variance stabilization based criterion $\mathcal{S}$ can outperform $\mathcal{L}_G$. When the noise level is weak, the problem of weight definition is less a problem of detecting identical patches under noise than a matter of selecting patches with "close" noise-free patches (the noise component becomes negligible). Compared to $\mathcal{L}_G$, the properties provided by Euclidean distances can then be preferable in this context, since it defines a reasonable metric on the space of noiseless patches. A generalized likelihood ratio testing that $\boldsymbol{\theta}_1$ is close to $\boldsymbol{\theta}_2$ could be more adapted to the denoising problem, i.e.: $\mathscr{H}_0 : \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^2 < \epsilon$, where ϵ is a real positive value. This different definition of similarity could be the topic for future work.

Figure 2 provides a visual comparison of the use of the Gaussian kernel $\mathcal{G}$ and the generalized likelihood ratio $\mathcal{L}_G$ on real data. The first one is a synthetic aperture radar (SAR) image of two buildings. SAR data suffers from speckle noise modeled by a gamma distribution. The second one is an X-ray image of a supernova explosion in the Milky Way of the supernova remnant G1.9+0.3. Due to low-light conditions, such images suffer from Poisson noise. Without knowledge of $\boldsymbol{\theta}$, the methodology of Van De Ville and Kocher (2009) has been used to automatically select the value of h that maximizes an estimate of the mean square error. We have already proposed in previous work an extension of this approach for Poisson noise (Deledalle et al., 2010) and an extension for gamma noise has been derived in the same vein following (Hudson, 1978). In both cases $\mathcal{G}$ blurs dark areas and leaves noise in bright areas, GLR allows to reduce the noise level everywhere in the image with a similar amount of smoothing.

Note that the results provided here could be improved by refining weights using the similarity between pre-estimated patches as done in (Deledalle et al., 2009b, 2010). Our motivation here is only to provide a fair comparison between similarity criteria, and therefore we have chosen not to refine weights to avoid interferences with pre-estimation procedures. Note that the performance of GLR for denoising SAR images has also been demonstrated in collaborative filtering (Parrilli et al., 2010).

4.3 Application to stereo-vision

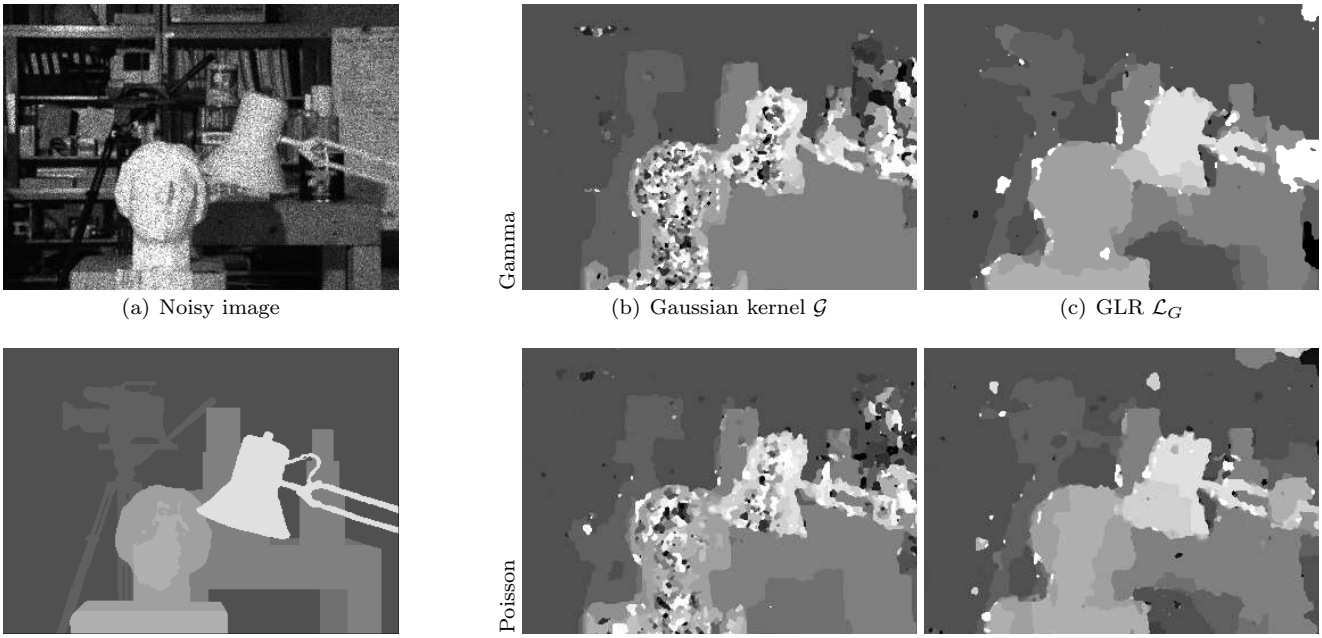
Stereo-vision is one of the tasks in computer vision which extensively uses patches. Given two images of the same scene, the purpose is to estimate the depth of the image parts. Using epipolar geometry, each pixel $p \in \Omega$ of one image has a corresponding pixel q at the same line in the other image (omitting the occlusion issues). The horizontal shift between these two pixels is called the disparity. The initial problem is then reduced to the estimation of a disparity map $\boldsymbol{d}$ (see Hartley and Zisserman, 2000). Given the disparity map, each patch $\boldsymbol{x}_1(p)$ should be similar to the patch $\boldsymbol{x}_2(p + \boldsymbol{d}(p)\overrightarrow{h})$ where $\overrightarrow{h}$ is a unit vector directed on the horizontal orientation.

The definition of patch similarity is then central to stereo-vision. Note however that two patches $\boldsymbol{x}_1(p)$ and $\boldsymbol{x}_2(q)$ can be similar while p and q are not corresponding pixels (e.g. in homogeneous regions or on repetitive patterns). As a consequence, many works introduce a prior knowledge on the solution to regularize the disparity map. Boykov et al. (1998) suggest that disparity maps are piece-wise constant. An estimate of the disparity map

Noisy	$\mathcal{Q}_G$	$\mathcal{L}_G$	$\mathcal{S}$	$\mathcal{G}$
Gamma				
8.47	3.21	**2.06**	**2.06**	3.67
13.18	2.15	**1.57**	1.61	2.24
21.46	1.45	**1.42**	1.47	1.75

Noisy	$\mathcal{Q}_B$	$\mathcal{Q}_G$	$\mathcal{L}_B$	$\mathcal{L}_G$	$\mathcal{K}_B$	$\mathcal{S}$	$\mathcal{G}$
Poisson							
8.37	2.79	2.79	**2.14**	**2.14**	2.18	2.19	2.37
13.97	2.15	2.17	1.86	1.80	**1.74**	1.80	2.02
20.96	1.49	1.51	1.52	**1.46**	1.47	1.50	1.69

Table 4 RMSE values for disparity maps computed with each of the similarity criteria. The disparity maps are regularized with a Potts prior. Different levels of gamma or Poisson noise, expressed in terms of noisy image PSNR, are considered.



(a) Noisy image (b) Gaussian kernel $\mathcal{G}$ (c) GLR $\mathcal{L}_G$

(d) Ground truth (e) Gaussian kernel $\mathcal{G}$ (f) GLR $\mathcal{L}_G$

Fig. 3 Results of a stereo vision approach on a standard pair of noisy stereo views. (a) One of the noisy input images, (d) the ground truth (i.e. the target disparity map) and the estimated disparity maps obtained on the pair damaged by (b-c) gamma and (e-f) Poisson noise. The method is based on energy minimization using either (b,e) the Gaussian kernel $\mathcal{G}$ or (c,f) the generalized likelihood ratio $\mathcal{L}_G$.

can then be obtained by solving the following optimization problem:

$$\hat{\boldsymbol{d}} = \underset{d}{\operatorname{argmax}} \sum_{p} -\log \mathscr{C}\big(\boldsymbol{x}_1(p), \boldsymbol{x}_2(p + \boldsymbol{d}(p)\,\overrightarrow{h}) \big) \\ -\lambda \sum_{p \sim q} \delta(\boldsymbol{d}(p) - \boldsymbol{d}(q)). \tag{22}$$

where $p \sim q$ denotes two neighboring pixels, $\delta(.)$ is the Dirac delta function and the Lagrangian multiplier $\lambda > 0$ acts as a regularization parameter. Thanks to the patch similarity criteria $\mathscr{C}$, the first term measures the data fidelity of the solution. The second term assesses the regularity of the solution: it corresponds to the Potts model which penalizes transitions in $\boldsymbol{d}$. Satisfying solutions of such discrete optimization problems can be iteratively obtained by graph cuts with the α-β swap strategy described in (Boykov et al., 2001).

While the patch-similarity is usually defined by the Gaussian kernel $\mathcal{G}$, or equivalently by the Euclidean distance usually referred to as the sum of squared differences (SSD), we suggest comparing stereo-vision performance of the model of eq. (22) when the similarity criterion is substituted by one of the seven aforementioned criteria.

We evaluate our estimator using each of the 7 similarity criteria on the classical *tusbaka* pair of stereo images. The patches are of size 7×7 and the optimal disparities are researched between 0 and 15 pixels. For the same reason as with the image denoising task, we have decided to use the true disparity map $\boldsymbol{d}$ to select the best possible value of λ for each criterion. In practice, an exhaustive research has been done. This allows comparing similarity criteria in the most favorable case when each estimator reaches its optimal performance.

As a numerical performance criterion, we have computed the root mean square error (RMSE), defined by

$$RMSE(\hat{\boldsymbol{d}}, \boldsymbol{d}) = \sqrt{\frac{1}{|\Omega|}\|\boldsymbol{d} - \hat{\boldsymbol{d}}\|_2^2} \tag{23}$$

for the results obtained by the use of the seven similarity criteria. Table 4 displays the RMSE values. Three levels of noise are considered, the first one, strong, leads to a noisy image with a PSNR around 8dB, the second one, medium, provides a PSNR around 13.5dB, and the third one, low, provides a PSNR around 21dB. The generalized likelihood ratio $\mathcal{L}_G$ challenges $\mathcal{L}_B$, $\mathcal{K}_B$ and $\mathcal{S}$ whatever the noise level and clearly outperforms $\mathcal{Q}_B$, $\mathcal{Q}_G$ and $\mathcal{G}$. Unlike for denoising, $\mathcal{L}_G$ behaves in medium/low levels of noise as good as in strong levels since here the problem is directly related to detecting identical patches rather a matter of selecting patches with "close" noise-free patches.

Figure 3 shows the visual comparison on this standard pair of stereo views damaged by gamma or Poisson noise. In both cases, the use of $\mathcal{G}$, i.e., SSD, leads to a disparity map over-regularized in dark areas and under-regularized in bright areas: there is no global regularization parameter λ offering the same amount of smoothing everywhere in the image. Since GLR is CFAR, we get the same level of regularization both in dark and bright areas for a global regularization parameter λ.

4.4 Application to motion tracking

Motion tracking, object tracking or optical flow estimation are classical problems involving the matching of image parts (e.g. Horn and Schunck, 1981; Lowe, 1992; Comaniciu et al., 2003). Here, we focus on the velocity estimation problem of a flowing Alpine glacier using a pair of synthetic aperture radar (SAR) images. SAR images provide scattering information which can be used under any weather conditions for glacier monitoring. Such images present a multiplicative speckle noise commonly modeled by gamma distributions (Goodman, 1976). The use of a similarity criterion robust to the statistics of the SAR intensity is then essential for the estimation of the displacement field.

Given two registered images of the same glacier sensed at different dates, the purpose is to estimate a displacement field characterizing at each position the local velocity of the glacier. Assuming that the movement is collinear to the glacier orientation, we only have to estimate the magnitude of the velocity. This quantity can be estimated by researching the patches of one acquisition which are similar to those present in the other acquisition along the glacier movement direction.

