Estimation of Kullback-Leibler losses for noisy recovery problems within the exponential family

Charles-Alban Deledalle[*]

IMB, CNRS, Université de Bordeaux, Bordeaux INP
351 cours de la Libération, F-33405 Talence, France
e-mail: charles-alban.deledalle@math.u-bordeaux.fr

Abstract: We address the question of estimating Kullback-Leibler losses rather than squared losses in recovery problems where the noise is distributed within the exponential family. Inspired by Stein unbiased risk estimator (SURE), we exhibit conditions under which these losses can be unbiasedly estimated or estimated with a controlled bias. Simulations on parameter selection problems in applications to image denoising and variable selection with Gamma and Poisson noises illustrate the interest of Kullback-Leibler losses and the proposed estimators.

Contents

1. Introduction

We consider the problem of predicting an unknown d-dimensional vector $\mu \in \mathbb{R}^d$ from its noisy measurements $V \in \mathbb{R}^d$. Given a collection of parametric predictors of μ, we focus on the selection of the predictor $\hat{\mu}$ that minimizes the discrepancy with the unknown vector μ. For instance, this includes the problem of selecting the best predictors from the set of Least Absolute Shrinkage and Selection Operator (LASSO) solutions [44] obtained for all possible choices of regularization parameters. To this end, the common approach is to select $\hat{\mu}$ that minimizes an unbiased estimate of the expected squared loss $\mathbb{E}\|\mu - \hat{\mu}\|^2$, typically, with the Stein unbiased risk estimator (SURE) [43]. Such estimators are classically built on some statistical modeling of the noise, e.g., as being distributed within the exponential family. In this context, we investigate the interest of going beyond squared losses by rather estimating a loss function grounded on an information based criterion, namely, the Kullback-Leibler divergence. We will first recall some basic properties of the exponential family, give a quick review on risk estimation and motivate the use of the Kullback-Leibler divergence.

Exponential family.   We assume that in the aforementioned recovery problem the noise distribution belongs to the exponential family. Formally, the recovery problem can be reparametrized using two one-to-one mappings $\psi : \mathbb{R}^d \to \mathbb{R}^d$ and $\phi : \mathbb{R}^d \to \mathbb{R}^d$ such that $Y = \psi(V)$ has a probability measure P_θ characterized by a probability density or mass function with respect to the Lebesgue measure $\mathrm{d}y$ of the following form

$$p(y; \theta) = h(y) \exp\left(\langle y, \theta \rangle - A(\theta)\right) \tag{1.1}$$

where $\theta = \phi(\mu) \in \mathbb{R}^d$. The distribution P_θ is said to be within the natural exponential family. We call θ the natural parameter, Y a sufficient statistic for θ, $h : \mathbb{R}^d \to \mathbb{R}^+$ the base measure, and $A : \mathbb{R}^d \to \mathbb{R}$ the log-partition function. Classical and important properties of the exponential family include A that is convex, $\mathbb{E}[Y] = \nabla A(\theta)$ and $\mathrm{Var}[Y] = \nabla\nabla^t A(\theta)$ (see, e.g., [3]). Here and in the following, $\mathbb{E}[Y] = \int Y \mathrm{d}P_\theta$ denotes the expectation of the random vector Y with respect to the measure $\mathrm{d}P_\theta$, and $\mathrm{Var}[Y] = \mathbb{E}[(Y - \mathbb{E}[Y])(Y - \mathbb{E}[Y])^t]$ is its so-called variance-covariance matrix.

Without loss of generality, we consider that Y is a minimal sufficient statistic. As a consequence, ∇A is one-to-one and we can choose ϕ as the canonical link function satisfying $\phi = (\nabla A)^{-1}$ (as coined in the language of generalized linear models). An immediate consequence is that Y has expectation $\mathbb{E}[Y] = \mu$ and its variance is a function of μ given by $\mathrm{Var}[Y] = \Lambda(\mu)$ where $\Lambda = (\nabla\nabla^t A) \circ \phi$. The function $\Lambda : \mathbb{R}^d \to \mathbb{R}^{d \times d}$ is the so-called variance function (see, e.g., [36]), also known as the noise level function (in the language of signal processing).

Table 1 gives five examples of univariate distributions of the exponential family – two of them are defined in a continuous domain, the other three are defined in a discrete domain.

<div align="center">

TABLE 1

Examples of univariate distributions of the exponential family. The variable y denotes an outcome of the random variable Y, $\mu = \mathbb{E}[Y]$ is the unknown (location) parameter of interest, and the variables in brackets are known nuisance (scale and shape) parameters.

</div>

Distribution law	$\theta = \phi(\mu)$	$\Lambda(\mu)$	$h(y)$	$A(\theta)$
Gaussian $(\sigma > 0)$ $\dfrac{1}{\sqrt{2\pi}\sigma} \exp\left(-\dfrac{(y-\mu)^2}{2\sigma^2}\right)$ $\mu \in \mathbb{R}$	$\dfrac{\mu}{\sigma^2}$	σ^2	$\dfrac{e^{-\frac{y^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}$	$\dfrac{\sigma^2\theta^2}{2}$
Gamma $(L > 0)$ $\dfrac{L^L y^{L-1}}{\Gamma(L)\mu^L} e^{-\frac{Ly}{\mu}} \mathbb{1}_{\mathbb{R}^+}(y)$ $\mu > 0$	$\dfrac{-L}{\mu}$	$\dfrac{\mu^2}{L}$	$\dfrac{L^L y^{L-1}}{\Gamma(L)}\mathbb{1}_{\mathbb{R}^+}(y)$	$-L\log(-\theta/L)$
Poisson $\dfrac{\mu^y e^{-\mu}}{y!}\mathbb{1}_{\mathbb{N}}(y)$ $\mu > 0$	$\log\mu$	μ	$\dfrac{\mathbb{1}_{\mathbb{N}}(y)}{y!}$	$\exp\theta$
Binomial $(n > 0)$ $\dbinom{n}{y}\mathsf{p}^y(1-\mathsf{p})^{n-y}\mathbb{1}_{[0,n]}(y)$ $\mu = n\mathsf{p},\ \mathsf{p} \in [0,1]$	$\log\dfrac{\mu}{n-\mu}$	$-\dfrac{\mu^2}{n}+\mu$	$\dbinom{n}{y}\mathbb{1}_{[0,n]}(y)$	$n\log(1+e^\theta)$
Negative Binomial $(r > 0)$ $\dfrac{\Gamma(r+y)}{y!\Gamma(r)}\mathsf{p}^y(1-\mathsf{p})^r\mathbb{1}_{\mathbb{N}}(y)$ $\mu = r\mathsf{p}/(1-\mathsf{p}),\ \mathsf{p} \in [0,1]$	$\log\dfrac{\mu}{r+\mu}$	$\dfrac{\mu^2}{n}-\mu$	$\dfrac{\Gamma(r+y)}{y!\Gamma(r)}\mathbb{1}_{\mathbb{N}}(y)$	$-r\log(1-e^\theta)$

Risk estimation. We now assume that the predictor $\hat{\mu}$ of μ is a function of Y only, hence, we write it $\hat{\mu}(Y)$, and we focus on estimating the loss associated to $\hat{\mu}(Y)$ with respect to μ. When the noise has a Gaussian distribution with independent entries, SURE [43] can be used to estimate the mean squared error (MSE), or in short the risk, defined as: $\text{MSE}_\mu = \mathbb{E}\|\mu - \hat{\mu}(Y)\|^2$. The resulting estimator, being independent on the unknown predictor μ, can serve in practice as an objective for parameter selection. Eldar [15] builds on Stein's lemma [43], a generalization of SURE valid for some continuous distributions of the exponential family. It provides an unbiased estimate of the "natural" risk, defined as: $\text{MSE}_\theta = \mathbb{E}\|\phi(\mu) - \phi(\hat{\mu}(Y))\|^2$, i.e., the risk with respect to $\theta = \phi(\mu)$. In the same vein, when the distribution is discrete, Hudson [26] provides another result for estimating the "exp-natural" risk: $\text{MSE}_\eta = \mathbb{E}\|\exp\phi(\mu) - \exp\phi(\hat{\mu}(Y))\|^2$, i.e., the risk with respect to $\eta = \exp\theta$, where $\exp: \mathbb{R}^d \to \mathbb{R}^d$ is the entry-wise exponential. As ϕ is assumed one-to-one, there is no doubt that if such loss functions cancel then $\hat{\mu}(Y) = \mu$. In this sense, they provide good objectives for selecting $\hat{\mu}(Y)$. However, within a family of parametric predictors and without strong assumptions on μ, such a loss function might never cancel. In such a case, it becomes unclear what its minimization leads it to select, all the more when ϕ or $\exp\circ\phi$ are non-linear. Furthermore, even when they are linear (e.g., $\exp\circ\phi = \text{id}$ for Poisson noise), minimizing $\text{MSE}_\mu = \mathbb{E}\|\mu - \hat{\mu}(Y)\|^2$ might not even be relevant

as it does not compensate for the heteroscedasticity of the noise (this will be made clear in our experiments). Estimating the reweighted or Mahanalobis risk given by $\mathbb{E}\|\Lambda(\mu)^{-1/2}(\mu - \hat{\mu}(Y))\|^2$ could be more relevant in this case, but its estimation is more intricate.

Kullback-Leibler divergence. The Kullback-Leibler (KL) divergence [27] is a measure of information loss when an alternative distribution P_1 is used to approximate the underlying one P_0. Its formal definition is given by $\mathcal{D}(P_0\|P_1) = \int \mathrm{d}P_0 \log \frac{\mathrm{d}P_0}{\mathrm{d}P_1}$. Unlike squared losses, it does not measure the discrepancy between an unknown parameter and its estimate, but between the unknown distribution P_0 of Y and its estimate P_1. As a consequence, it is invariant with one-to-one reparametrization of the parameters and, hence, becomes a serious competitor to squared losses. Remark that it is also invariant under one-to-one transformations of Y because such transforms do not affect the quantity of information carried by Y. Interestingly, provided P_0 and P_1 belongs to the same member of the natural exponential family respectively with parameters θ_0 and θ_1, the KL divergence can be written in terms of the Bregman divergence associated with A for points θ_0 and θ_1, i.e.,

$$\mathcal{D}(P_0\|P_1) = A(\theta_1) - A(\theta_0) - \langle \nabla A(\theta_0), \theta_1 - \theta_0 \rangle. \qquad (1.2)$$

While squared losses are defined irrespective of the noise distribution, the KL divergence adjusts its penalty with respect to the scales and the shapes of the deviations. In particular, it accounts for heteroscedasticity.

Contributions. In this paper, we address the problem of estimating KL losses, i.e., losses based on the KL divergence. As it is a non symmetric discrepancy measure, we can define two KL loss functions. The first one

$$\mathrm{MKLA} = \mathbb{E}[\mathcal{D}(P_\theta\|P_{\hat{\theta}(Y)})] \qquad (\mathrm{MKLA})$$

will be referred to as the mean KL analysis loss as it can be given the following interpretation: "how well might $P_{\hat{\theta}(Y)}$ explain independent copies of Y". The mean KL analysis loss is inherent to many statistical problems as it takes as reference the true underlying distribution. It is at the heart of the maximum likelihood estimator and is typically involved in non-parametric density estimation, oracle inequalities, mini-max control, etc. (see, e.g., [22, 17, 41]). The second one will be referred to as the mean KL synthesis loss given by

$$\mathrm{MKLS} = \mathbb{E}[\mathcal{D}(P_{\hat{\theta}(Y)}\|P_\theta)] \qquad (\mathrm{MKLS})$$

which can be given the following interpretation: "how well might $P_{\hat{\theta}(Y)}$ generate independent copies of Y". The mean KL synthesis loss has also been considered in different statistical studies. For instance, the authors of [48] consider this loss function to design a James Stein-like shrinkage predictor. Hannig and Lee address a very similar problem to ours, by designing a consistent estimator of MKLS used as an objective for bandwidth selection in kernel smoothing

TABLE 2
Summary of what can be estimated provided $y \mapsto \hat{\mu}(y)$ is sufficiently regular.

	Continuous	Discrete	
MSE_μ	if $\phi(\mu) = \alpha\mu + \beta$ (Gaussian)	if $\phi(\mu) = \log(\alpha\mu + \beta)$ (Poisson)	Stein & Hudson
MSE_θ	yes		
MSE_η		yes	
MKLA	if $\phi(\mu) = \alpha\mu + \beta$ (Gaussian)	if $\phi(\mu) = \log(\alpha\mu + \beta)$ (Poisson)	Our contributions
	yes, when Y results from a large sample mean (Gaussian, Gamma, ...)	(Poisson, NegBin, Binomial, ...)	
MKLS*	yes		

*consistently for kernel smoothing under Gamma [24] and Poisson noises [25].

problems subject to Gamma [24] and Poisson noise [25]. Table 2 gives a summary of our contributions. It highlights which loss can be estimated and under which conditions of the exponential family. The main contributions of our paper are:

1. provided $y \mapsto \hat{\mu}(y)$ and the base measure h are both weakly differentiable, MKLS can be unbiasedly estimated (Theorem 4.1),
2. for any mapping $y \mapsto \hat{\mu}(y)$, MKLA can be unbiasedly estimated for Poisson variates (Theorem 4.2),
3. provided $y \mapsto \hat{\mu}(y)$ is $k \geq 3$ times differentiable with bounded k-th derivative, MKLA can be estimated with vanishing bias when Y results from a large sample mean of independent random vectors with finite k-th order moments (Theorem 4.3).

It is worth mentioning that a symmetrized version of the mean Kullback-Leibler loss: MKLA + MKLS, can be estimated as soon as MKLA and MKLS can both be estimated (e.g., for continuous distributions according to Table 2).

2. Risk estimation under Gaussian noise

This section recalls important properties of the MSE and the definition of SURE under additive noise models of the form $Y = \mu + Z$ where $Z \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_d)$ and Id_d denotes the $d \times d$ identity matrix.

Before turning to the unbiased estimation of MSE_μ, it is important to recall that for any additive models and zero-mean noise with variance $\sigma^2 \mathrm{Id}_d$, provided the following quantities exists, we have

$$\mathrm{MSE}_\mu = \underbrace{\mathbb{E}\|Y - \hat{\mu}(Y)\|^2 - d\sigma^2}_{\text{expected data fidelity}} + 2\underbrace{\mathrm{tr}\,\mathrm{Cov}(Y, \hat{\mu}(Y))}_{\text{model complexity}} \tag{2.1}$$

where $\mathrm{Cov}(Y, \hat{\mu}(Y)) = \mathbb{E}[(Y - \mathbb{E}[Y])(\hat{\mu}(Y) - \mathbb{E}[\hat{\mu}(Y)])^t]$ is the cross-covariance matrix between Y and $\hat{\mu}(Y)$. Equation (2.1) gives a variational interpretation of the minimization of the MSE as the optimization of a trade-off between overfitting (first term) and complexity (second term). In fact, $\sigma^{-2}\mathrm{tr}\,\mathrm{Cov}(Y, \hat{\mu}(Y))$ is a

classical measure of the complexity of a statistical modeling procedure, known as the degrees of freedom (DOF), see, e.g., [13]. The DOF plays an important role in model validation and model selection rules, such as, Akaike information criteria (AIC) [1], Bayesian information criteria (BIC) [42], and the generalized cross-validation (GCV) [20].

For linear predictors of the form $\hat{\mu}(y) = Wy$, $W \in \mathbb{R}^{d \times d}$ (think of least-square or ridge regression), the DOF boils down to $\operatorname{tr} W$. As a consequence, the random quantity $\|Y - \hat{\mu}(Y)\|^2 - d\sigma^2 + 2\operatorname{tr} W$ becomes an unbiased estimator of MSE_μ, that depends solely on Y without *prior* knowledge of μ. If W is a projector, the DOF corresponds to the dimension of the target space, and we retrieve the well known Mallows' C_p statistic [35] as well as the aforementioned AIC. The SURE provides a generalization of these results that is not only restricted to linear predictors but can be applied to weakly differentiable mappings. A comprehensive account on weak differentiability can be found in e.g., [16, 18]. Let us now recall Stein's lemma [43].

Lemma 1 (Stein lemma). *Assume f is weakly differentiable with essentially bounded weak partial derivatives on $\mathbb{R}^d$ and $Y \sim \mathcal{N}(\mu, \sigma^2 \mathrm{Id}_d)$, then*

$$\operatorname{tr} \operatorname{Cov}(Y, f(Y)) = \sigma^2 \mathbb{E} \left[\operatorname{tr} \left. \frac{\partial f(y)}{\partial y} \right|_Y \right].$$

A direct consequence of Stein's Lemma, provided $\hat{\mu}$ fulfills the assumptions of Lemma 1, is that

$$\mathrm{SURE} = \|Y - \hat{\mu}(Y)\|^2 - d\sigma^2 + 2\sigma^2 \operatorname{tr} \left. \frac{\partial \hat{\mu}(y)}{\partial y} \right|_Y \tag{2.2}$$

satisfies $\mathbb{E}\mathrm{SURE} = \mathrm{MSE}_\mu$. Applications of SURE emerged for choosing the smoothing parameters in families of linear predictors [30] such as for model selection, ridge regression, smoothing splines, etc. After its introduction in the wavelet community with the SURE-Shrink algorithm [11], it has been widely used to various image restoration problems, e.g., with sparse regularizations [2, 38, 6, 37, 5, 32, 39, 45] or with non-local filters [46, 12, 9, 47].

3. Risk estimation for the exponential family and beyond

In this section, we recall how SURE has been extended beyond Gaussian noises towards noises distributed within the natural exponential family.

Continuous exponential family.   We first consider continuous noise models, e.g., Gamma noise. To begin, we recall a well known result derived by Eldar [14], that can be traced back to Hudson[1] in the case of independent entries [26], and that can be seen as a generalization of Stein's lemma.

---

[1]In his paper, Hudson mentioned that Stein already knew about this result.

Lemma 2 (Generalized Stein's lemma). *Assume f is weakly differentiable with essentially bounded weak partial derivatives on $\mathbb{R}^d$ and Y follows a distribution of the natural exponential family with natural parameter θ, provided h is also weakly differentiable on $\mathbb{R}^d$, we have*

$$\mathbb{E}\langle \theta, f(Y) \rangle = -\mathbb{E}\left[\left\langle \frac{\nabla h(Y)}{h(Y)}, f(Y) \right\rangle + \operatorname{tr} \left. \frac{\partial f(y)}{\partial y} \right|_Y \right].$$

Lemma 2, whose proof can be found in [14], provides an estimator of the dot product $\mathbb{E}\langle \theta, f(Y) \rangle$ that solely depends on Y without reference to θ. As a consequence, the Generalized SURE (as coined by [14]) defined by

$$\mathrm{GSURE} = \|\hat{\theta}(Y)\|^2 + 2 \left\langle \frac{\nabla h(Y)}{h(Y)}, \hat{\theta}(Y) \right\rangle + 2 \operatorname{tr} \left. \frac{\partial \hat{\theta}(y)}{\partial y} \right|_Y + \frac{1}{h(Y)} \operatorname{tr} \left. \frac{\partial^2 h(y)}{\partial y^2} \right|_Y \tag{3.1}$$

is an unbiased estimator of MSE_θ, i.e., $\mathbb{E}\mathrm{GSURE} = \mathrm{MSE}_\theta$, provided $\hat{\theta}$, h and ∇h are weakly differentiable[2]. Note that omitting the last term in (3.1) leads to the seminal definition of GSURE given in [14] which provides an unbiased estimate of $\mathrm{MSE}_\theta - \|\theta\|^2$, even though ∇h is not weakly differentiable.