For the same reasons as in the stereo-vision problem, the solution has to be regularized. Since glacier movement is assumed to be smooth, we propose here to use the total-variation (TV) model[4] whose penalization depends on the height of the transitions. This leads to the following optimization problem:

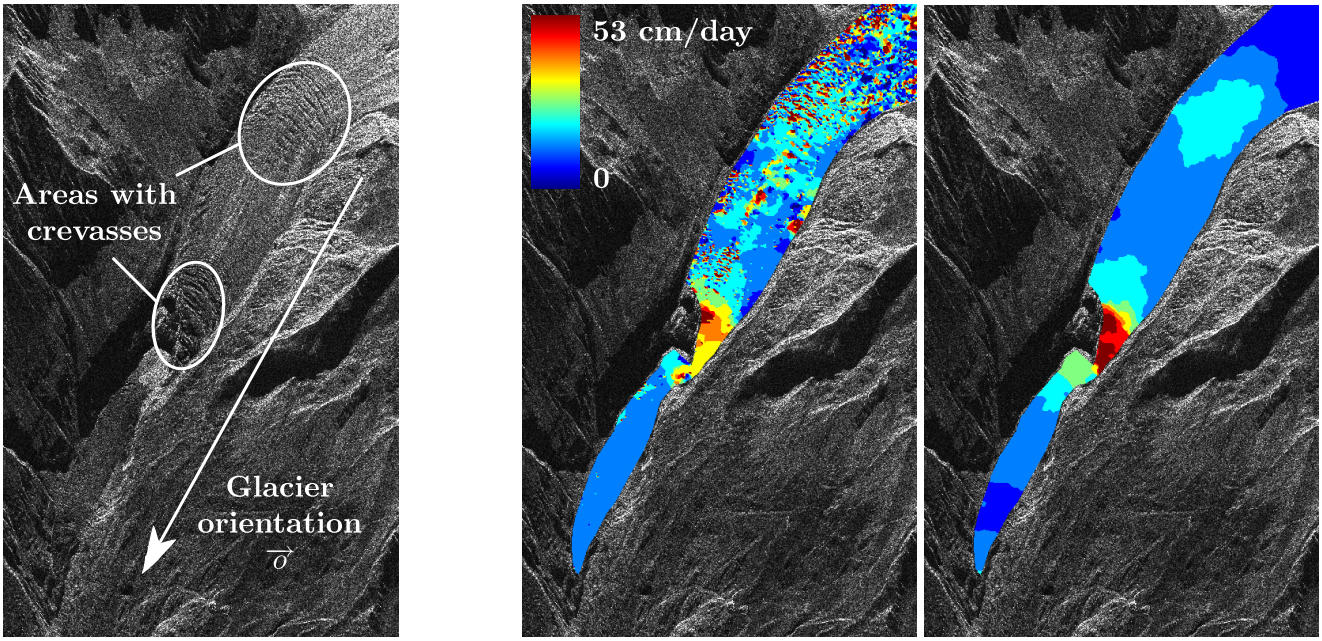
$$\begin{aligned}\hat{\boldsymbol{d}} = \operatorname*{argmax}_{\boldsymbol{d}} & \sum_p -\log \mathscr{C}\big(\, \boldsymbol{x}_1(p),\ \boldsymbol{x}_2(p + \boldsymbol{d}(p)\overrightarrow{o})\,\big) \\ & + \lambda \sum_{p\sim q} |\boldsymbol{d}(p) - \boldsymbol{d}(q)|\end{aligned} \tag{24}$$

where $p \sim q$ denotes two neighboring pixels, the Lagrangian multiplier $\lambda > 0$ acts as a regularization parameter and $\overrightarrow{o}$ is a unit vector directed along the glacier orientation. Optimal solutions of such discrete optimization problems can be obtained by graph cuts using the graph construction described in (Ishikawa, 2003).

We suggest now to compare the quality of the estimated displacement fields obtained by solving (24), when using either the Gaussian kernel $\mathcal{G}$ or the generalized likelihood ratio $\mathcal{L}_G$.

Figure 4 shows the estimated magnitude of the displacement field obtained on two SAR images of the lower part of the glacier of Argentière (French Alps) sensed by TerraSAR-X on September 29th, 2008 and October 21th, 2008 respectively. The two SAR images have been previously co-registered on static areas. They have a resolution cell of 1.36×2.04 meters in line of sight and azimuth directions respectively. The displacement along the orientation $\overrightarrow{o}$ is searched in a range of magnitude from 0 to 10 pixels. This corresponds to a maximum displacement of about 111 cm/day. Patches of size 3×3 were chosen, i.e. about 4 m and 6 m in ground geometry. A binary mask was provided to localize the glacier surface. Only corresponding pixels which are both on the glacier surface are

[4] we use anisotropic TV corresponding to the sum of the ℓ^1 norm of the gradient of $\boldsymbol{d}$ so that minimization problem (24) can be solved by graph-cuts



(a) Noisy image (b) Gaussian kernel $\mathcal{G}$ (c) GLR $\mathcal{L}_G$

Fig. 4 Results of motion tracking on a pair of SAR images of the glacier of Argentière. (a) One of the noisy input images and (b-c) the estimated magnitudes of the vector field. The method is based on energy minimization using either (b) the Gaussian kernel $\mathcal{G}$ or (c) the generalized likelihood ratio $\mathcal{L}_G$. The estimated speeds have an average over the surface of 12.27 cm/day and a maximum of 41.12 cm/day in the breaking slope (called "serac falls") for the estimation with GLR compared to 20.7 cm/day with a maximum of 67.2 cm/day for the Gaussian kernel $\mathcal{G}$.

used in patch comparisons. At each position is represented the magnitude of the local displacement estimated by both similarity criteria. According to experts and GPS measurements, the estimated velocities obtained with the generalized likelihood ratio $\mathcal{L}_G$ better reflects the ground truth with an average over the surface of 15.4 cm/day and a maximum of 53.8 cm/day in the breaking slope (called "serac falls") compared to 20.7 cm/day with a maximum of 67.2 cm/day for the Gaussian kernel $\mathcal{G}$. The use of $\mathcal{G}$ leads to a vector field over-regularized in dark areas and under-regularized in bright areas: there is no regularization parameter λ offering the same amount of smoothing everywhere in the vector field. Once again, since GLR is CFAR, we get the same amount of regularization of the field map both in dark and bright areas for a global regularization parameter λ.

Finally let us mention that neither criterion is optimal for this task due to illumination variations between the two observations. Correlation-based criteria could then be more adapted for such a task or a generalized likelihood ratio testing that $\boldsymbol{\theta}_1$ is within an affine transform of $\boldsymbol{\theta}_2$, i.e.: $\mathscr{H}_0 : \boldsymbol{\theta}_1 = \alpha\boldsymbol{\theta}_2 + \beta$, where α and β are real unknown values considered as nuisance parameters. Such an extension of the definition of similarity could be the topic of future work.

5 Conclusion

We have presented and compared seven similarity criteria taken from different research fields. Their theoretical grounding has been discussed as well as the different properties that they fulfil. In particular, it has been shown that some criteria are not invariant to the choice of the data space, and should thus be discarded. Others require signal-adaptive thresholds which restricts their usability in image processing applications. It has then been shown on patches extracted from a natural image that, under high levels of gamma or Poisson noise, the similarity criterion based on generalized likelihood ratio (GLR) is the most powerful. It also led to the best denoising performance when used as the criterion for patch similarity in NL-Means filtering, as assessed on a denoising benchmark made of twelve standard images synthetically damaged with strong gamma or Poisson noise. While GLR clearly outperforms techniques based on variance stabilization (such as the homomorphic approach or Anscombe transform) for low SNR images, our experiments show that variance stabilization is preferable for better SNR. With high SNR, patch

comparison probably requires further modeling of noiseless patch distances. In the absence of such a model, the Euclidean distance used after variance stabilization is probably the best choice.

We have illustrated the improvements brought by a suitable similarity criterion to denoise real-world images: a synthetic aperture radar image corrupted by multiplicative speckle noise, and an X-ray image of a supernova explosion with Poisson noise. With a similarity criterion adapted to the noise distribution, noise is smoothed out equally well in dark and bright regions. We then illustrated the wide applicability of the proposed similarity criterion in vision by considering a stereo-vision reconstruction problem and the estimation of displacement of a glacier with remote sensing.

Based on this study, we recommend a broader use of GLR for measuring patch similarity in computer vision. This criterion is both easy to implement and theoretically well grounded. With its very general definition based on hypothesis testing, this criterion is flexible and can easily be adapted to other problems of matching image parts. Two extensions could be derived in future work. Similarity criteria invariant to some transforms of the noise-free patch (e.g., change of illumination) could be derived, which would increase robustness in application such as motion tracking, stereo vision or flickering reduction. The modeling of a metric in the space of noise-free patches could also improve denoising performance, as suggested by our experiments with high SNR.

Acknowledgements

The authors would like to thank Julie Delon, Vincent Duval, Joseph Salmon and the anonymous reviewer for their help, comments and criticisms as well as the Centre National d'Études Spatiales, the Office Nationale d'Études et de Recherches Aérospatiales and the Délégation Générale pour l'Armement for providing the RAMSES data, the German Aerospace Center (DLR) and the french Agence Nationale de Recherche for providing the TerraSAR-X data in the framework of the project EFIDIR, and the Chandra X-ray Observatory for making publicly available the X-ray data.

Loïc Denis has been supported by project MiTiV funded by the French National Research Agency (ANR DEFI 09-EMER-008-01).

A Discussion about prior dependency and invariance in the case of multiplicative noise

Consider a multiplicative noise described by a variable x following a gamma distribution:

$$p(x|\theta) = \frac{L^L x^{L-1} e^{-\frac{Lx}{\theta}}}{\Gamma(L)\theta^L}. \tag{25}$$

This noise distribution corresponds to speckle noise as encountered in radar imaging. We have introduced in Sec. 4.1 a Bayesian joint likelihood criterion with an (improper) Jeffrey's prior $p(\theta = t) = \sqrt{|\mathcal{I}(\theta)|} = \frac{\sqrt{L}}{t}$ leading to the following criterion:

$$\mathcal{Q}_B^1(x_1, x_2) = \int p(x_1|\theta_1 = t)p(x_2|\theta_2 = t)\frac{\sqrt{L}}{t}\,\mathrm{d}t \tag{26}$$

$$\equiv \frac{1}{x_1 x_2}\left(\frac{x_1 x_2}{(x_1 + x_2)^2}\right)^L \tag{27}$$

where $\equiv$ means equivalence in terms of detection performances. We have seen that this criterion clearly does not fulfil the desirable properties and provides poor detection performance. However, we could have consider instead an (improper) uniform prior $p(\theta = t) = 1$ leading to:

$$\mathcal{Q}_B^2(x_1, x_2) = \int p(x_1|\theta_1 = t)p(x_2|\theta_2 = t)\,\mathrm{d}t \tag{28}$$

$$\equiv \frac{1}{\sqrt{x_1 x_2}}\left(\frac{x_1 x_2}{(x_1 + x_2)^2}\right)^{2L-1} \tag{29}$$

which has also poor performance, or a linear prior of the form $p(\theta = t) = t$ leading to:

$$\mathcal{Q}_B^3(x_1, x_2) = \int p(x_1|\theta_1 = t)p(x_2|\theta_2 = t)\; t\,\mathrm{d}t \tag{30}$$

$$\equiv \frac{x_1 x_2}{(x_1 + x_2)^2} \tag{31}$$

which is in this case equivalent to GLR. If Bayesian joint criteria generally provide poor performances, they sometimes lead to powerful criteria when the prior is well chosen.