The GSURE can be specified for Gaussian noise, and in this case $\mathrm{GSURE} = \sigma^{-4}\mathrm{SURE}$ and the "natural" risk boils down to the risk as $\mathrm{MSE}_\theta = \sigma^{-4}\mathrm{MSE}_\mu$. In general, such a linear relationship between the "natural" risk and the risk of interest might not be met. For instance, under Gamma noise[3] with scale parameter L (see Table 1), with expectation μ and independent entries, the GSURE reads as

$$\underset{\mathrm{Gamma}}{\mathrm{GSURE}} = \sum_{i=1}^d \frac{L^2}{\hat{\mu}_i(Y)^2} - \frac{2L(L-1)}{Y_i\hat{\mu}_i(Y)} + \frac{2L}{\hat{\mu}_i(Y)^2} \left. \frac{\partial \hat{\mu}_i(y)}{\partial y_i} \right|_Y + \frac{(L-1)(L-2)}{Y_i^2} \tag{3.2}$$

which, as soon as $L > 2$ and $\hat{\mu}$ fulfills the assumptions of Lemma 2, unbiasedly estimates $\mathrm{MSE}_\theta = L^2\mathbb{E}\|\mu^{-1} - \hat{\mu}(Y)^{-1}\|^2$, where $(\cdot)^{-1}$ is the entry-wise inversion[4]. We will see in practice that minima of MSE_θ can strongly depart from those of interest. As the GSURE can only measure discrepancy in the "natural" parameter space, its applicability in real scenarios can thus be seriously limited.

---

[2]Eq. (3.1) is obtained by applying Lemma 2 on $\left\langle \theta, \hat{\theta}(Y) \right\rangle$, $\langle \theta, \theta \rangle$ and $\langle h(Y)^{-1}\nabla h(Y), \theta \rangle$.

[3]A random variable Y follows a Gamma distribution with scale parameter L if it results from the mean of L independent and identically distributed exponential random variables. For this reason, L is often referred to as the number of looks and controls the spread of the distribution as $\mathrm{Var}[Y] = \Lambda(\mu) = \frac{\mu^2}{L}$. This distribution is widely used to describe fluctuations of speckle in coherent laser imagery [21].

[4]$L > 2$ implies that h and ∇h are weakly differentiable. By omitting the last term of GSURE, an unbiased estimate of $L^2\mathbb{E}\|\mu^{-1} - \hat{\mu}(Y)^{-1}\|^2 - L^2\|\mu^{-1}\|^2$ is obtained as soon as $L > 1$.

Discrete exponential family. We now consider discrete noises distributed within the natural exponential family, e.g., Poisson or binomial. Before turning to the general result, let us focus on Poisson noise with mean μ and independent entries for which the Poisson unbiased risk estimator (PURE) defined as

$$\text{PURE} = \|\hat{\mu}(Y)\|^2 - 2\langle Y, \hat{\mu}_\downarrow(Y)\rangle + \langle Y, Y-1\rangle \quad \text{where} \quad \hat{\mu}_\downarrow(Y)_i = \hat{\mu}_i(Y-e_i), \tag{3.3}$$

unbiasedly estimates MSE_μ, see, e.g., [7, 26]. The vector e_i is defined as $(e_i)_i = 1$ and $(e_i)_j = 0$ for $j \neq i$. The PURE is in fact the consequence of the following lemma also due to Hudson [26].

Lemma 3 (Hudson's lemma). *Assume Y follows a discrete distribution on $\mathbb{Z}^d$ of the natural exponential family with natural parameter θ, then*

$$\mathbb{E}\langle \exp\theta, f(Y)\rangle = \mathbb{E}\left[\left\langle \frac{h_\downarrow(Y)}{h(Y)}, f_\downarrow(Y)\right\rangle\right] \quad \text{where} \quad h_\downarrow(Y)_i = h(Y - e_i)$$

holds for every mapping $f : \mathbb{Z}^d \to \mathbb{R}$ where $\exp$ is the entry-wise exponential.

Hudson's lemma provides an estimator of the dot product $\mathbb{E}\langle \exp\theta, f(Y)\rangle$ that solely depends on Y without reference to the parameter $\eta = \exp\theta$. As a consequence, we can define a Generalized PURE (GPURE) as

$$\text{GPURE} = \|\exp\hat{\theta}(Y)\|^2 - 2\left\langle \frac{h_\downarrow(Y)}{h(Y)}, \exp\hat{\theta}_\downarrow(Y)\right\rangle + \left\langle \frac{h_\downarrow(Y)}{h(Y)}, \left(\frac{h_\downarrow}{h}\right)_\downarrow(Y)\right\rangle \tag{3.4}$$

which unbiasedly estimates MSE_η for the discrete natural exponential family[5].

As for GSURE, GPURE cannot in general measure discrepancy in the parameter space of interest, and for this reason, its applicability in real scenarios can also be limited. However, under Poisson noise, the "exp-natural" space coincides with the parameter space of interest as $\eta = \exp(\phi(\mu)) = \mu$, hence, leading to the PURE. Another interesting case, already investigated in [26], is the one of noise with a negative binomial distribution with mean μ and independent entries, for which the "exp-natural" space does not match with the one of μ but with the one of the underlying probability vector $\mathtt{p} \in [0,1]^d$ as defined in Table 1 (we have $\theta_i = \log \mathtt{p}_i$). In such a case, GPURE reads, for $r \in \mathbb{R}_*^+/\{1,2\}$, as

$$\text{GPURE}_{\text{negbin}} = \|\hat{\mathtt{p}}(Y)\|^2 - 2\sum_{i=1}^d \frac{Y_i \hat{\mathtt{p}}_i(Y_i - 1)}{Y_i + r - 1} + \sum_{i=1}^d \frac{Y_i(Y_i - 1)}{(Y_i + r - 1)(Y_i + r - 2)} \tag{3.5}$$

and is an unbiased estimator of $\mathbb{E}\left[\|\hat{\mathtt{p}}(Y) - \mathtt{p}\|^2\right]$.

---

[5] Eq. (3.4) is obtained by applying three times Lemma 3.

Other related works. It is worth mentioning that there have been several works focusing on estimating mean squared errors in other scenarios. For instance, when Y has an elliptical-contoured distribution with a finite known covariance matrix Σ, the works of [28, 23] provide a generalization of Stein's lemma that can also be used to estimate the risk associated to μ. In [40], the authors provide a versatile approach that provides unbiased risk estimators in many cases, including, all members of the exponential family (continuous or discrete), the Cauchy distribution, the Laplace distribution, and the uniform distribution [40]. The authors of [33] use a similar approach to design such an estimator in the case of the non-centered χ^2 distribution [33].

4. Kullback-Leibler loss estimation for the exponential family

We now turn to our first contribution that provides, for continuous distributions of the natural exponential family, an unbiased estimator of the Kullback-Leibler synthesis loss.

Theorem 4.1 (Stein Unbiased KLS estimator). *Assume $y \mapsto \hat{\mu}(y)$ is weakly differentiable with essentially bounded weak partial derivatives on $\mathbb{R}^d$ and Y follows a distribution of the natural exponential family with natural parameter θ, provided h is weakly differentiable on $\mathbb{R}^d$, the following*

$$\text{SUKLS} = \left\langle \hat{\theta}(Y) + \frac{\nabla h(Y)}{h(Y)}, \, \hat{\mu}(Y) \right\rangle + \text{tr} \left. \frac{\partial \hat{\mu}(y)}{\partial y} \right|_Y - A(\hat{\theta}(Y))$$

where $\hat{\theta}(Y) = \phi(\hat{\mu}(Y))$, is an unbiased estimator of $\text{MKLS} - A(\theta)$.

Proof. Remark that $\text{MKLS} = \mathbb{E}\left[\left\langle \hat{\theta}(Y) - \theta, \, \hat{\mu}(Y) \right\rangle - A(\hat{\theta}(Y)) \right] + A(\theta)$ since $\nabla A(\hat{\theta}(Y)) = \hat{\mu}(Y)$. Hence, Lemma 2 leads to

$$\mathbb{E}\left[\langle \theta, \, \hat{\mu}(Y) \rangle \right] = -\mathbb{E}\left[\left\langle \frac{\nabla h(Y)}{h(Y)}, \, \hat{\mu}(Y) \right\rangle + \text{tr} \left. \frac{\partial \hat{\mu}(y)}{\partial y} \right|_Y \right], \tag{4.1}$$

which concludes the proof. $\square$

As GSURE, SUKLS can be specified for Gaussian noise, and in this case $\text{SUKLS} = (2\sigma^2)^{-1}(\text{SURE} - \|Y\|^2 + d\sigma^2)$ and the Kullback-Leibler synthesis loss boils down to the risk as $\text{MKLS} = (2\sigma^2)^{-1}\text{MSE}_\mu$. More interestingly, consider the following example of Gamma noise.

Example 1. *Under Gamma noise with expectation μ, shape parameter L (as defined in Table 1) and independent entries, SUKLS reads as*

$$\underset{Gamma}{\text{SUKLS}} = \sum_{i=1}^{d} \left[\frac{(L-1)\hat{\mu}(Y)_i}{Y_i} - L\log(\hat{\mu}(Y)_i) - L \right] + \text{tr} \left. \frac{\partial \hat{\mu}(y)}{\partial y} \right|_Y \tag{4.2}$$

which, up to a constant, and provided $L > 1$, unbiasedly estimates

$$\underset{Gamma}{\mathrm{MKLS}} = \sum_{i=1}^{d} \mathbb{E}\left[L\frac{\hat{\mu}(Y)_i}{\mu_i} - L\log\left(\frac{\hat{\mu}(Y)_i}{\mu_i}\right) - L\right]. \qquad (4.3)$$

In our experiments, we will see that minimizing MKLS (or its SUKLS estimate) leads to relevant selections, unlike minimizing MSE_θ (or its GSURE estimate). Note that the authors of [24] have proposed a consistent estimator of MKLS when $L = 1$ (they did not study the case where $L > 1$), their estimator has been however designed only for kernel smoothing problems.