Consider now the case of a variable x' following a Nakagami-Rayleigh distribution defined by:

$$p(x'|\theta) = \frac{2L^L x'^{2L-1} e^{-\frac{Lx'^2}{\theta}}}{\Gamma(L)\theta^L} \tag{32}$$

and define the Bayesian joint likelihood criterion with an (improper) uniform prior:

$$\mathcal{Q}_B^4(x'_1, x'_2) = \int p(x'_1|\theta_1 = t)p(x'_2|\theta_2 = t)\,\mathrm{d}t \tag{33}$$

$$\equiv \frac{x'_1 x'_2}{({x'_1}^2 + {x'_2}^2)}. \tag{34}$$

Consider also the case of a variable x'' following a Fisher-Tippet distribution defined by:

$$p(x''|\theta'') = \frac{L^L e^{L(x''-\theta'')-Le^{(x''-\theta'')}}}{\Gamma(L)} \tag{35}$$

and define the Bayesian joint likelihood criterion with an (improper) uniform prior:

$$\mathcal{Q}_B^5(x''_1, x''_2) = \int p(x''_1|\theta''_1 = t)p(x''_2|\theta''_2 = t)\,\mathrm{d}t \tag{36}$$

$$\equiv \frac{e^{\frac{x''_1+x''_2}{2}}}{(e^{x''_1} + e^{x''_2})} \tag{37}$$

where $\equiv$ means equivalence in terms of detection performances. Now observe that the Nakagami-Rayleigh and the Fisher-distribution corresponds respectively to the distribution of $\sqrt{x}|\theta$ and $\log x|\log\theta$ when $x|\theta$ has a gamma distribution. Using the change of variable $x' = \sqrt{x}$ and $x'' = \log x$ and $\theta'' = \log\theta$, we get that $\mathcal{Q}_B^4$ and $\mathcal{Q}_B^5$ are also criteria for gamma random variables:

$$\mathcal{Q}_B^4(x_1, x_2) = \int p(\sqrt{x}_1|\theta_1 = t)p(\sqrt{x}_2|\theta_2 = t)\,\mathrm{d}t \tag{38}$$

$$= 4\sqrt{x_1 x_2}\int p(x_1|\theta_1 = t)p(x_2|\theta_2 = t)\,\mathrm{d}t \tag{39}$$

$$= 4\sqrt{x_1 x_2}\mathcal{Q}_B^2(x_1, x_2) \tag{40}$$

$$\equiv \frac{x_1 x_2}{(x_1 + x_2)^2} \tag{41}$$

$$\mathcal{Q}_B^5(x_1, x_2) = \int p(\log x_1|\log\theta_1 = t)p(\log x_2|\log\theta_2 = t)\,\mathrm{d}t \tag{42}$$

$$= x_1 x_2 \int p(x_1|\theta_1 = t)p(x_2|\theta_2 = t)\frac{1}{t}\,\mathrm{d}t \tag{43}$$

$$= x_1 x_2 \mathcal{Q}_B^1(x_1, x_2) \tag{44}$$

$$\equiv \frac{x_1 x_2}{(x_1 + x_2)^2}. \tag{45}$$

which are both equivalent to GLR. If joint criteria generally provide poor performances, they can be redefined with suitable changes of variables to fulfill some desirable properties. For instance, if $\mathscr{C}$ is a joint likelihood criterion, we have:

$$\mathscr{C}_{\boldsymbol{g}(\boldsymbol{X}_1),\boldsymbol{g}(\boldsymbol{X}_2)}(\boldsymbol{g}(\boldsymbol{x}_1), \boldsymbol{g}(\boldsymbol{x}_2)) = \left|\frac{\mathrm{d}\boldsymbol{g}(\boldsymbol{x}_1)}{\mathrm{d}\boldsymbol{x}_1}\right|^{-1}\left|\frac{\mathrm{d}\boldsymbol{g}(\boldsymbol{x}_2)}{\mathrm{d}\boldsymbol{x}_2}\right|^{-1}\mathscr{C}_{\boldsymbol{X}_1,\boldsymbol{X}_2}(\boldsymbol{x}_1, \boldsymbol{x}_2)\,. \tag{46}$$

and one can search for a mapping $\boldsymbol{g}$ such that the Jacobian terms normalize the resulting criteria in the same way as the denominator of likelihood-ratio based criteria. This was the constructive argument we used in (Deledalle et al., 2011a) to obtain a suitable Bayesian joint likelihood criterion for a variant of circular complex Gaussian distributions.

Finally, note that by definition (33), $\mathcal{Q}_B^4$ is the criterion introduced in (Deledalle et al., 2009b) and by the relation (42), $\mathcal{Q}_B^5$ is the criterion introduced in (Teuber and Lang, 2011) and they both have the same performance which is as good as GLR. Note that Teuber and Lang (2011) compare their approach, i.e. $\mathcal{Q}_B^5$, with ours introduced in (Deledalle et al., 2009b), but they consider $\mathcal{Q}_B^2$ instead of $\mathcal{Q}_B^4$. We believe that the confusion comes from the similitude between eq. (28) and eq. (33).

Due to its non-invariance and prior-dependency, the definition of Bayesian joint likelihood criteria is ambiguous, and can lead arbitrarily to poor or good performances (see Table 5). Let us highlight again that GLR is invariant and priorless, which makes its definition non-ambiguous.

Criteria	mapping $\boldsymbol{g}$	prior $p(\theta = t)$	Properties
$\mathcal{Q}_B^1$	$x \mapsto x$	Jeffrey's prior	None
$\mathcal{Q}_B^2$	$x \mapsto x$	1	None
$\mathcal{Q}_B^3$	$x \mapsto x$	t	$\equiv$ GLR
$\mathcal{Q}_B^4$	$x \mapsto \sqrt{x}$	1	$\equiv$ GLR
$\mathcal{Q}_B^5$	$x \mapsto \log x$	$\frac{1}{t}$	$\equiv$ GLR

Table 5 Five different Bayesian likelihood criteria in the case of gamma distributions . Depending of the choice of the mapping $\boldsymbol{g}$ and the prior $p(\theta = t)$, the resulting criteria can fulfill none of the desirable properties or be equivalent to GLR in terms of detection performance.

B Derivation of closed-form expressions of similarity criteria

We derive in this section the closed-form expression of the 7 different similarity criteria between patches $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$:

- $\mathcal{G}$, the usual similarity criterion based on squared differences: $\mathcal{G}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\left(-\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2/h\right)$,
- $\mathcal{S}$, based on variance stabilizing transform $\boldsymbol{s}$: $\mathcal{S}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathcal{G}(\boldsymbol{s}(\boldsymbol{x}_1), \boldsymbol{s}(\boldsymbol{x}_2))$,
- $\mathcal{L}_B$, the Bayesian likelihood ratio: $\mathcal{L}_B(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{\int p(\boldsymbol{x}_1|\boldsymbol{\theta}_{12}=\boldsymbol{t}) p(\boldsymbol{x}_2|\boldsymbol{\theta}_{12}=\boldsymbol{t}) p(\boldsymbol{\theta}_{12}=\boldsymbol{t})\, \mathrm{d}\boldsymbol{t}}{\int p(\boldsymbol{x}_1|\boldsymbol{\theta}_1=\boldsymbol{t}_1) p(\boldsymbol{\theta}_1=\boldsymbol{t}_1) \mathrm{d}\boldsymbol{t}_1 \int p(\boldsymbol{x}_2|\boldsymbol{\theta}_2=\boldsymbol{t}_2) p(\boldsymbol{\theta}_2=\boldsymbol{t}_2) \mathrm{d}\boldsymbol{t}_2}$,
- $\mathcal{L}_G$, the generalized likelihood ratio: $\mathcal{L}_G(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{\sup_{\boldsymbol{t}} p(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{\theta}_{12}=\boldsymbol{t}, \mathscr{H}_0)}{\sup_{\boldsymbol{t}_1, \boldsymbol{t}_2} p(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{\theta}_1=\boldsymbol{t}_1, \boldsymbol{\theta}_2=\boldsymbol{t}_2, \mathscr{H}_1)}$,
- $\mathcal{Q}_B$, the Bayesian joint likelihood: $\mathcal{Q}_B(\boldsymbol{x}_1, \boldsymbol{x}_2) = \int p(\boldsymbol{x}_1|\boldsymbol{\theta}_1 = \boldsymbol{t})\, p(\boldsymbol{x}_2|\boldsymbol{\theta}_2 = \boldsymbol{t})\, p(\boldsymbol{\theta}_{12} = \boldsymbol{t})\, \mathrm{d}\boldsymbol{t}$,
- $\mathcal{Q}_G$, the maximum joint likelihood: $\mathcal{Q}_G(\boldsymbol{x}_1, \boldsymbol{x}_2) = p(\boldsymbol{x}_1; \boldsymbol{\theta}_1 = \hat{\boldsymbol{t}}_{12})\, p(\boldsymbol{x}_2; \boldsymbol{\theta}_2 = \hat{\boldsymbol{t}}_{12})$,
- $\mathcal{K}_B$, the mutual information kernel: $\mathcal{K}_B(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathcal{Q}_B(\boldsymbol{x}_1, \boldsymbol{x}_2)/\sqrt{\mathcal{Q}_B(\boldsymbol{x}_1, \boldsymbol{x}_1)\mathcal{Q}_B(\boldsymbol{x}_2, \boldsymbol{x}_2)}$.

where, $\boldsymbol{x}$ denotes the available (i.e., noisy) data, while $\boldsymbol{\theta}$ are the parameters of interest that are to be recovered.

We consider uncorrelated noise, so that patch similarity is the product over the patch of similarity between pixels. We study first Gaussian noise, then Gamma noise, Poisson noise, and finally Cauchy-distributed noise.

B.1 Gaussian noise case

Given $\sigma \in \mathbb{R}_*^+$, a Gaussian random variable X follows the probability density function (pdf):

$$p(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\theta)^2}{2\sigma^2}\right], \tag{47}$$

with expectation $\mathbb{E}[X] = \theta$ and variance $\mathrm{Var}[X] = \sigma^2$. Gaussian fluctuations are additive, it is straightforward to show that X can be decomposed as $\theta + N$ with N a zero mean Gaussian random variable.

B.1.1 Fisher information

Fisher information associated with a Gaussian pdf is given by:

$$\mathcal{I}(\theta) \triangleq \mathbb{E}_X\left[\left(\frac{\partial}{\partial\theta}\log p(x|\theta)\right)^2\right] = \int \left(\frac{\partial}{\partial\theta}\log p(x|\theta)\right)^2 p(x|\theta)\, \mathrm{d}x \tag{48}$$

$$= \int \left(\frac{x-\theta}{\sigma^2}\right)^2 \frac{e^{-\frac{(x-\theta)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}\, \mathrm{d}x = \underbrace{\frac{1}{\sigma^4}\int (x-\theta)^2 \frac{e^{-\frac{(x-\theta)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}\, \mathrm{d}x}_{\text{by definition of variance}} = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}. \tag{49}$$

B.1.2 Jeffreys' prior

Jeffreys' prior follows from Fisher information:

$$p(\theta) \triangleq \sqrt{|\mathcal{I}(\theta)|} = \frac{1}{\sigma}. \tag{50}$$

B.1.3 Bayesian joint likelihood

With Jeffreys' prior, we can derive the Bayesian joint likelihood as follows:

$$\mathcal{Q}_B(x_1,x_2) = \int p(x_1|\theta_1=t)p(x_2|\theta_2=t)p(\theta_{12}=t)\mathrm{d}t = \int \left(\frac{e^{-\frac{(x_1-t)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}\right)\left(\frac{e^{-\frac{(x_2-t)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}\right)\left(\frac{1}{\sigma}\right)\mathrm{d}t \tag{51}$$

$$= \underbrace{\frac{1}{2\pi\sigma^3}\int e^{-\frac{(x_1-t)^2}{2\sigma^2}} e^{-\frac{(x_2-t)^2}{2\sigma^2}}\,\mathrm{d}t = \frac{e^{-\frac{(x_1-x_2)^2}{4\sigma^2}}}{2\pi\sigma^3}}_{\text{by convolution of two Gaussian functions}} . \tag{52}$$