Theorem 4.1 is a straightforward application of Lemma 2 that applies since $\mathrm{MKLS} - A(\theta)$ depends only on θ through a dot product $\langle \theta, f(Y) \rangle$ for some mappings f. For discrete distributions, Lemma 3 only provides an estimate of $\langle \exp(\theta), f(Y) \rangle$ and hence cannot be applied to estimate MKLS. Alternatively, we can focus on estimating the Kullback-Leibler analysis loss MKLA. To this end, a formula that provides an estimate of $\langle \nabla A(\theta), f(Y) \rangle$ for some mappings f is needed. Of course, if $\nabla A(\theta) = \theta$ for some continuous distributions, Lemma 2 applies and can be used to design an estimator of MKLA. However, the only distribution satisfying $\nabla A(\theta) = \theta$ is the normal distribution, for which SURE can already be used to estimate $\mathrm{MKLA} = (2\sigma^2)^{-1}\mathrm{MSE}_\mu$. More interestingly, if $\nabla A(\theta) = \exp(\theta)$ for some discrete distributions, Lemma 3 applies and can be used to design an unbiased estimator of MKLA. The Poisson distribution satisfies this relation leading us to state the following theorem.

Theorem 4.2 (Poisson Unbiased KLA estimator). *Assume Y follows a Poisson distribution with expectation μ and independent entries, then*

$$\mathrm{PUKLA} = \|\hat{\mu}(Y)\|_1 - \langle Y,\, \log \hat{\mu}_\downarrow(Y) \rangle,$$

is an unbiased estimator of $\underset{Poisson}{\mathrm{MKLA}} + \|\mu\|_1 - \langle \mu,\, \log \mu \rangle$ *where*

$$\underset{Poisson}{\mathrm{MKLA}} = \mathbb{E}\left[\|\hat{\mu}(Y)\|_1 - \langle \mu,\, \log \hat{\mu}(Y) - \log \mu \rangle\right] - \|\mu\|_1$$

and $\log$ *is the entry-wise logarithm.*

Proof. The expression of MKLA follows directly from Table 1 and Equation (1.2) since $\exp\theta = \mu$. From Lemma 3, we get

$$\mathbb{E}\left[\langle \mu,\, \log(\hat{\mu}(Y))\rangle\right] = \mathbb{E}\left[\left\langle \exp\theta,\, \hat{\theta}(Y)\right\rangle\right] = \mathbb{E}\left[\left\langle \frac{h_\downarrow(Y)}{h(Y)},\, \hat{\theta}_\downarrow(Y)\right\rangle\right], \qquad (4.4)$$

which concludes the proof as $h_\downarrow(y)/h(y) = y$ and $\hat{\theta}_\downarrow(Y) = \log\hat{\mu}_\downarrow(Y)$. $\qquad \square$

With such results at hand, only the Poisson distribution admits an unbiased estimator of the mean Kullback-Leibler analysis loss. In order to design an estimator of MKLA for a larger class of natural exponential distributions, we will make use of the following proposition.

Proposition 1. *For any probability density or mass function $y \mapsto p(y; \theta)$ of the natural exponential family of parameter θ, the Kullback-Leibler analysis loss associated to $y \mapsto \hat{\theta}(y)$ can be decomposed as follows*

$$\text{MKLA} = \underbrace{-\mathbb{E} \log \frac{p(Y; \hat{\theta}(Y))}{p(Y; \theta)}}_{expected\ data\ fidelity\ loss} + \underbrace{\text{tr Cov}\left(\hat{\theta}(Y), Y\right)}_{model\ complexity},$$

where $\quad - \mathbb{E} \log \frac{p(Y; \hat{\theta}(Y))}{p(Y; \theta)} = \mathbb{E}\left[A(\hat{\theta}(Y)) - A(\theta) - \left\langle Y, \hat{\theta}(Y) - \theta \right\rangle\right]$

and $\quad \text{tr Cov}\left(\hat{\theta}(Y), Y\right) = \mathbb{E}\left[\left\langle Y - \mu, \hat{\theta}(Y) \right\rangle\right]$.

Proof. Subtracting and adding $\left\langle Y, \hat{\theta}(Y) - \theta \right\rangle$ in the MKLA definition leads to

$$\text{MKLA} = \mathbb{E}\left[A(\hat{\theta}(Y)) - A(\theta) - \left\langle Y, \hat{\theta}(Y) - \theta \right\rangle + \left\langle Y - \nabla A(\theta), \hat{\theta}(Y) - \theta \right\rangle\right].$$

As $-\log p(Y; \theta) = -\log h(Y) - \langle Y, \theta \rangle + A(\theta)$ and $\nabla A(\theta) = \mu = \mathbb{E}[Y]$, this concludes the proof. □

In the same vein as for the decomposition (2.1), Proposition 1 provides a variational interpretation of the minimization of MKLA, valid for noise distributions within the exponential family. Minimizing MKLA leads to a maximum *a posteriori* selection promoting faithful models with low complexity. It boils down to (2.1) when specified for Gaussian noise. As for the MSE, the fidelity term can always be unbiasedly estimated, up to an additive constant, without knowledge of θ. Only the complexity term $\text{tr Cov}(\hat{\theta}(Y), Y)$, which generalizes the notion of degrees of freedom, is required to be estimated. Except for the Poisson distribution, none of the previous lemmas can be applied to unbiasedly estimate this term. However, we will show that it can be biasedly estimated, with vanishing bias depending on both the "smoothness" of $\hat{\theta}$ and the behavior of the moments of Y. Towards this goal, let us first recall the Delta method.

Lemma 4 (Delta method)**.** *Let $Y_n = \frac{1}{n}(Z_1 + \ldots + Z_n)$, $n \geq 1$, where $Z_1, Z_2, \ldots$ is an infinite sequence of independent and identically distributed random vectors in $\mathbb{R}^d$ with $\mathbb{E}Z_i = \mu$, $\text{Var}[Z_i] = \Sigma$ and finite moments up to order $k \geq 3$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be k times totally differentiable with bounded k-th derivative, then*

$$\mathbb{E}\left[f(Y_n) - f(\mu)\right] = \frac{1}{2n} \text{tr}\left(\Sigma \left.\frac{\partial^2 f(y)}{\partial y^2}\right|_\mu\right) + O(n^{-2}) = O(n^{-1}).$$

Lemma 4 is a direct d-dimensional extension of [29] (Theorem 5.1a, page 109), that allows us to introduce our biased estimator of MKLA.

Theorem 4.3 (Delta KLA estimator)**.** *Let $Y_n = \frac{1}{n}(Z_1 + \ldots + Z_n)$, $n \geq 1$, where $Z_1, Z_2, \ldots$ is an infinite sequence of independent random vectors in $\mathbb{R}^d$*

identically distributed within the natural exponential family with natural parameter θ, log-partition function A, expectation μ, variance function Λ and finite moments up to order $k \geq 3$. As a result, the distribution of Y_n is also in the natural exponential family parametrized by $\theta_n = n\theta$ with log-partition function $A_n(\theta_n) = nA(\theta_n/n)$, expectation μ and variance function $\Lambda_n = \Lambda/n$. Provided $\hat{\theta}_n$ reads as $\hat{\theta}_n = n\hat{\theta}$, and $\hat{\theta} : \mathbb{R}^d \to \mathbb{R}^d$ is k times totally differentiable with bounded k-th derivative, then

$$
\mathrm{DKLA}_n = A_n(\hat{\theta}_n(Y_n)) - \left\langle Y_n, \hat{\theta}_n(Y_n) \right\rangle + \mathrm{tr}\left(\Lambda_n(Y_n) \left. \frac{\partial \hat{\theta}_n(y)}{\partial y} \right|_{Y_n} \right)
$$

satisfies $\quad \mathbb{E}\mathrm{DKLA}_n = \mathrm{MKLA}_n - \langle \mu, \theta_n \rangle + A_n(\theta_n) + O(n^{-1})$

where MKLA_n is the KL analysis loss associated to $\hat{\theta}_n$ with respect to θ_n.