B.1.4 Bayesian likelihood ratio

Let $\mathcal{D}_B$ be the denominator term appearing in the Bayesian likelihood ratio and expressed as:

$$\mathcal{D}_B(x) = \int p(x|\theta=t)p(\theta=t)\mathrm{d}t = \int \left(\frac{e^{-\frac{(x-t)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}\right)\left(\frac{1}{\sigma}\right)\mathrm{d}t = \frac{1}{\sigma} . \tag{53}$$

Using the expression of $\mathcal{Q}_B(x_1,x_2)$ and $\mathcal{D}_B(x)$, it results that the Bayesian likelihood ratio is given by:

$$\mathcal{L}_B(x_1,x_2) = \frac{\mathcal{Q}_B(x_1,x_2)}{\mathcal{D}_B(x_1)\mathcal{D}_B(x_2)} = \frac{\frac{e^{-\frac{(x_1-x_2)^2}{4\sigma^2}}}{2\pi\sigma^3}}{\frac{1}{\sigma}\frac{1}{\sigma}} = \frac{e^{-\frac{(x_1-x_2)^2}{4\sigma^2}}}{2\pi\sigma} . \tag{54}$$

B.1.5 Mutual information kernel

Using the expression of $\mathcal{Q}_B(x_1,x_2)$ and $\mathcal{Q}_B(x,x)$, it comes that the mutual information kernel is:

$$\mathcal{K}_B(x_1,x_2) = \frac{\mathcal{Q}_B(x_1,x_2)}{\sqrt{\mathcal{Q}_B(x_1,x_1)\mathcal{Q}_B(x_2,x_2)}} = \frac{\frac{e^{-\frac{(x_1-x_2)^2}{4\sigma^2}}}{2\pi\sigma^3}}{\sqrt{\frac{e^0}{2\pi\sigma^3}\frac{e^0}{2\pi\sigma^3}}} = e^{-\frac{(x_1-x_2)^2}{4\sigma^2}} . \tag{55}$$

B.1.6 Maximum joint likelihood

The priorless extension of $\mathcal{Q}_B(x_1,x_2)$, i.e. the maximum joint likelihood is obtained as follows:

$$\mathcal{Q}_G(x_1,x_2) = \sup_t p(x_1|\theta_1=t)p(x_2|\theta_2=t) = \left(\frac{e^{-\frac{(x_1-\frac{x_1+x_2}{2})^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}\right)\left(\frac{e^{-\frac{(x_2-\frac{x_1+x_2}{2})^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}\right) \tag{56}$$

$$= \frac{1}{2\pi\sigma^2} e^{-\frac{(x_1-x_2)^2}{8\sigma^2}} e^{-\frac{(x_-x_2)^2}{8\sigma^2}} = \frac{e^{-\frac{(x_1-x_2)^2}{4\sigma^2}}}{2\pi\sigma^2} . \tag{57}$$

since under Gaussian noise the maximum likelihood estimator (MLE) is the mean.

B.1.7 Generalized likelihood ratio

Let $\mathcal{D}_G$ be the denominator term appearing in the generalized likelihood ratio and expressed as:

$$\mathcal{D}_G(x) = \sup_t p(x|\theta=t) = \frac{e^0}{\sqrt{2\pi}\sigma} = \frac{1}{\sqrt{2\pi}\sigma} . \tag{58}$$

Using the expression of $\mathcal{Q}_G(x_1,x_2)$ and $\mathcal{D}_G(x)$, it results that the generalized likelihood ratio is given by:

$$\mathcal{L}_G(x_1,x_2) = \frac{\mathcal{Q}_G(x_1,x_2)}{\mathcal{D}_G(x_1)\mathcal{D}_G(x_2)} = \frac{\frac{e^{-\frac{(x_1-x_2)^2}{4\sigma^2}}}{2\pi\sigma^2}}{\frac{1}{\sqrt{2\pi}\sigma}\frac{1}{\sqrt{2\pi}\sigma}} = e^{-\frac{(x_1-x_2)^2}{4\sigma^2}} . \tag{59}$$

B.2 Gamma noise case

Given the positive integer $L \in \mathbb{N}^*$, a Gamma random variable X can be described by the following pdf:

$$p(x|\theta) = \frac{L^L x^{L-1} e^{-\frac{Lx}{\theta}}}{\Gamma(L)\theta^L}. \tag{60}$$

Its expectation is $\mathbb{E}[X] = \theta$ and variance $\mathrm{Var}[X] = \frac{\theta^2}{L}$. The relation $\mathrm{Var}[X] \propto \mathbb{E}[X]^2$ indicates a multiplicative behaviour. Indeed, it is straightforward to show that X can be decomposed as $\theta \times S$ with S a Gamma random variable of parameter $\theta_S = 1$.

B.2.1 Fisher information

Fisher information associated with a Gamma pdf is given by:

$$\mathcal{I}(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\log p(x|\theta)\right)^2 |\theta\right] = \int \left(\frac{Lx}{\theta^2} - \frac{L}{\theta}\right)^2 \frac{L^L x^{L-1} e^{-\frac{Lx}{\theta}}}{\Gamma(L)\theta^L}\mathrm{d}x \tag{61}$$

$$= \underbrace{\frac{L^2}{\theta^4}\int (x-\theta)^2 \frac{L^L x^{L-1} e^{-\frac{Lx}{\theta}}}{\Gamma(L)\theta^L}\mathrm{d}x}_{\text{by definition of variance}} = \frac{L^2}{\theta^4}\frac{\theta^2}{L} = \frac{L}{\theta^2}\,. \tag{62}$$

B.2.2 Jeffreys' prior

Fisher information allows to define Jeffreys' prior as:

$$p(\theta) \triangleq \sqrt{|\mathcal{I}(\theta)|} = \frac{\sqrt{L}}{\theta} \tag{63}$$

B.2.3 Bayesian joint likelihood

With Jeffreys prior, we can derive the Bayesian joint likelihood as follows:

$$\mathcal{Q}_B(x_1,x_2) = \int p(x_1|t_1=t)p(x_2|t_2=t)p(t_{12}=t)\mathrm{d}t = \int \left(\frac{L^L x_1^{L-1} e^{-\frac{Lx_1}{t}}}{\Gamma(L)t^L}\right)\left(\frac{L^L x_2^{L-1} e^{-\frac{Lx_2}{t}}}{\Gamma(L)t^L}\right)\left(\frac{\sqrt{L}}{t}\right)\mathrm{d}t \tag{64}$$

$$= \frac{L^{2L+1/2} x_1^{L-1} x_2^{L-1}}{\Gamma(L)^2}\int \frac{e^{-\frac{L(x_1+x_2)}{t}}}{t^{2L+1}}\mathrm{d}t = \frac{L^{2L+1/2} x_1^{L-1} x_2^{L-1}}{\Gamma(L)^2}\frac{\Gamma(2L)}{(L(x_1+x_2))^{2L}} \tag{65}$$

$$= \frac{\sqrt{L}\Gamma(2L)}{\Gamma(L)^2}\left(\frac{1}{x_1x_2}\left(\frac{x_1x_2}{(x_1+x_2)^2}\right)^L\right) \tag{66}$$

by using

$$\int \frac{e^{-\frac{A}{t}}}{t^N}\mathrm{d}t = \frac{\Gamma(N-1)}{A^{N-1}}\,. \tag{67}$$

B.2.4 Bayesian likelihood ratio

Let $\mathcal{D}_B$ be the denominator term appearing in the Bayesian likelihood ratio and expressed as:

$$\mathcal{D}_B(x) = \int p(x|\theta=t)p(\theta=t)\mathrm{d}t = \int \left(\frac{L^L x^{L-1} e^{-\frac{Lx}{t}}}{\Gamma(L)t^L}\right)\left(\frac{\sqrt{L}}{t}\right)\mathrm{d}t \tag{68}$$

$$= \frac{L^{L+1/2} x^{L-1}}{\Gamma(L)}\int \frac{e^{-\frac{Lx}{t}}}{t^{L+1}}\mathrm{d}t = \frac{L^{L+1/2} x^{L-1}}{\Gamma(L)}\frac{\Gamma(L)}{(Lx)^L} = \frac{\sqrt{L}}{x}\,. \tag{69}$$

Using the expression of $\mathcal{Q}_B(x_1,x_2)$ and $\mathcal{D}_B(x)$, it comes that the Bayesian likelihood ratio is given by:

$$\mathcal{L}_B = \frac{\mathcal{Q}_B(x_1,x_2)}{\mathcal{D}_B(x_1)\mathcal{D}_B(x_2)} = \frac{\frac{\sqrt{L}\Gamma(2L)}{\Gamma(L)^2}\frac{x_1^{L-1}x_2^{L-1}}{(x_1+x_2)^{2L}}}{\frac{\sqrt{L}}{x}\frac{\sqrt{L}}{x}} = \frac{\Gamma(2L)}{\sqrt{L}\Gamma(L)^2}\left(\frac{x_1x_2}{(x_1+x_2)^2}\right)^L\,. \tag{70}$$

B.2.5 Mutual information kernel

Using the expression of $\mathcal{Q}_B(x_1, x_2)$ and $\mathcal{Q}_B(x, x)$, it results that the mutual information kernel is given by:

$$\mathcal{K}_B(x_1, x_2) = \frac{\mathcal{Q}_B(x_1, x_2)}{\sqrt{\mathcal{Q}_B(x_1, x_1)\mathcal{Q}_B(x_2, x_2)}} = \frac{\frac{\sqrt{L}\Gamma(2L)}{\Gamma(L)^2}\frac{x_1^{L-1}x_2^{L-1}}{(x_1+x_2)^{2L}}}{\sqrt{\frac{\sqrt{L}\Gamma(2L)}{\Gamma(L)^2}\frac{x_1^{2L-2}}{(2x_1)^{2L}}\frac{\sqrt{L}\Gamma(2L)}{\Gamma(L)^2}\frac{x_2^{2L-2}}{(2x_2)^{2L}}}} = 2^{2L}\left(\frac{x_1x_2}{(x_1+x_2)^2}\right)^L . \tag{71}$$

B.2.6 Maximal joint likelihood

The priorless extension of $\mathcal{Q}_B(x_1, x_2)$, i.e. the maximum joint likelihood is obtained as follows:

$$\mathcal{Q}_G(x_1, x_2) = \int \sup_t p(x_1|t_1 = t)p(x_2|t_2 = t)\mathrm{d}t = \left(\frac{2^L L^L x_1^{L-1} e^{-\frac{2Lx_1}{x_1+x_2}}}{\Gamma(L)(x_1+x_2)^L}\right)\left(\frac{2^L L^L x_2^{L-1} e^{-\frac{2Lx_2}{x_1+x_2}}}{\Gamma(L)(x_1+x_2)^L}\right) \tag{72}$$

$$= \frac{2^{2L}L^{2L}x_1^{L-1}x_2^{L-1}e^{-2L}}{\Gamma(L)^2(x_1+x_2)^{2L}} = \frac{2^{2L}L^{2L}e^{-2L}}{\Gamma(L)^2}\left(\frac{1}{x_1x_2}\left(\frac{x_1x_2}{(x_1+x_2)^2}\right)^L\right) . \tag{73}$$

since under Gamma noise the MLE is the mean.

B.2.7 Generalized likelihood ratio

Let $\mathcal{D}_G$ be the denominator term appearing in the generalized likelihood ratio and expressed as:

$$\mathcal{D}_G(x) = \sup_t p(x|\theta = t) = \frac{L^L e^{-L}}{\Gamma(L)x} . \tag{74}$$

Using the expression of $\mathcal{Q}_G(x_1, x_2)$ and $\mathcal{D}_G(x)$, it results that the generalized likelihood ratio is given by:

$$\mathcal{L}_G = \frac{\mathcal{Q}_G(x_1, x_2)}{\mathcal{D}_G(x_1)\mathcal{D}_G(x_2)} = \frac{\frac{2^{2L}L^{2L}e^{-2L}}{\Gamma(L)}\frac{x_1^{L-1}x_2^{L-1}}{(x_1+x_2)^{2L}}}{\frac{L^Le^{-L}}{\Gamma(L)x}\frac{L^Le^{-L}}{\Gamma(L)x}} = 2^{2L}\left(\frac{x_1x_2}{(x_1+x_2)^2}\right)^L . \tag{75}$$

B.2.8 Variance stabilization criterion

Variance stabilization of Gamma random values can be performed using a log transform:

$$\boldsymbol{s}(X) = \log X \Rightarrow \mathrm{Var}[\boldsymbol{s}(X)] = \mathrm{Var}[\log X] = \Psi(1, L) \tag{76}$$

where $\Psi(1, L)$ is the first-order Polygamma function of L (e.g. Xie et al., 2002). The resulting similarity criterion is then given by:

$$\mathcal{S}(x_1, x_2) = \exp\left[-\left(\log\frac{x_1}{x_2}\right)^2\right] . \tag{77}$$

B.3 Poisson noise case

A Poisson random variable X can be described by the following pdf:

$$p(x|\theta) = \frac{\theta^x e^{-\theta}}{x!} . \tag{78}$$

Its expectation is $\mathbb{E}[x] = \theta$ and variance $\mathrm{Var}[X] = \theta$. Note that the relation $\mathrm{Var}[X] = \mathbb{E}[x]$ is non-homogeneous, which is challenging, since, as a consequence, X cannot be related to θ through additive or multiplicative decomposition.

B.3.1 Fisher information

Fisher information associated with a Poissonian pdf is given by:

$$\mathcal{I}(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\log p(x|\theta)\right)^2|\theta\right] = \int\left(\frac{\partial}{\partial\theta}\log p(x|\theta)\right)^2 p(x|\theta)\mathrm{d}x \tag{79}$$

$$= \int\left(\frac{x}{\theta}-1\right)^2\frac{\theta^x e^{-\theta}}{x!}\mathrm{d}x = \underbrace{\frac{1}{\theta^2}\int(x-\theta)^2\frac{\theta^x e^{-\theta}}{x!}\mathrm{d}x = \frac{\theta}{\theta^2}}_{\text{by definition of variance}} = \frac{1}{\theta} . \tag{80}$$

B.3.2 Jeffreys' prior

The corresponding Jeffreys' prior is:

$$p(\theta) \triangleq \sqrt{|\mathcal{I}(\theta)|} = \frac{1}{\sqrt{\theta}} \,. \tag{81}$$

B.3.3 Bayesian joint likelihood

With Jeffreys' prior, we can derive the Bayesian joint likelihood as follow:

$$\mathcal{Q}_B(x_1, x_2) = \int p(x_1|\theta_1 = t)p(x_2|\theta_2 = t)p(\theta_{12} = t)\mathrm{d}t = \int \left(\frac{t^{x_1}e^{-t}}{x_1!}\right)\left(\frac{t^{x_2}e^{-t}}{x_2!}\right)\left(\frac{1}{\sqrt{t}}\right)\mathrm{d}t \tag{82}$$

$$= \frac{1}{x_1!x_2!}\int t^{x_1+x_2-1/2}e^{-2t}\mathrm{d}t = \frac{1}{\sqrt{2}}\frac{\Gamma(x1+x2+1/2)}{2^{x1+x2}x_1!x_2!} \tag{83}$$

by using

$$\int t^N e^{-At}\mathrm{d}t = \int \frac{e^{-\frac{A}{t}}}{t^{N+2}}\mathrm{d}t = \frac{\Gamma(N+1)}{A^{N+1}} \,. \tag{84}$$

B.3.4 Bayesian likelihood ratio

Let $\mathcal{D}_B$ be the denominator term appearing in the Bayesian likelihood ratio and expressed as:

$$\mathcal{D}_B(x) = \int p(x|\theta = t)p(\theta = t)\mathrm{d}t = \int \left(\frac{t^x e^{-t}}{x!}\right)\left(\frac{1}{\sqrt{t}}\right)\mathrm{d}t = \frac{1}{x!}\int t^{x-1/2}e^{-t}\mathrm{d}t \tag{85}$$

$$= \frac{\Gamma(x+1/2)}{x!} \,. \tag{86}$$

Using the expression of $\mathcal{Q}_B(x_1, x_2)$ and $\mathcal{D}_B(x)$, it results that the Bayesian likelihood ratio is given by:

$$\mathcal{L}_B = \frac{\mathcal{Q}_B(x_1, x_2)}{\mathcal{D}_B(x_1)\mathcal{D}_B(x_2)} = \frac{\frac{1}{\sqrt{2}}\frac{\Gamma(x_1+x_2+1/2)}{2^{x1+x2}x_1!x_2!}}{\frac{\Gamma(x_1+1/2)}{x_1!}\frac{\Gamma(x_2+1/2)}{x_2!}} = \frac{1}{\sqrt{2}}\frac{\Gamma(x_1+x_2+1/2)}{2^{x_1+x_2}\Gamma(x_1+1/2)\Gamma(x_2+1/2)} \,. \tag{87}$$

B.3.5 Mutual information kernel

Using the expression of $\mathcal{Q}_B(x_1, x_2)$ and $\mathcal{Q}_B(x, x)$, the mutual information kernel can be written as:

$$\mathcal{K}_B(x_1, x_2) = \frac{\mathcal{Q}_B(x_1, x_2)}{\sqrt{\mathcal{Q}_B(x_1, x_1)\mathcal{Q}_B(x_2, x_2)}} = \frac{\frac{1}{\sqrt{2}}\frac{\Gamma(x1+x2+1/2)}{2^{x1+x2}x_1!x_2!}}{\sqrt{\frac{1}{\sqrt{2}}\frac{\Gamma(2x_1+1/2)}{2^{2x_1}x_1!^2}\frac{1}{\sqrt{2}}\frac{\Gamma(2x_2+1/2)}{2^{2x_2}x_2!^2}}} = \frac{\Gamma(x_1+x_2+1/2)}{\sqrt{\Gamma(2x_1+1/2)\Gamma(2x_2+1/2)}} \,. \tag{88}$$

B.3.6 Maximal joint likelihood

The priorless extension of $\mathcal{Q}_B(x_1, x_2)$, i.e. the maximum joint likelihood is obtained as follows:

$$\mathcal{Q}_G(x_1, x_2) = \int \sup_t p(x_1|t_1 = t)p(x_2|t_2 = t)\mathrm{d}t = \left(\frac{\frac{(x_1+x_2)^{x_1}}{2^{x_1}}e^{-\frac{x_1+x_2}{2}}}{x_1!}\right)\left(\frac{\frac{(x_1+x_2)^{x_2}}{2^{x_2}}e^{-\frac{x_1+x_2}{2}}}{x_2!}\right) \tag{89}$$

$$= \frac{(x_1+x_2)^{x_1+x_2}}{(2e)^{x_1+x_2}x_1!x_2!} \,. \tag{90}$$

since once again, the MLE for Poisson noise is the mean.

B.3.7 Generalized likelihood ratio

Let $\mathcal{D}_G$ be the denominator term appearing in the generalized likelihood ratio and expressed as:

$$\mathcal{D}_G(x) = \sup_t p(x|\theta = t) = \frac{x^x e^{-x}}{x!} \,. \tag{91}$$

Using the expression of $\mathcal{Q}_G(x_1, x_2)$ and $\mathcal{D}_G(x)$, it comes that the generalized likelihood ratio is:

$$\mathcal{L}_G = \frac{\mathcal{Q}_G(x_1, x_2)}{\mathcal{D}_G(x_1)\mathcal{D}_G(x_2)} = \frac{\frac{(x_1+x_2)^{x_1+x_2}}{(2e)^{x_1+x_2}x_1!x_2!}}{\frac{x_1^{x_1}e^{-x_1}}{x_1!}\frac{x_2^{x_2}e^{-x_2}}{x_2!}} = \frac{(x_1+x_2)^{x_1+x_2}}{2^{x_1+x_2}x_1^{x_1}x_2^{x_2}} \,. \tag{92}$$

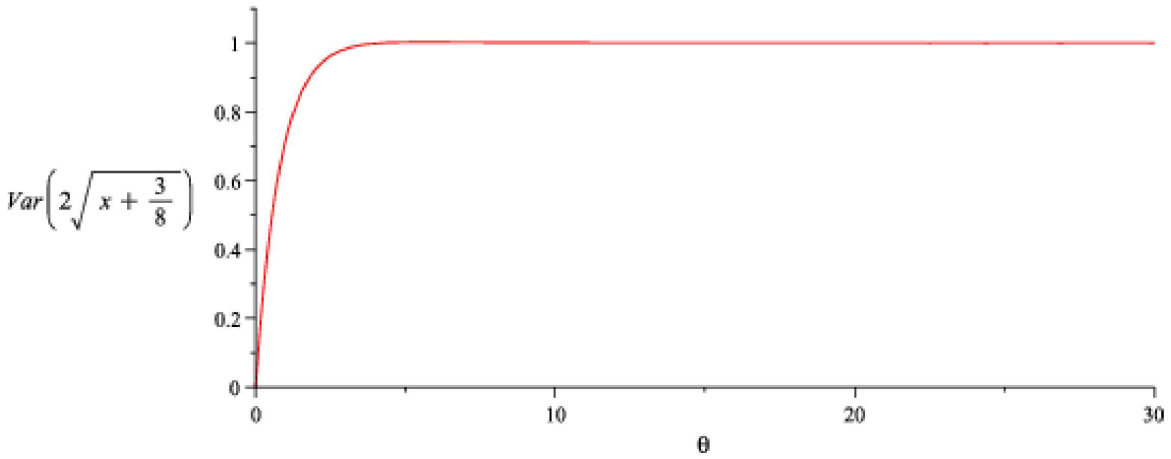


Fig. 5 Variance of the Anscombe transform of Poisson random variables wrt to the parameter θ. For θ sufficiently high, the variance is independent of θ and equal to 1.

B.3.8 Variance stabilization criterion

Approximated variance stabilization of Poisson random values can be performed using Anscombe transform:

$$s(X) = 2\sqrt{X+\frac{3}{8}} \Rightarrow (\theta \gg 0 \Rightarrow \mathrm{Var}[s(X)] = 1)\ . \tag{93}$$

Figure 5 describes the relation between θ and the variance of the Anscombe transform. The resulting similarity criterion is then given by:

$$\mathcal{S}(x_1, x_2) = \exp\left(-4\left(\sqrt{x_1+\frac{3}{8}} - \sqrt{x_2+\frac{3}{8}}\right)^2\right). \tag{94}$$

B.4 Cauchy noise case

A Cauchy random variable X can be described by the following pdf:

$$p(x|\theta) = \frac{1}{\pi\gamma\left[1+\left(\frac{x-\theta}{\gamma}\right)^2\right]}\ . \tag{95}$$

where θ is the mode and γ is a shape parameter. Cauchy fluctuations are additive, it is straightforward to show that X can be decomposed as $\theta + N$ with N a Cauchy random variable with a mode in 0 and the scale parameter γ. The particularity of Cauchy random variables is that their expectation and variance do not exist. A consequence is that the sample mean and the sample variance do not converge wrt the number of observations. Surprisingly, all criteria are still defined in this case, except the variance stabilization criterion since we have not found a transformation g such as $g(X)$ has a finite and constant variance whatever θ.