Proof. Let $f(y) = \left\langle \hat{\theta}(y), y - \mu \right\rangle$. We have $f(\mu) = 0$ and $\left. \frac{\partial^2 f(y)}{\partial y^2} \right|_\mu = 2 \left. \frac{\partial \hat{\theta}(y)}{\partial y} \right|_\mu$. Under the assumptions on $\hat{\theta}$, the second-order approximation of Lemma 4 applies

$$
\mathrm{tr}\,\mathrm{Cov}(\hat{\theta}(Y_n), Y_n) \triangleq \mathbb{E}\left[f(Y_n) - f(\mu) \right] = \frac{1}{n}\,\mathrm{tr}\left(\Lambda(\mu) \left. \frac{\partial \hat{\theta}(y)}{\partial y} \right|_\mu \right) + O(n^{-2})\ . \tag{4.5}
$$

Moreover, under the assumptions on $\hat{\theta}$ and as Λ is in $\mathcal{C}^\infty$, the first-order approximation of Lemma 4 applies and

$$
\mathbb{E}\left[\mathrm{tr}\left(\Lambda(Y_n) \left. \frac{\partial \hat{\theta}(y)}{\partial y} \right|_{Y_n} \right) \right] = \mathrm{tr}\left(\Lambda(\mu) \left. \frac{\partial \hat{\theta}(y)}{\partial y} \right|_\mu \right) + O(n^{-1})\ . \tag{4.6}
$$

Subsequently, we have

$$
\mathbb{E}\mathrm{DKLA}_n - \mathrm{MKLA}_n + \langle \nabla A_n(\theta_n), \theta_n \rangle - A_n(\theta_n)
$$

$$
= \mathbb{E}\left[\mathrm{tr}\left(\Lambda_n(Y_n) \left. \frac{\partial \hat{\theta}_n(y)}{\partial y} \right|_{Y_n} \right) - \mathrm{tr}\,\mathrm{Cov}\left(\hat{\theta}_n(Y_n), Y_n \right) \right] \tag{4.7}
$$

$$
= n\mathbb{E}\left[\frac{1}{n}\,\mathrm{tr}\left(\Lambda(Y_n) \left. \frac{\partial \hat{\theta}(y)}{\partial y} \right|_{Y_n} \right) - \mathrm{tr}\,\mathrm{Cov}\left(\hat{\theta}(Y_n), Y_n \right) \right] = O(n^{-1}) \tag{4.8}
$$

and as $\nabla A_n(\theta_n) = \mu$, this concludes the proof. $\qquad\square$

It is worth mentioning that Theorem 4.3 can be applied to Gaussian noise, with DKLA boiling down to SURE, as $\mathrm{DKLA} = (2\sigma^2)^{-1}(\mathrm{SURE} - \|Y\|^2 + d\sigma^2)$. However, the conclusion is not as strong, as by virtue of Lemma 1, DKLA would be in fact an unbiased estimator provided only that $\hat{\mu}$ is weakly differentiable. More interestingly, consider the two following examples.

Example 2. *Gamma random vectors Y_n with expectation $\mu \in (\mathbb{R}_*^+)^d$ and shape parameter $L_n = n$ (as defined in Table 1) results from the sample mean of n independent exponential random vectors with expectation μ (entries of the vectors are supposed to be independent). As exponential random vectors have finite moments, provided $\hat{\mu}$ is sufficiently smooth and since ϕ is continuously differentiable in $(\mathbb{R}_*^+)^d$, Theorem 4.3 applies and we get*

$$\underset{Gamma}{\mathrm{DKLA}_n} = \sum_{i=1}^{d} -L_n \log \hat{\mu}_i(Y_n) + \frac{L_n(Y_n)_i}{\hat{\mu}_i(Y_n)} + \frac{(Y_n)_i^2}{\hat{\mu}_i(Y_n)^2} \left. \frac{\partial \hat{\mu}_i(y)}{\partial y_i} \right|_{Y_n}$$

satisfies $\quad \underset{Gamma}{\mathbb{E}\mathrm{DKLA}_n} = \underset{Gamma}{\mathrm{MKLA}_n} + L_n \sum_{i=1}^{d} \log(\mu_i) - L_n + O(n^{-1})$ (4.9)

where $\quad \underset{Gamma}{\mathrm{MKLA}_n} = L_n \sum_{i=1}^{d} \mathbb{E}\left[-\log(\hat{\mu}_i(Y_n)) + \frac{\mu_i}{\hat{\mu}_i(Y_n)} + \log(\mu_i) - 1 \right].$

Example 3. *Consider Y_n the sample mean of n independent Poisson random vectors with expectation $\mu \in (\mathbb{R}_*^+)^d$. We have that Y_n, for all n, belongs to the natural exponential family with $A_n(\theta_n) = n \exp(\theta_n/n)$ and $\theta_n = n \log \mu$ (entries of the vectors are supposed to be independent). As Poisson random vectors have finite moments, provided $\hat{\mu}$ is sufficiently smooth and since ϕ is continuously differentiable in $(\mathbb{R}_*^+)^d$, Theorem 4.3 applies and we get*

$$\underset{Poisson}{\mathrm{DKLA}_n} = n\|\hat{\mu}(Y_n)\|_1 - \left\langle Y_n,\, n \log \hat{\mu}(Y_n) + \mathrm{diag}\left(\left. \frac{\partial \log \hat{\mu}(y)}{\partial y} \right|_{Y_n} \right) \right\rangle$$

satisfies $\quad \underset{Poisson}{\mathbb{E}\mathrm{DKLA}_n} = \underset{Poisson}{\mathrm{MKLA}_n} - n \langle \mu, \log \mu \rangle + n\|\mu\|_1 + O(n^{-1})$ (4.10)

where $\quad \underset{Poisson}{\mathrm{MKLA}_n} = n\mathbb{E}\left[\|\hat{\mu}(Y_n)\|_1 - \langle \mu, \log \hat{\mu}(Y_n) - \log \mu \rangle - \|\mu\|_1 \right].$

Interestingly, remark that $\mathrm{PUKLA}(\hat{\mu}, Y) \approx \mathrm{DKLA}(\hat{\mu}, Y)$, as soon as we have both $\hat{\mu}(Y - 1) \approx \hat{\mu}(Y) - \hat{\mu}'(Y)$ and $|\hat{\mu}(Y)| \gg |\hat{\mu}'(Y)|$.

5. Reliability study

In this section, we aim at studying and comparing the sensitivity of the previously studied risk estimators. Little is known about the variance of SURE: $\mathrm{Var}[\mathrm{SURE}] = \mathbb{E}\left[(\mathrm{SURE} - \mathrm{MSE})^2 \right]$. It is in general an intricate problem and some studies [37, 31] focus instead on the reliability $\mathbb{E}\left[(\mathrm{SURE} - \mathrm{SE})^2 \right]$ where $\mathrm{SE} = \|\mu - \hat{\mu}(Y)\|^2$ (note that $\mathrm{MSE} = \mathbb{E}[\mathrm{SE}]$). Here, we do not aim at providing tight bounds on the reliability as this would require specific extra assumptions for each pair of loss functions and estimators. The next proposition provides only crude bounds on the reliability of each estimator.

Proposition 2. *Assume $y \mapsto \hat{\theta}(y)$ is weakly differentiable. Then, provided the following quantities are finite, we have*

$$\frac{1}{2}\mathbb{E}\left[\left(\overline{\mathrm{GSURE}} - \overline{\mathrm{SE}}_\theta\right)^2\right]^{1/2} \le \mathbb{E}\left[\left\langle \frac{\nabla h(Y)}{h(Y)} + \theta, \hat{\theta}(Y)\right\rangle^2\right]^{1/2} + \mathbb{E}\left[\left(\mathrm{tr}\left.\frac{\partial\hat{\theta}(y)}{\partial y}\right|_Y\right)^2\right]^{1/2}$$

$$\mathbb{E}\left[\left(\overline{\mathrm{SUKLS}} - \overline{\mathrm{KLS}}\right)^2\right]^{1/2} \le \mathbb{E}\left[\left\langle \frac{\nabla h(Y)}{h(Y)} + \theta, \hat{\mu}(Y)\right\rangle^2\right]^{1/2} + \mathbb{E}\left[\left(\mathrm{tr}\left.\frac{\partial\hat{\mu}(y)}{\partial y}\right|_Y\right)^2\right]^{1/2}$$

$$\frac{1}{2}\mathbb{E}\left[\left(\overline{\mathrm{PURE}} - \overline{\mathrm{SE}}_\mu\right)^2\right]^{1/2} \le \mathbb{E}\left[\langle\mu, \hat{\mu}(Y)\rangle^2\right]^{1/2} \qquad + \mathbb{E}\left[\langle Y, \hat{\mu}_\downarrow(Y)\rangle^2\right]^{1/2}$$

$$\mathbb{E}\left[\left(\overline{\mathrm{PUKLA}} - \overline{\mathrm{KLA}}\right)^2\right]^{1/2} \le \mathbb{E}\left[\langle\mu, \log\hat{\mu}(Y)\rangle^2\right]^{1/2} \qquad + \mathbb{E}\left[\langle Y, \log\hat{\mu}_\downarrow(Y)\rangle^2\right]^{1/2}$$

$$\mathbb{E}\left[\left(\overline{\mathrm{DKLA}} - \overline{\mathrm{KLA}}\right)^2\right]^{1/2} \le \mathbb{E}\left[\left\langle Y - \mu, \hat{\theta}(Y)\right\rangle^2\right]^{1/2} \qquad + \mathbb{E}\left[\mathrm{tr}\left(\Lambda(Y)\left.\frac{\partial\hat{\theta}(y)}{\partial y}\right|_Y\right)^2\right]^{1/2}$$

where $\mathrm{KLA} = \mathcal{D}(P_\theta\|P_{\hat{\theta}(Y)})$ *(note that $\mathbb{E}[\mathrm{KLA}] = \mathrm{MKLA}$), and* KLS *is defined similarly. The over-line refers to quantities for which additive constant with respect to $\hat{\mu}(Y)$ are skipped, e.g.,* $\overline{\mathrm{SE}}_\mu = \mathrm{SE}_\mu - \|\mu\|^2 = \|\hat{\mu}(Y)\|^2 - 2\langle\hat{\mu}(Y), \mu\rangle$ *and* $\overline{\mathrm{KLA}} = \mathrm{KLA} + A(\theta) - \langle\nabla A(\theta), \theta\rangle$.

Proof. This is a straightforward consequence of Cauchy-Schwartz's inequality. $\qquad\square$

Proposition 2 allows us to compare the relative sensitivities of the different estimators. Comparing GSURE and SUKLS, one can notice that the bounds are similar but the first one is controlled by $\hat{\theta}(Y)$ while the second one is controlled by $\hat{\mu}(Y)$. While it is difficult to make a general statement, we believe SUKLS estimates might be more stable than GSURE since $\hat{\mu}(Y)$ has usually better control than $\hat{\theta}(Y)$, given the non-linearity of the canonical link function ϕ.

6. Implementation details for the proposed estimators

In this section, we explain how the proposed risk estimators can be evaluated in practice within a reasonable computation time.