B.4.1 Fisher information

Fisher information associated with a Gaussian pdf is given by:

$$\mathcal{I}(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\log p(x|\theta)\right)^2 |\theta\right] = \int\left(\frac{\partial}{\partial\theta}\log p(x|\theta)\right)^2 p(x|\theta)\mathrm{d}x \tag{96}$$

$$= \int\left(\frac{2(x-\theta)}{\gamma^2\left[1+\left(\frac{x-\theta}{\gamma}\right)^2\right]}\right)^2 \frac{1}{\pi\gamma\left[1+\left(\frac{x-\theta}{\gamma}\right)^2\right]}\mathrm{d}x = \frac{1}{2\gamma^2}\ . \qquad \text{(derived with Maple)} \tag{97}$$

B.4.2 Jeffreys' prior

Fisher information gives Jeffreys' prior as:

$$p(\theta) \triangleq \sqrt{|\mathcal{I}(\theta)|} = \frac{1}{\sqrt{2}\gamma}\ . \tag{98}$$

B.4.3 Bayesian joint likelihood

With Jeffreys' prior, we can derive the Bayesian joint likelihood as follows:

$$\mathcal{Q}_B(x_1,x_2) = \int p(x_1|\theta_1=t)p(x_2|\theta_2=t)p(\theta_{12}=t)\mathrm{d}t \tag{99}$$

$$= \int \left(\frac{1}{\pi\gamma\left[1+\left(\frac{x_1-t}{\gamma}\right)^2\right]}\right)\left(\frac{1}{\pi\gamma\left[1+\left(\frac{x_2-t}{\gamma}\right)^2\right]}\right)\left(\frac{1}{\sqrt{2}\gamma}\right)\mathrm{d}t \tag{100}$$

$$= \frac{\sqrt{2}}{\pi\gamma^2\left[4+\left(\frac{x_1-x_2}{\gamma}\right)^2\right]}\,. \tag{101}$$

B.4.4 Bayesian likelihood ratio

Let $\mathcal{D}_B$ be the denominator term appearing in the Bayesian likelihood ratio and expressed as:

$$\mathcal{D}_B(x) = \int p(x|\theta=t)p(\theta=t)\mathrm{d}t = \int \left(\frac{1}{\pi\gamma\left[1+\left(\frac{x-t}{\gamma}\right)^2\right]}\right)\left(\frac{1}{\sqrt{2}\gamma}\right)\mathrm{d}t = \frac{1}{\sqrt{2}\gamma}\,. \tag{102}$$

Using the expression of $\mathcal{Q}_B(x_1,x_2)$ and $\mathcal{D}_B(x)$, it results that the Bayesian likelihood ratio is given by:

$$\mathcal{L}_B = \frac{\mathcal{Q}_B(x_1,x_2)}{\mathcal{D}_B(x_1)\mathcal{D}_B(x_2)} = \frac{\frac{\sqrt{2}}{\pi\gamma^2\left[4+\left(\frac{x_1-x_2}{\gamma}\right)\right]}}{\frac{1}{\sqrt{2}\gamma}\frac{1}{\sqrt{2}\gamma}} = \frac{2\sqrt{2}}{\pi\left[4+\left(\frac{x_1-x_2}{\gamma}\right)^2\right]}\,. \tag{103}$$

B.4.5 Mutual information kernel

Using the expression of $\mathcal{Q}_B(x_1,x_2)$ and $\mathcal{Q}_B(x,x)$, it results that the mutual information kernel is given by:

$$\mathcal{K}_B(x_1,x_2) = \frac{\mathcal{Q}_B(x_1,x_2)}{\sqrt{\mathcal{Q}_B(x_1,x_1)\mathcal{Q}_B(x_2,x_2)}} = \frac{\frac{\sqrt{2}}{\pi\gamma^2\left[4+\left(\frac{x_1-x_2}{\gamma}\right)^2\right]}}{\sqrt{\frac{\sqrt{2}}{\pi\gamma^2\left[4+\left(\frac{x_1-x_1}{\gamma}\right)^2\right]}\frac{\sqrt{2}}{\pi\gamma^2\left[4+\left(\frac{x_2-x_2}{\gamma}\right)^2\right]}}} = \frac{1}{1+\left(\frac{x_1-x_2}{2\gamma}\right)^2} \tag{104}$$

B.4.6 Maximal joint likelihood

The priorless extension of $\mathcal{Q}_B(x_1,x_2)$, i.e. the maximum joint likelihood is obtained as follows:

$$\mathcal{Q}_G(x_1,x_2) = \int \sup_t p(x_1|t_1=t)p(x_2|t_2=t)\mathrm{d}t = \frac{1}{\pi\gamma\left[1+\left(\frac{x_1-\frac{x_1+x_2}{2}}{\gamma}\right)^2\right]}\frac{1}{\pi\gamma\left[1+\left(\frac{x_1-\frac{x_1+x_2}{2}}{\gamma}\right)^2\right]} \tag{105}$$

$$= \frac{1}{\pi^2\gamma^2\left[1+\left(\frac{x_1-x_2}{2\gamma}\right)^2\right]^2} \tag{106}$$

$$\tag{107}$$

since for a dataset of one or two elements the mean is the MLE (note that its no more the case for larger datasets).

B.4.7 Generalized likelihood ratio

Let $\mathcal{D}_G$ be the denominator term appearing in the generalized likelihood ratio and expressed as:

$$\mathcal{D}_G(x) = \sup_t p(x|\theta=t) = \frac{1}{\pi\gamma\left[1+\left(\frac{x-x}{\gamma}\right)^2\right]} = \frac{1}{\pi\gamma}\,. \tag{108}$$

Using the expression of $\mathcal{Q}_G(x_1,x_2)$ and $\mathcal{D}_G(x)$, it results that the generalized likelihood ratio is given by:

$$\mathcal{L}_G = \frac{\mathcal{Q}_G(x_1,x_2)}{\mathcal{D}_G(x_1)\mathcal{D}_G(x_2)} = \frac{\frac{1}{\pi^2\gamma^2\left[1+\left(\frac{x_1-x_2}{2\gamma}\right)^2\right]^2}}{\frac{1}{\pi\gamma}\frac{1}{\pi\gamma}} = \frac{1}{\left[1+\left(\frac{x_1-x_2}{2\gamma}\right)^2\right]^2}\,. \tag{109}$$

B.4.8 Variance stabilization criterion

Cauchy random variables have no expectation nor variance. Our attempts to transform Cauchy r.v. into random variables with constant variance did not succeed.

C Proof sketches for similarity criteria properties

C.1 Bayesian joint likelihood

× **Max. self-similarity:** Assume X is Gamma distributed with $L = 1$ and $x_1 = 2x_2$:

$$\mathcal{Q}_B(x_1, x_2) = \frac{1}{(x_1 + x_2)^2} = \frac{1}{9x_2^2} > \frac{1}{16x_2^2} = \frac{1}{(2x_2 + 2x_2)^2} = \frac{1}{(x_1 + x_1)^2} = \mathcal{Q}_B(x_1, x_1) \tag{110}$$

which breaks the property of max. self-similarity.

× **Eq. self-similarity:** Assume X is Gamma distributed with $L = 1$ and $x_1 = 2x_2$:

$$\mathcal{Q}_B(x_1, x_1) = \frac{1}{(x_1 + x_1)^2} = \frac{1}{(2x_2 + 2x_2)^2} = \frac{1}{16x_2^2} < \frac{1}{4x_2^2} = \frac{1}{(x_2 + x_2)^2} = \mathcal{Q}_B(x_2, x_2) \tag{111}$$

which breaks the property of eq. self-similarity.

× **Id. of indiscernible:** It requires the eq. self-similarity property.

× **Invariance:** Assume X is Gamma distributed with $L = 1$ and consider $X' = \sqrt{X}$, i.e., the mapping function $g(.) = \sqrt{.}$, then:

$$\mathcal{Q}_{BX_1,X_2}(x_1, x_2) = \frac{1}{(x_1 + x_1)^2} \tag{112}$$

$$\mathcal{Q}_{BX_1',X_2'}\mathcal{Q}_B(\sqrt{x_1}, \sqrt{x_2}) = \int p(\sqrt{x_1}|\theta_1 = t)p(\sqrt{x_2}|\theta_2 = t)p(\theta_{12} = t)\mathrm{d}t \tag{113}$$

$$= \left|\frac{\mathrm{d}\sqrt{x_1}}{\mathrm{d}x_1}\right|^{-1}\left|\frac{\mathrm{d}\sqrt{x_2}}{\mathrm{d}x_2}\right|^{-1}\int p(x_1|\theta_1 = t)p(x_2|\theta_2 = t)p(\theta_{12} = t)\mathrm{d}t \tag{114}$$

$$= 4\sqrt{x_1 x_2}\mathcal{Q}_{BX_1,X_2}(x_1, x_2)\ . \tag{115}$$

The equality does not hold for any value $x_1 > 0$ or $x_2 > 0$.

× **Asymp. CFAR:** The closed-from expression of $\mathcal{Q}_B$ obtained for Gamma distribution is clearly not asymptotically CFAR, since the expectation of the similarity criterion is inversely proportional to the underlying parameters.

× **Asymp. UMPI:** $\mathcal{L}_G$ being UMPI, it defeats $\mathcal{S}$.

C.2 Maximum joint likelihood

Since $\mathcal{Q}_G$ corresponds to $\mathcal{Q}_B$ in the Gamma case, we can use the same counter-examples as above.

C.3 Bayesian likelihood ratio

× **Max. self-similarity:** Assume X to take values in $\{x_1, x_2, x_3\}$ and $\theta \in \{a, b, c\}$. Assume the distribution of X to be defined by:

$$p(x_1|a) = 5/8 \qquad p(x_1|b) = 2/8 \qquad p(x_1|c) = 1/8 \tag{116}$$

$$p(x_2|a) = 2/8 \qquad p(x_2|b) = 4/8 \qquad p(x_2|c) = 3/8 \tag{117}$$

$$p(x_3|a) = 1/8 \qquad p(x_3|b) = 2/8 \qquad p(x_3|c) = 4/8\ . \tag{118}$$

Note that the observations are statistically identifiable through their likelihood and their MLE. Assume $p(\theta)$ to be described by

$$p(\theta = a) = 0/2 \tag{119}$$

$$p(\theta = b) = 1/2 \tag{120}$$

$$p(\theta = c) = 1/2\ . \tag{121}$$

The self Bayesian likelihood ratio for x_2 is given by

$$\mathcal{L}_B(x_2, x_2) = \frac{p(x_2|a)^2p(a) + p(x_2|b)^2p(b) + p(x_2|c)^2p(c)}{(p(x_2|a)p(a) + p(x_2|b)p(b) + p(x_2|c)p(c))^2} = \frac{\frac{2\times2\times0}{8\times8\times2} + \frac{4\times4\times1}{8\times8\times2} + \frac{3\times3\times1}{8\times8\times2}}{(\frac{2\times0}{8\times2} + \frac{4\times1}{8\times2} + \frac{3\times1}{8\times2})^2} = \frac{50}{49}\ . \tag{122}$$

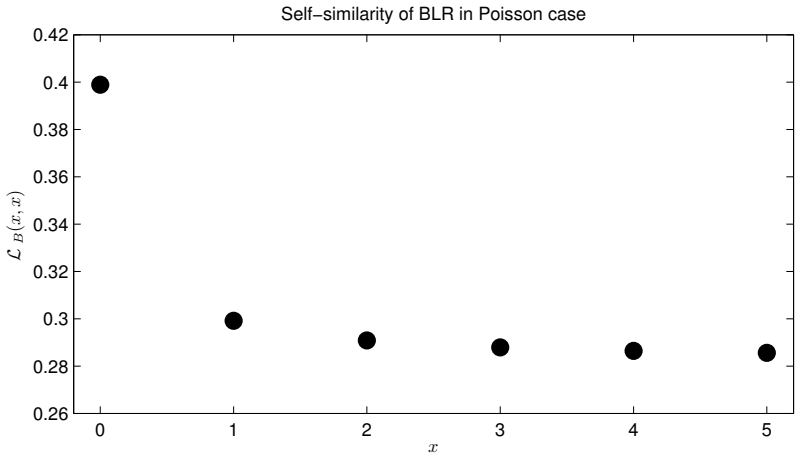


Fig. 6 Self Bayesian likelihood ratio $\mathcal{L}_B(x,x)$ with respect to the value x in the case of Poisson noise.