All risk estimators designed for continuous distributions rely on the computation of $\mathrm{tr}\left[g(y)\left.\frac{\partial f(y)}{\partial y}\right|_y\right]$ for some mappings $g : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ and $f : \mathbb{R}^d \to \mathbb{R}^d$. For instance, SURE requires to compute such a a quantity with $g(y) = \mathrm{Id}_d$ and $f = \hat{\mu}$ (see eq. (2.2)). In general, the computation of these terms requires at least $O(d^2)$ operations and thus prevents the use of such risk estimators in practice. Fortunately, following [19, 38], we can approximate such terms by using Monte-Carlo simulations, thanks to the following relation

$$\mathrm{tr}\left[g(y)\left.\frac{\partial f(y)}{\partial y}\right|_y\right] = \mathbb{E}\left\langle\zeta, g(y)\left.\frac{\partial f(y)}{\partial y}\right|_y\zeta\right\rangle \quad \text{for} \quad \zeta \sim \mathcal{N}(0, \mathrm{Id}_d), \qquad (6.1)$$

where the directional derivatives in the direction $\zeta \in \mathbb{R}^d$ can be computed by using finite differences or algorithm differentiations as described in [10]. This leads in general to a much faster evaluation in $O(d)$ operations.

In the Poisson setting, risk estimators rely on the computation of $\langle y, f_{\downarrow}(y) \rangle$ for some mapping $f : \mathbb{R}^d \to \mathbb{R}^d$. For instance, PUKLA requires to compute such a quantity with $f = -\log \hat{\mu}$ (see Theorem (4.2)). Again, the computation of such terms requires at least $O(d^2)$ operations in general. Based on first order expansions, we have empirically chosen to perform Monte-Carlo simulations on the following approximation

$$\langle y, f_{\downarrow}(y) \rangle \approx \left\langle y, f(y) - \text{diag}\left(\left(\left. \frac{\partial f(y)}{\partial y} \right|_y \zeta \right) \zeta^t \right) \right\rangle, \tag{6.2}$$

where $\zeta \in \{-1, +1\}^d$ is Bernoulli distributed with $\mathtt{p} = 0.5$. In our numerical experiments, this approximation led to $O(d)$ operations and satisfactory results even though f was chosen to be non-linear. This approximation clearly deserves more attention but is considered here to be beyond the scope of this study.

7. Numerical experiments

In this section, we will perform numerical experiments showing the interest of the proposed Kullback-Leibler risk estimators in two different applications.

7.1. *Application to image denoising*

We first consider that Y and μ are d dimensional vectors representing images on a discrete grid of d pixels, such that entries with index i are located at pixel location $\delta_i \in \Delta \subset \mathbb{Z}^2$. A realization y of Y represents a noisy observation of the image μ. The estimate $\hat{\mu}$ of μ is a denoised version of y.

Performance evaluation. In order to evaluate the proposed loss functions and their estimates, visual inspection will be considered to assess the image quality in terms of noise variance reduction and image content preservation. In order to provide an objective measure of performance, taking into account heteroscedasticity and tails of the noise, we will evaluate the mean normalized absolute deviation error defined as $\text{MNAE} = d^{-1}\sqrt{\pi/2}\|\Lambda(\mu)^{-1/2}(\mu - \hat{\mu}(Y))\|_1$. The MNAE measures to which extent $\hat{\mu}(Y)$ might belong to a confident interval around μ with dispersion related to $\Lambda(\mu)$. The MNAE is expected to be 1 when $\hat{\mu}(Y) \sim \mathcal{N}(\mu, \Lambda(\mu))$, and should get closer to 0 when $\hat{\mu}(Y)$ improves on Y itself.

Simulations in linear filtering. We consider here that $\hat{\mu}$ is the linear filter

$$\hat{\mu}(y) = Wy \quad \text{with} \quad W_{i,j} = \frac{\exp(-\|\delta_i - \delta_j\|^2/\tau^2)}{\sum_j \exp(-\|\delta_i - \delta_j\|^2/\tau^2)}, \tag{7.1}$$

(a) Original image μ (b) Noisy image y

(c) MNAE = 0.972   (d) MNAE = 0.496   (e) MNAE = 0.500

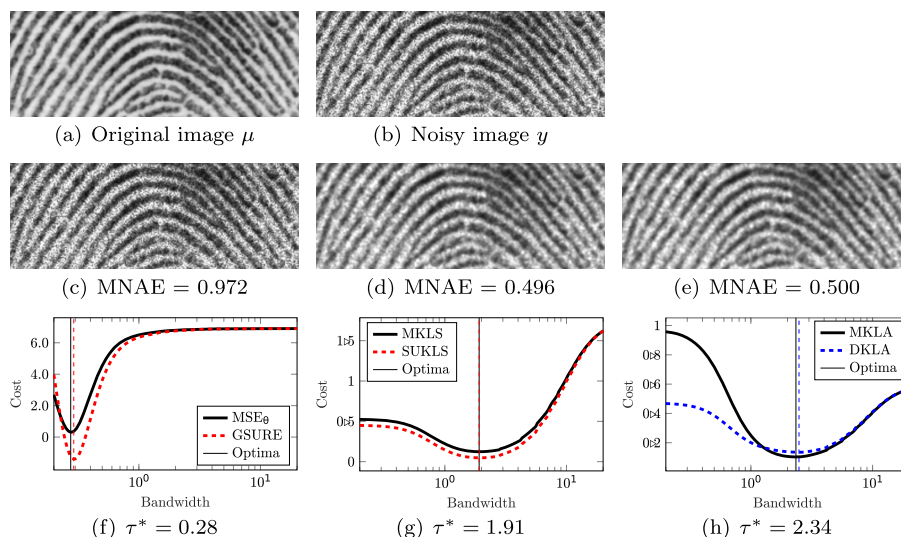(f) $\tau^* = 0.28$ (g) $\tau^* = 1.91$ (h) $\tau^* = 2.34$

FIG 1. *(a, b) Original and noisy image contaminated by Gamma noise with $L = 3$ (square-root of the images are displayed for better visual inspection). (c, d, e) Results of linear filtering for the optimal bandwidth with respect to the natural risk MSE_θ MKLS and MKLA. The MNAE is indicated. (f,g,h) Risks and their estimates as a function of the bandwidth. The optimal bandwidth $\tau^\star$ is indicated. Red shows unbiased estimation and blue biased estimation.*

where $W \in \mathbb{R}^{d \times d}$ is a circulant matrix encoding a discrete convolution with a Gaussian kernel of bandwidth $\tau > 0$. In this context, we will evaluate the relevance of the different proposed loss functions and their estimates as objectives to select a bandwidth τ offering a satisfying denoising.

Figure 1 gives an example of a noisy observation y of an image μ representing fingerprints whose pixel values are independently corrupted by Gamma noise with shape parameter $L = 3$. We have evaluated the relevance of the natural risk MSE_θ given by $\|\mu^{-1} - \hat{\mu}(Y)^{-1}\|^2$, MKLS and MKLA in selecting the bandwidth τ. Visual inspection of the results obtained at the optimal bandwidth for each criterion shows that the natural risk MSE_θ fails in selecting a relevant bandwidth while MKLS and MKLA both provide a better trade-off. The natural risk strongly penalizes small discrepancies at the lowest intensities while not being sensitive enough for discrepancies at higher intensities. As the noisy image has several isolated pixel values approaching 0, the natural risk will strongly penalize smoothing effects of such isolated structures preventing satisfying noise variance reduction. The Kullback-Leibler loss functions take into account that Gamma noise has a constant signal to noise ratio. Hence, it does not favor the restoration of either bright or dark structures more, allowing satisfying smoothing for both, as assessed by the MNAE. Finally, estimators of these loss functions with respectively GSURE, SUKLS and DKLA are given. Note that for $L = 3$, the Gamma distribution is far from reaching the asymptotic conditions of Theorem 4.3. As a result, bias is not negligible (it becomes obvious for the lowest values of
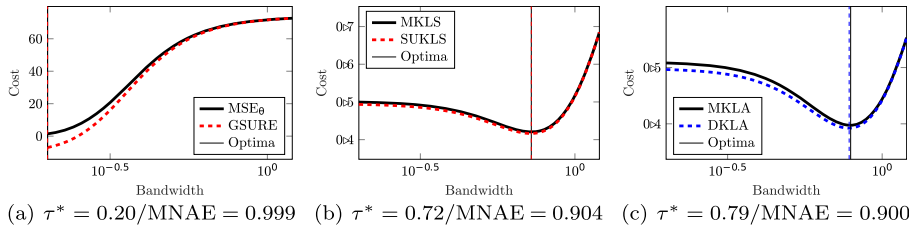
(a) $\tau^* = 0.20/\text{MNAE} = 0.999$ (b) $\tau^* = 0.72/\text{MNAE} = 0.904$ (c) $\tau^* = 0.79/\text{MNAE} = 0.900$

FIG 2. *(a,b,c) Risks and their estimates as a function of the bandwidth in the same setting as in Figure 1 but for Gamma noise with $L = 100$. The optimal bandwidth $\tau^\star$ and the MNAE are indicated. Red shows unbiased estimation and blue biased estimation.*

τ in Figure 1.h). Nevertheless, minimizing DKLA can still provide an accurate location of the optimal parameter for MKLA.

Figure 2 reproduces the same experiment but with Gamma noise with $L = 100$, i.e., with a much better signal to noise ratio. Interestingly, the bias of DKLA gets much smaller than in the previous experiment. This was indeed expected as with $L = 100$, the Gamma distribution fulfills much better the asymptotic conditions of Theorem 4.3. Remark that MNAE values are still in favor of Kullback-Leibler objectives, but the gains are much smaller. In fact, all MNAE values get closer to 1 since noise reduction with signal preservation using linear filtering becomes much harder in such a low signal to noise ratio setting.

Simulations in non-linear filtering. We consider here that $\hat{\mu}$ is the non-local filter [4] defined by

$$\hat{\mu}(y) = W(y)y \quad \text{with} \quad W_{i,j}(y) = \exp\left(-\frac{d(\mathcal{P}_i y, \mathcal{P}_j y)}{\tau}\right) \qquad (7.2)$$

where $\mathcal{P}_i \in \mathbb{R}^{p \times n}$ is a linear operator extracting a patch (a small window of fixed size) at location δ_i, $d : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^+$ is a dissimilarity measure (infinitely differentiable and adapted to the exponential family following [8]) and $\tau > 0$ a bandwidth parameter. Remark as $W(y) \in \mathbb{R}^{d \times d}$ depends on y, $\hat{\mu}(y)$ is non-linear. In this context, we will evaluate again the relevance of the proposed loss functions and their estimates as objectives to select the bandwidth τ.

Figure 3 gives an example of a noisy observation y of an image μ representing a bright two dimensional chirp signal shaded gradually into a dark homogeneous region. The noisy observation y is contaminated by noise following a Gamma distribution with shape parameter $L = 3$. We have again evaluated the relevance of the natural risk MSE_θ given by $\|\mu^{-1} - \hat{\mu}(Y)^{-1}\|^2$, MKLS and MKLA in selecting the bandwidth parameter. Visual inspection of the results obtained at the optimal bandwidth for each criterion shows that the natural risk fails in selecting a relevant bandwidth while MKLS and MKLA both provide more satisfying results. As the image μ is very smooth in the darker region, the natural risk favors strong variance reduction leading to a strong smoothing of

(a) Original image μ (b) Noisy image y

(c) MNAE = 0.853   (d) MNAE = 0.249   (e) MNAE = 0.249

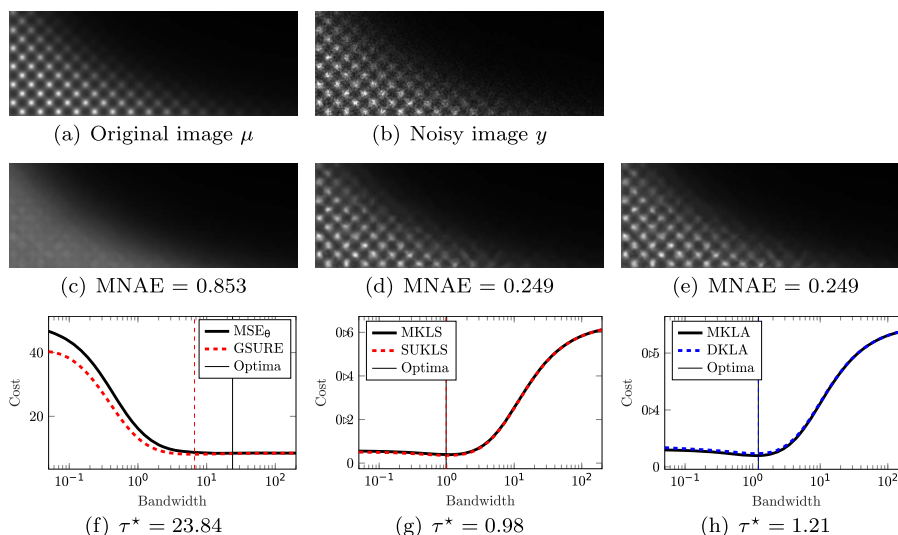(f) $\tau^\star = 23.84$ (g) $\tau^\star = 0.98$ (h) $\tau^\star = 1.21$

FIG 3. *(a, b) Original and noisy image contaminated by Gamma noise with $L = 3$ (squared root versions of the images are displayed for better visual inspection). (c, d, e) Results of non-local filtering for the optimal bandwidth with respect to the natural risk MSE_θ MKLS and MKLA. The MNAE is indicated. (f,g,h) Risks and their estimates as a function of the bandwidth. The optimal bandwidth $\tau^\star$ is indicated. Red shows unbiased estimation and blue biased estimation.*

the texture in the brightest area. Again, the Kullback-Leibler loss functions find a good trade-off preserving simultaneously the bright texture and reducing the noise in the dark homogeneous region, as assessed by the MNAE. Finally, estimators of these loss functions with respectively GSURE, SUKLS and DKLA are given.

Figure 4 gives a similar example where the image μ represents a two dimensional chirp signal shaded gradually into a bright homogeneous region. The image is displayed in log-scale to better assess the variations of the texture in the darkest region. The noisy observation y is corrupted by independent noise following a Poisson distribution. We have evaluated the relevance of the risks MSE_μ, MKLS and MKLA in selecting the bandwidth parameter. Visual inspection of y shows that darker regions are more affected by noise than brighter ones. This is due to the fact that Poisson corruptions lead to a signal to noise ratio evolving as $\sqrt{\mu}$. It follows that the MSE essentially penalizes the residual variance of the brightest region hence leading to a strong smoothing of the texture in the darkest area. Again, Kullback-Leibler losses lead to selecting a more relevant bandwidth, smoothing less the brightest area but preserving better the texture, as assessed by the MNAE. Finally, estimators of the MSE with PURE and MKLA with PUKLA and DKLA are given. Note that estimators of MKLS are not available for non-local filtering under Poisson noise.
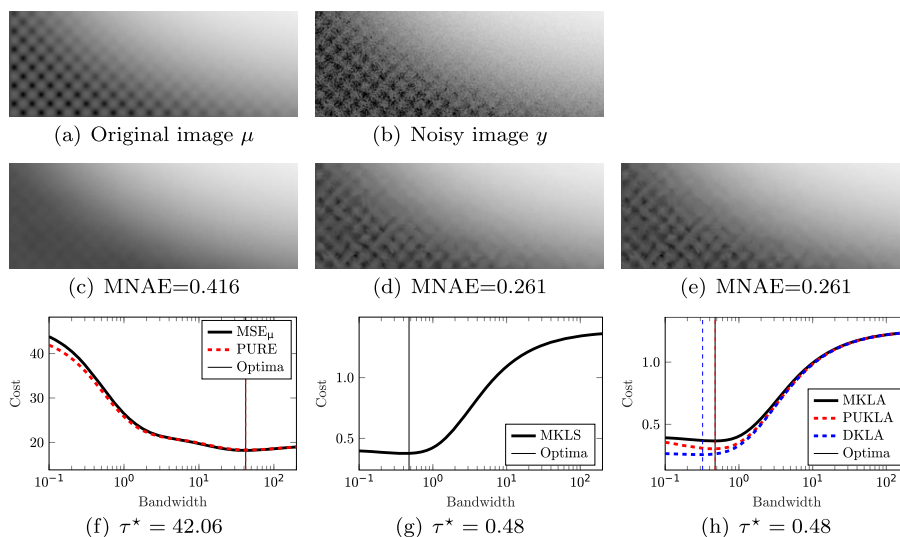
(a) Original image μ

(b) Noisy image y

(c) MNAE=0.416

(d) MNAE=0.261

(e) MNAE=0.261

(f) $\tau^\star = 42.06$

(g) $\tau^\star = 0.48$

(h) $\tau^\star = 0.48$

FIG 4. *(a, b) Original and noisy image contaminated by Poisson noise (log of the images are displayed for better visual inspection). (c, d, e) Results of non-local filtering for the optimal bandwidth with respect to the risk* MSE_μ MKLS *and* MKLA. *The* MNAE *is indicated. (f,g,h) Risks and their estimates (when available) as a function of the bandwidth. The optimal bandwidth* $\tau^\star$ *is indicated. Red shows unbiased estimation and blue biased estimation.*

7.2. Application to variable selection

We now consider the problem of variable selection in linear regression problems, i.e., in finding the non-zero components of a vector $\beta \in \mathbb{R}^q$ that fulfills the assumption that an observed vector $y \in \mathbb{R}^d$ has expectation $\mu = X\beta$ where $X \in \mathbb{R}^{d \times q}$ is the so-called design matrix. To this aim, we consider the Least Absolute Shrinkage and Selection Operator (LASSO) [44] given, for $\lambda > 0$, by

$$\hat{\beta}(y) \in \underset{\beta \in \mathbb{R}^q}{\operatorname{argmin}} \ -\log p(y; \theta = \phi(X\beta)) + \lambda \|\beta\|_1 \ .$$

In this case the predictor $\hat{\mu}$ is given by $\hat{\mu}(y) = X\hat{\beta}(y)$. The LASSO is known to promote sparse solutions, i.e., such that the number of non-zero entries of $\hat{\beta}$ is small compared to q. The level of sparsity is indirectly controlled by the regularization parameter λ, the larger λ is, the sparser $\hat{\beta}$ will be. Finding the optimal parameter λ, and then selecting the relevant variables (columns of X) explaining y, is a challenging problem that can be addressed by minimizing an estimator of the risk. In this context, we will evaluate again the relevance of the different proposed loss functions and their estimates as objectives to select a regularization parameter λ offering a relevant selection of variables.

Figure 5 and Table 3 provide results obtained on such a linear regression problem where X is an orthogonal matrix and $d = q = 16,384$. The vector β was chosen such that 28% of its entries are non-zero. We have generated 200
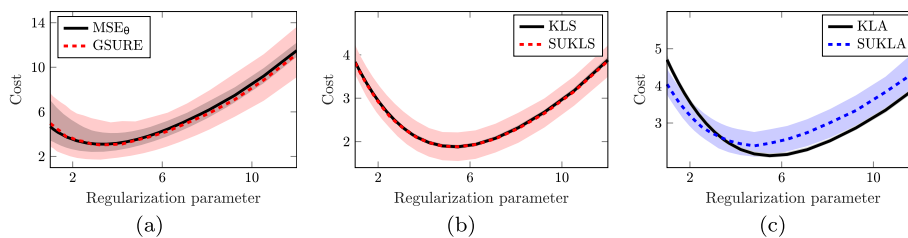
FIG 5. *Risks and their estimates as a function of the regularization parameter λ averaged on 200 realizations (their corresponding 90% confidence intervals are also indicated in shaded colors). Red shows unbiased estimation and blue biased estimation.*

TABLE 3
Performance in terms of errors, false negatives (FN) and false positives (FP) of the LASSO in solving a variable selection problem contaminated with Gamma noise (with scale parameter $L = 8$) and guided by different specific objectives. The noise model is considered correctly specified ($\hat{L} = L$) or misspecified ($|1 - \hat{L}/L| = 0.1$). Results are given in average with plus and minus their standard deviations on 200 noise realizations.