The Bayesian likelihood ratio between x_1 and x_2 is given by

$$\mathcal{L}_B(x_1,x_2) = \frac{(p(x_1|a)p(x_2|a)p(a) + p(x_1|b)p(x_2|b)p(b) + p(x_1|c)p(x_2|c)p(c))}{(p(x_1|a)p(a) + p(x_1|b)p(b) + p(x_1|c)p(c))(p(x_2|a)p(a) + p(x_2|b)p(b) + p(x_2|c)p(c))} \tag{123}$$

$$= \frac{\frac{5\times2\times0}{8\times8\times2} + \frac{2\times4\times1}{8\times8\times2} + \frac{1\times3\times1}{8\times8\times2}}{(\frac{5\times0}{8\times2} + \frac{2\times1}{8\times2} + \frac{1\times1}{8\times2})(\frac{2\times0}{8\times2} + \frac{4\times1}{8\times2} + \frac{3\times1}{8\times2})} = \frac{22}{21}\,. \tag{124}$$

Since $50/49 < 22/21$ then $\mathcal{L}_B(x_2,x_2) < \mathcal{L}_B(x_1,x_2)$. The max. self-similarity does not hold.

Open question: what are the sufficient and necessary conditions on the likelihood p to ensure the max. self similarity of $\mathcal{L}_B$?

× **Eq. self-similarity:** Consider the case of Poisson noise, the eq. self similarity is given by:

$$\mathcal{L}_B(x,x) = \frac{1}{\sqrt{2}}\frac{\Gamma(x+x+1/2)}{2^{x+x}\Gamma(x+1/2)\Gamma(x+1/2)} = \frac{1}{\sqrt{2}}\frac{\Gamma(2x+1/2)}{2^{2x}\Gamma(x+1/2)^2} \tag{125}$$

which depends, as illustrated on Figure 6, on the value of x.

× **Id. of indiscernible:** It requires the eq. self-similarity property.

√ **Invariance:** Let g be an invertible and differentiable mapping function of the rv X to X', then:

$$\mathcal{L}_{BX'_1,X'_2}(g(x_1),g(x_2)) = \frac{\int p(g(x_1)|\theta_{12}=t)p(g(x_2)|\theta_{12}=t)p(\theta_{12}=t)\,\mathrm{d}t}{\int p(g(x_1)|\theta_1=t)p(\theta_1=t)\,\mathrm{d}t \int p(g(x_2)|\theta_2=t)p(\theta_2=t)\,\mathrm{d}t} \tag{126}$$

$$= \frac{\left|\frac{\mathrm{d}g(x_1)}{\mathrm{d}x_1}\right|^{-1}\left|\frac{\mathrm{d}g(x_2)}{\mathrm{d}x_2}\right|^{-1}\int p(x_1|\theta_{12}=t)p(x_2|\theta_{12}=t)p(\theta_{12}=t)\,\mathrm{d}t}{\left|\frac{\mathrm{d}g(x_1)}{\mathrm{d}x_1}\right|^{-1}\left|\frac{\mathrm{d}g(x_2)}{\mathrm{d}x_2}\right|^{-1}\int p(x_1|\theta_1=t)p(\theta_1=t)\,\mathrm{d}t \int p(x_2|\theta_2=t)p(\theta_2=t)\,\mathrm{d}t} \tag{127}$$

$$= \mathcal{L}_{BX_1,X_2}(x_1,x_2) \tag{128}$$

The Bayesian likelihood ratio fulfils the invariance property.

× **Asymp. CFAR:** We can always choose a prior on the underlying parameters, favouring the similarity for a range of underlying parameters, implying that $\mathcal{L}_B$ would not be CFAR.

× **Asymp. UMPI:** $\mathcal{L}_G$ being UMPI, it defeats $\mathcal{S}$.

C.4 Generalized likelihood ratio

√ **Eq. self-similarity:** The self generalized likelihood ratio is always equal to one:

$$\mathcal{L}_G(x,x) = \frac{\sup_t p(x|\theta=t)^2}{(\sup_t p(x|\theta=t))^2} = 1 \tag{129}$$

since the superior bound is reached at the same value(s) t for $p(x|\theta=t)$ and $p(x|\theta=t)^2$.

√ **Max. self-similarity:** The superior bound of a product is always inferior to the product of the superior bounds, then:

$$\mathcal{L}_G(x_1,x_2) = \frac{\sup_t p(x_1|\theta_1=t)p(x_2|\theta_2=t)}{\sup_t p(x_1|\theta_1=t)\sup_t p(x_2|\theta_2=t)} \leq 1\,. \tag{130}$$

√ [5] **Id. of indiscernible:** Assume the observations are statistically identifiable through their MLE. Let two observations $x_1 \neq x_2$. Let $\hat{t}_1$ and $\hat{t}_2$ be respectively the maximum likelihood estimates of x_1 and x_2, and $\hat{t}_{12}$ be the maximum likelihood estimator of $\{x_1, x_2\}$. Since $x_1 \neq x_2$ and observations are statistically identifiable through their MLE, $\hat{t}_1 \neq \hat{t}_2$. Since the MLE is unique, then, either:

$$p(x_1|\theta_1 = \hat{t}_1) \geq p(x_1|\theta_1 = \hat{t}_{12}) > 0 \tag{131}$$
$$p(x_2|\theta_2 = \hat{t}_2) > p(x_2|\theta_2 = \hat{t}_{12}) > 0 \tag{132}$$

or

$$p(x_1|\theta_1 = \hat{t}_1) > p(x_1|\theta_1 = \hat{t}_{12}) > 0 \tag{133}$$
$$p(x_2|\theta_2 = \hat{t}_2) \geq p(x_2|\theta_2 = \hat{t}_{12}) > 0 \tag{134}$$

Then, in any case, $p(x_1|\theta_1 = \hat{t}_1)p(x_2|\theta_2 = \hat{t}_1) > p(x_1|\theta_1 = \hat{t}_{12})p(x_2|\theta_2 = \hat{t}_{12})$, i.e., $x_1 \neq x_2 \Rightarrow \mathcal{L}_G(x_1, x_2) < 1$.

√ **Invariance:** Let g be an invertible and differentiable mapping function of the rv X to X', then:

$$\mathcal{L}_{GX'_1,X'_2}(g(x_1), g(x_2)) = \frac{\sup_t p(g(x_1)|\theta_1 = t)p(g(x_2)|\theta_2 = t)}{\sup_t p(g(x_1)|\theta_1 = t)\sup_t p(g(x_2)|\theta_2 = t)} \tag{135}$$
$$= \frac{\left|\frac{\mathrm{d}g(x_1)}{\mathrm{d}x_1}\right|^{-1}\left|\frac{\mathrm{d}g(x_2)}{\mathrm{d}x_2}\right|^{-1}\sup_t p(x_1|\theta_1 = t)p(x_2|\theta_2 = t)}{\left|\frac{\mathrm{d}g(x_1)}{\mathrm{d}x_1}\right|^{-1}\left|\frac{\mathrm{d}g(x_2)}{\mathrm{d}x_2}\right|^{-1}\sup_t p(x_1|\theta_1 = t)\sup_t p(x_2|\theta_2 = t)} \tag{136}$$
$$= \mathcal{L}_{GX_1,X_2}(x_1, x_2) \tag{137}$$

The generalized likelihood ratio fulfils the invariance property (see also Kay and Gabriel, 2003)).

√ **Asymp. CFAR:** According to (Kay, 1998).

√ **Asymp. UMPI:** Due to its convergence to the likelihood ratio $\mathcal{L}$, which is Neyman-Pearson optimal, $\mathcal{L}_G$ is UMPI (Lehmann, 1959).

C.5 Mutual information kernel

√ **Eq. self-similarity:** The self mutual information kernel is always equal to one:

$$\mathcal{K}_B(x, x) = \frac{\mathcal{Q}_B(x, x)}{\sqrt{\mathcal{Q}_B(x, x)\mathcal{Q}_B(x, x)}} = 1\ . \tag{138}$$

√ **Max. self-similarity:** This property derived directly from the Cauchy-Schwartz inequality.

√ [6] **Id. of indiscernible:** Assume the observations are statistically identifiable though their likelihood. See θ as a random variable with distribution $p(\theta)$. Let $P_1 = p(x_1|\theta)$ and $P_2 = p(x_2|\theta)$ be the two r.v. resulting of the evaluation of the likelihood of the r.v. θ. We can rewrite the mutual information kernel as the correlation between P_1 and P_2:

$$\mathcal{K}_B(x, x) = \frac{\mathbb{E}[P_1P_2]}{\sqrt{\mathbb{E}[P_1^2]\mathbb{E}[P_2^2]}} \tag{139}$$

We get that the mutual information is maximal if the correlation between P_1 and P_2 is equal to one:

$$\mathcal{K}_B(x, x) = 1 \Rightarrow \quad \frac{\mathbb{E}[P_1P_2]}{\sqrt{\mathbb{E}[P_1^2]\mathbb{E}[P_2^2]}} = 1 \tag{140}$$

i.e., for all θ, $p(x_1|\theta) = a\,p(x_2|\theta)$ with $a > 0$ since a pdf is a positive function. Under normalization constraint and since the observations are statistically identifiable though their likelihood, $x_1 = x_2$.

√ **Invariance:** Let g be an invertible and differentiable mapping function of the rv X to X', then:

$$\mathcal{Q}_{GX'_1,X'_2}(g(x_1), g(x_2)) = \frac{\int p(g(x_1)|\theta_1 = t)p(g(x_2)|\theta_2 = t)p(\theta_{12} = t)\mathrm{d}t}{\sqrt{\int p(g(x_1)|\theta_1 = t)^2p(\theta_{12} = t)\mathrm{d}t\int p(g(x_2)|\theta_2 = t)^2p(\theta_{12} = t)\mathrm{d}t}} \tag{141}$$
$$= \frac{\left|\frac{\mathrm{d}g(x_1)}{\mathrm{d}x_1}\right|^{-1}\left|\frac{\mathrm{d}g(x_2)}{\mathrm{d}x_2}\right|^{-1}\int p(x_1|\theta_1 = t)p(x_2|\theta_2 = t)p(\theta_{12} = t)\mathrm{d}t}{\left|\frac{\mathrm{d}g(x_1)}{\mathrm{d}x_1}\right|^{-1}\left|\frac{\mathrm{d}g(x_2)}{\mathrm{d}x_2}\right|^{-1}\sqrt{\int p(x_1|\theta_1 = t)^2p(\theta_{12} = t)\mathrm{d}t\int p(x_2|\theta_2 = t)^2p(\theta_{12} = t)\mathrm{d}t}} \tag{142}$$
$$= \mathcal{Q}_{GX_1,X_2}(x_1, x_2) \tag{143}$$

The mutual information kernel fulfils the invariance property.

× **Asymp. CFAR:** We can always choose a prior on the underlying parameters, favouring the similarity for a range of underlying parameters, implying that $\mathcal{Q}_G$ would not be CFAR.

× **Asymp. UMPI:** $\mathcal{L}_G$ being UMPI, it defeats $\mathcal{S}$.

[5] Holds true under the assumption that the observations are statistically identifiable through their MLE.

[6] Holds true under the assumption that the observations are statistically identifiable through their likelihood.

C.6 Variance stabilization criterion

It is important to note that all properties below require that a variance stabilizer s exists.

√ **Eq. self-similarity:** Thanks to the Gaussian kernel, the self similarity of $\mathcal{S}$ is always equal to one:

$$\mathcal{S}(x,x) = \exp\left(\frac{\|s(x)-s(x)\|_2^2}{h}\right) = 1\ . \tag{144}$$

√ **Max. self-similarity:** This property follows from the property of the Euclidean distance:

$$\begin{aligned}
&\|s(x_1)-s(x_2)\|_2^2 \geq 0 && (145)\\
\Leftrightarrow\ &\|s(x_1)-s(x_2)\|_2^2 \geq \|s(x_1)-s(x_1)\|_2^2 && (146)\\
\Leftrightarrow\ &\exp\left(-\frac{\|s(x_1)-s(x_2)\|_2^2}{h}\right) \leq \exp\left(-\frac{\|s(x_1)-s(x_1)\|_2^2}{h}\right) && (147)\\
\Leftrightarrow\ &\mathcal{S}(x_1,x_2) \leq \mathcal{S}(x_1,x_1)\ . && (148)
\end{aligned}$$

√ **Id. of indiscernible:** This property is obtained as follows:

$$\begin{aligned}
&\mathcal{S}(x_1,x_2) = 1 && (149)\\
\Rightarrow\ &\exp\left(-\frac{\|s(x_1)-s(x_2)\|_2^2}{h}\right) = 1 && (150)\\
\Rightarrow\ &\|s(x_1)-s(x_2)\|_2^2 = 0 && (151)\\
\Rightarrow\ &s(x_1) = s(x_2) && (152)\\
\Rightarrow\ &x_1 = x_2 \quad \text{since } s \text{ is invertible}\ . && (153)
\end{aligned}$$

√ **Invariance:** If s stabilizes the variance of X then $s \circ g^{-1}$ stabilizes the variance of $g(X)$. Hence:

$$\mathcal{S}_{X_1',X_2'}(g(X_1),g(X_2)) = \mathcal{N}((s\circ g^{-1})(g(X_1)),(s\circ g^{-1})(g(X_2))) = \mathcal{N}(s(X_1),s(X_2)) = \mathcal{S}_{X_1,X_2}(X_1,X_2)\ . \tag{154}$$

√ **Asymp. CFAR:** If s stabilizes the variance of X, and given that $\mathbb{E}[\|s(X)-s(X)\|_2^2] = 2\mathrm{Var}[s(X)]$, then $\mathcal{S}$ is asymptotically CFAR.

× **Asymp. UMPI:** $\mathcal{L}_G$ being UMPI, it defeats $\mathcal{S}$.

References

Alter F, Matsushita Y, Tang X (2006) An intensity similarity measure in low-light conditions. Lecture Notes in Computer Science 3954:267

Baxter J (1995) Learning internal representations. In: Proceedings of the eighth annual conference on Computational learning theory, ACM, pp 311–320

Baxter J, Bartlett P (1998) The canonical distortion measure in feature space and 1-nn classification. In: Advances in neural information processing systems 10: proceedings of the 1997 conference, The MIT Press, vol 10, p 245

Boulanger J, Kervrann C, Bouthemy P, Elbau P, Sibarita J, Salamero J (2010) Patch-based nonlocal functional for denoising fluorescence microscopy image sequences. Medical Imaging, IEEE Transactions on 29(2):442–454

Boykov Y, Veksler O, Zabih R (1998) Markov random fields with efficient approximations. In: Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on, IEEE, pp 648–655

Boykov Y, Veksler O, Zabih R (2001) Fast approximate energy minimization via graph cuts. IEEE Transactions on pattern analysis and machine intelligence pp 1222–1239

Brown L, Cai T, Zhang R, Zhao L, Zhou H (2010) The root–unroot algorithm for density estimation as implemented via wavelet block thresholding. Probability theory and related fields 146(3):401–433

Buades A, Coll B, Morel J (2005a) A Non-Local Algorithm for Image Denoising. In: Proc. IEEE Computer Society Conf. CVPR, vol 2, pp 60–65

Buades A, Coll B, Morel J (2005b) A Review of Image Denoising Algorithms, with a New One. Multiscale Modeling and Simulation 4(2):490

Buades A, Coll B, Morel JM (2009) Non-local means denoising. Image Processing on Line URL http://www.ipol.im/pub/algo/bcm_non_local_means_denoising/

Chen J, Chen Y, An W, Cui Y, Yang J (2011) Nonlocal Filtering for Polarimetric SAR Data: A Pretest Approach. Geoscience and Remote Sensing, IEEE Transactions on 49(5):1744 –1754

Cho T, Avidan S, Freeman W (2009) The patch transform. IEEE Transactions on Pattern Analysis and Machine Intelligence pp 1489–1501

Comaniciu D, Ramesh V, Meer P (2003) Kernel-Based Object Tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence pp 564–577

Criminisi A, Pérez P, Toyama K (2004) Region filling and object removal by exemplar-based image inpainting. Image Processing, IEEE Transactions on 13(9):1200–1212

Dabov K, Foi A, Katkovnik V, Egiazarian K (2007) Image denoising by sparse 3-D transform-domain collaborative filtering. IEEE Transactions on image processing 16(8):2080

Dabov K, Foi A, Katkovnik V, Egiazarian K (2008) A Nonlocal and Shape-Adaptive Transform-Domain Collaborative Filtering. In: Int. Workshop on Local and Non-Local Approximation in Image Processing, LNLA

Deledalle C, Denis L, Tupin F (2009a) Débruitage Non-Local Itératif fondé sur un Critère de Similarité Probabiliste. In the proceedings of GRETSI, Dijon, France, September 2009

Deledalle C, Denis L, Tupin F (2009b) Iterative Weighted Maximum Likelihood Denoising with Probabilistic Patch-Based Weights. IEEE Transactions on Image Processing 18(12):2661–2672, DOI 10.1109/TIP.2009.2029593

Deledalle C, Tupin F, Denis L (2010) Poisson NL means: Unsupervised non local means for Poisson noise. In: Image Processing (ICIP), 2010 17th IEEE International Conference on, IEEE, pp 801–804

Deledalle C, Denis L, Tupin F (2011a) NL-InSAR : Non-Local Interferogram Estimation. IEEE Transaction on Geoscience and Remote Sensing 49(4)

Deledalle CA, Duval V, Salmon J (2011b) Non-local Methods with Shape-Adaptive Patches (NLM-SAP). Journal of Mathematical Imaging and Vision pp 1–18

Deledalle CA, Tupin F, Denis L (2011c) Patch similarity under non-gaussian noise. In: Image Processing (ICIP), 2011 18th IEEE International Conference on, IEEE

Efros A, Freeman W (2001) Image quilting for texture synthesis and transfer. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques, ACM, pp 341–346

Elad M, Aharon M (2006) Image denoising via sparse and redundant representations over learned dictionaries. Image Processing, IEEE Transactions on 15(12):3736–3745

Freeman W, Jones T, Pasztor E (2002) Example-based super-resolution. IEEE Computer Graphics and Applications pp 56–65

Gilboa G, Osher S (2008) Nonlocal linear image regularization and supervised segmentation. Multiscale Modeling and Simulation 6(2):595–630

Goodman J (1976) Some fundamental properties of speckle. J Opt Soc Am 66(11):1145–1150

Hartley R, Zisserman A (2000) Multiple view geometry, vol 642. Cambridge university press

Horn B, Schunck B (1981) Determining optical flow. Artificial intelligence 17(1-3):185–203

Hudson HM (1978) A natural identity for exponential families with applications in multiparameter estimation. Ann Statist 6(3):473–484

Hyvärinen A, Hurri J, Hoyer P (2009) Natural Image Statistics: A probabilistic approach to early computational vision. Springer-Verlag New York Inc

Ishikawa H (2003) Exact optimization for Markov random fields with convex priors. IEEE Transactions on Pattern Analysis and Machine Intelligence pp 1333–1336

Jain A (1989) Fundamentals of digital image processing. Prentice-Hall, Inc. Upper Saddle River, NJ, USA

Katkovnik V, Foi A, Egiazarian K, Astola J (2010) From local kernel to nonlocal multiple-model image denoising. International journal of computer vision 86(1):1–32

Kay S (1998) Fundamentals of statistical signal processing Volume 2: Detection theory. Prentice Hall

Kay S, Gabriel J (2003) An invariance property of the generalized likelihood ratio test. Signal Processing Letters, IEEE 10(12):352–355

Kendall M, Stuart A (1979) The advanced theory of statistics. Vol. 2: Inference and relationship. Charles Griffin and Co., Ltd., London

Kervrann C, Boulanger J (2008) Local Adaptivity to Variable Smoothness for Exemplar-Based Image Regularization and Representation. International Journal of Computer Vision 79(1):45–69

Kervrann C, Boulanger J, Coupé P (2007) Bayesian non-local means filter, image redundancy and adaptive dictionaries for noise removal. In: Proceedings of the 1st international conference on Scale space and variational methods in computer vision, Springer-Verlag, pp 520–532

Kim H, Hero III A (2001) Comparison of GLR and invariant detectors under structured clutter covariance. Image Processing, IEEE Transactions on 10(10):1509–1520

Kwatra V, Schödl A, Essa I, Turk G, Bobick A (2003) Graphcut textures: image and video synthesis using graph cuts. ACM Transactions on Graphics (TOG) 22(3):277–286

Lehmann E (1959) Optimum invariant tests. The Annals of Mathematical Statistics 30(4):881–884

Liang L, Liu C, Xu Y, Guo B, Shum H (2001) Real-time texture synthesis by patch-based sampling. ACM Transactions on Graphics (ToG) 20(3):127–150

Lowe D (1992) Robust model-based motion tracking through the integration of search and estimation. International Journal of Computer Vision 8(2):113–122

Mairal J, Bach F, Ponce J, Sapiro G, Zisserman A (2009) Non-local sparse models for image restoration. ICCV

Mäkitalo M, Foi A (2011) Optimal inversion of the anscombe transformation in low-count poisson image denoising. Image Processing, IEEE Transactions on 20(1):99–109

Mäkitalo M, Foi A, Fevralev D, Lukin V (2010) Denoising of single-look SAR images based on variance stabilization and nonlocal filters. In: Proc. Int. Conf. Math. Meth. Electromagn. Th., MMET 2010, Kiev, Ukraine

Matsushita Y, Lin S (2007) A Probabilistic Intensity Similarity Measure based on Noise Distributions. In: IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07, pp 1–8

Minka T (1998) Bayesian Inference, Entropy, and the Multinomial Distribution. Tech. rep., CMU

Minka T (2000) Distance measures as prior probabilities. Tech. rep., CMU

Parrilli S, Poderico M, Angelino C, Scarpa G, Verdoliva L (2010) A nonlocal approach for sar image denoising. In: Geoscience and Remote Sensing Symposium (IGARSS), 2010 IEEE International, IEEE, pp 726–729

Peyré G, Bougleux S, Cohen L (2008) Non-local regularization of inverse problems. In: Computer Vision–ECCV 2008, Springer, pp 57–68

Salmon J (2010) On two parameters for denoising with Non-Local Means. IEEE Signal Process Lett 17:269–272

Scharstein D, Szeliski R (2002) A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International journal of computer vision 47(1):7–42

Seeger M (2002) Covariance kernels from Bayesian generative models. In: Advances in neural information processing systems 14: proceedings of the 2001 conference, MIT Press, p 905

Teuber T, Lang A (2011) A new similarity measure for nonlocal filtering in the presence of multiplicative noise. Preprint University of Kaiserslautern

Van De Ville D, Kocher M (2009) SURE-Based Non-Local Means. IEEE Signal Processing Letters 16(11):973–976

Varma M, Zisserman A (2003) Texture classification: Are filter banks necessary? In: Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, IEEE, vol 2, pp II–691

Xie H, Pierce L, Ulaby F (2002) Statistical properties of logarithmically transformed speckle. IEEE Transactions on Geoscience and Remote Sensing 40(3):721–727

Yianilos P (1995) Metric learning via normal mixtures. Tech. rep., NEC Research Institute, Princeton, NJ

Zhang X, Burger M, Bresson X, Osher S (2010) Bregmanized nonlocal regularization for deconvolution and sparse reconstruction. SIAM Journal on Imaging Sciences 3(3):253–276, DOI 10.1137/090746379, URL http://link.aip.org/link/?SII/3/253/1

Zitova B, Flusser J (2003) Image registration methods: a survey. Image and vision computing 21(11):977–1000