	Errors	FN	FP
Correctly specified			
MSE_θ	34.64 ± 3.09	11.36 ± 1.28	23.27 ± 4.36
GSURE	34.00 ± 4.34	11.74 ± 1.95	22.25 ± 6.26
MKLS	26.24 ± 0.28	15.68 ± 0.16	10.56 ± 0.23
SUKLS	26.84 ± 1.17	15.31 ± 0.82	11.52 ± 1.95
MKLA	26.24 ± 0.28	15.68 ± 0.16	10.56 ± 0.23
DKLA	28.31 ± 1.62	14.37 ± 0.95	13.94 ± 2.54
Misspecified			
GSURE	35.94 ± 4.44	10.93 ± 1.81	25.01 ± 6.22
SUKLS	28.67 ± 1.45	14.13 ± 0.81	14.53 ± 2.22
DKLA	30.20 ± 1.90	13.34 ± 0.96	16.86 ± 2.84

independent realizations y of Y using a Gamma distribution model with scale parameter $L = 8$. We have again evaluated the relevance of the natural risk MSE_θ given by $\|\mu^{-1} - \hat{\mu}(Y)^{-1}\|^2$, MKLS and MKLA in selecting the regularization parameter. Figure 5 shows the evolution of these objectives as a function of λ. It shows that KL objectives lead to selecting a larger λ parameter than with the natural risk. Performance in terms of average percentages of false negatives (FN: $\hat{\beta}_i = 0$ and $\beta_i \neq 0$), false positives (FP: $\hat{\beta}_i = 0$ and $\beta_i \neq 0$) and errors (FP or FN) are reported in Table 3. It shows that tuning the parameter λ with respect to KL objectives leads to lower numbers of errors than with the natural risk. One can observe that the subsequent LASSO estimators work at different trade-offs: KL objectives favor FN over FP, while the natural risk favors FP over FN. Finally, performances with estimators of MSE_θ with GSURE, MKLS with SUKLS, and MKLA with DKLA are also given. It can be observed that risk estimators offer in average comparable results than their oracle counterparts but have higher variance. Note that the LASSO is not differentiable, such that DKLA is not guaranteed to be asymptotically unbiased (as the conditions

of Theorem 3 are not fulfilled), which explains the large discrepancies observed between the results obtained by MKLA and DKLA. Nevertheless, even though DKLA is not asymptotically unbiased in this case variable selections with the LASSO guided by DKLA still provides a good objective for variable selection, with similar results as if it was guided by the oracle MKLA objective.

A last important question is to know whether our risk estimators are robust against model misspecification, i.e., when the generative model (1.1) is only approximately known. Indeed, Lv and Liu [34] demonstrated the advantage of using KL divergence principle for model selection problems in both correctly specified and misspecified models. Along these lines, we have also shown in Table 3 the results obtained under misspecification. We have chosen to evaluate the performance of the LASSO guided by the aforementioned estimators of the risk when the shape parameter L of the Gamma distribution is misestimated by a factor 0.1%: $|1 - \hat{L}/L| = 0.1$. We found that the performance of all estimators drop in this case. Nevertheless, their relative performance are preserved: KL objectives lead to lower numbers of errors than with the natural risk.

8. Conclusion

We addressed the problem of using and estimating Kullback-Leibler losses for model selection in recovery problems involving noise distributed within the exponential family. Our conclusions are threefold: 1) Kullback-Leibler losses have shown empirically to be more relevant than squared losses for model selection in the considered scenarios; 2) Kullback-Leibler losses can be estimated in many cases unbiasedly or with controlled bias depending on the regularity of both the predictor and the noise; 3) Even though the estimation is subject to variance and bias, the subsequent selection has shown empirically to be close to the optimal one associated to the loss being estimated. Future works should focus on understanding under which conditions such a behavior can be guaranteed. This includes establishing tighter bounds on the reliability, consistency with respect to the data dimension d and asymptotic optimality results for some given class of predictors. Estimation of Kullback-Leibler losses and other discrepancies (e.g., Battacharyya, Hellinger, Mahanalobis, Rényi or Wasserstein distances/divergences) beyond the exponential family and requiring less regularity on the predictor should also be investigated.

References

[1] AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* **1** 267–281. Springer Verlag. MR0483125
[2] BLU, T. and LUISIER, F. (2007). The SURE-LET approach to image denoising. *IEEE Trans. Image Process.* **16** 2778–2786. MR2472419
[3] BROWN, L. D. (1986). Fundamentals of statistical exponential families with applications in statistical decision theory. *Lecture Notes–Monograph Series* i–279. MR0882001

[4] Buades, A., Coll, B. and Morel, J. M. (2005). A Review of Image Denoising Algorithms, with a New One. *Multiscale Modeling and Simulation* **4** 490. MR2162865

[5] Cai, T. T. and Zhou, H. H. (2009). A data-driven block thresholding approach to wavelet estimation. *The Annals of Statistics* **37** 569–595. MR2502643

[6] Chaux, C., Duval, L., Benazza-Benyahia, A. and Pesquet, J.-C. (2008). A nonlinear Stein-based estimator for multichannel image denoising. *IEEE Trans. on Signal Processing* **56** 3855–3870. MR2517073

[7] Chen, L. H. Y. (1975). Poisson approximation for dependent trials. *The Annals of Probability* **3** 534–545. MR0428387

[8] Deledalle, C.-A., Denis, L. and Tupin, F. (2012). How to compare noisy patches? Patch similarity beyond Gaussian noise. *International J. of Computer Vision* **99** 86–102. MR2917021

[9] Deledalle, C. A., Duval, V. and Salmon, J. (2011). Non-local Methods with Shape-Adaptive Patches (NLM-SAP). *J. of Mathematical Imaging and Vision* 1-18. MR2910876

[10] Deledalle, C.-A., Vaiter, S., Fadili, J. and Peyré, G. (2014). Stein Unbiased GrAdient estimator of the Risk (SUGAR) for multiple parameter selection. *SIAM J. Imaging Sci.* **7** 2448–2487. MR3281282

[11] Donoho, D. L. and Johnstone, I. M. (1995). Adapting to Unknown Smoothness Via Wavelet Shrinkage. *J. of the American Statistical Association* **90** 1200–1224. MR1379464

[12] Duval, V., Aujol, J.-F. and Gousseau, Y. (2011). A bias-variance approach for the Non-Local Means. *SIAM J. Imaging Sci.* **4** 760–788. MR2831078

[13] Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *J. of the American Statistical Association* **81** 461–470. MR0845884

[14] Eldar, Y. C. (2009). Generalized SURE for exponential families: Applications to regularization. *IEEE Trans. Signal Process.* **57** 471–481. MR2603376

[15] Eldar, Y. C. and Mishali, M. (2009). Robust recovery of signals from a structured union of subspaces. *IEEE Trans. on Information Theory* **55** 5302–5316. MR2596977

[16] Evans, L. C. and Gariepy, R. F. (1992). *Measure theory and fine properties of functions*. CRC Press. MR1158660

[17] George, E. I., Liang, F. and Xu, X. (2006). Improved minimax predictive densities under Kullback-Leibler loss. *The Annals of Statistics* 78–91. MR2275235

[18] Gilbarg, D. and Trudinger, N. S. (1998). *Elliptic Partial Differential Equations of Second Order*, 2nd ed. *Classics in Mathematics* **517**. Springer. MR1814364

[19] Girard, A. (1989). A fast Monte-Carlo cross-validation procedure for large least squares problems with noisy data. *Numerische Mathematik* **56** 1–23. MR1012701

[20] GOLUB, G. H., HEATH, M. and WAHBA, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* 215–223. MR0533250

[21] GOODMAN, J. W. (1976). Some fundamental properties of speckle. *J. of the Optical Society of America* **66** 1145–1150.

[22] HALL, P. (1987). On Kullback-Leibler loss and density estimation. *The Annals of Statistics* 1491–1519. MR0913570

[23] HAMADA, M. and VALDEZ, E. A. (2008). CAPM and option pricing with elliptically contoured distributions. *J. of Risk and Insurance* **75** 387–409.

[24] HANNIG, J. and LEE, T. (2004). Kernel smoothing of periodograms under Kullback–Leibler discrepancy. *Signal Processing* **84** 1255–1266.

[25] HANNIG, J. and LEE, T. (2006). On Poisson signal estimation under Kullback–Leibler discrepancy and squared risk. *J. of Statistical Planning and Inference* **136** 882–908. MR2181981

[26] HUDSON, H. M. (1978). A natural identity for exponential families with applications in multiparameter estimation. *The Annals of Statistics* **6** 473–484. MR0467991

[27] KULLBACK, S. and LEIBLER, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics* 79–86. MR0039968

[28] LANDSMAN, Z. and NEŠLEHOVÁ, J. (2008). Stein's Lemma for elliptical random vectors. *J. of Multivariate Analysis* **99** 912–927. MR2405098

[29] LEHMANN, E. (1983). Theory of point estimation. *Wiley publication.* MR0702834

[30] LI, K.-C. (1985). From Stein's unbiased risk estimates to the method of generalized cross validation. *The Annals of Statistics* **13** 1352–1377. MR0811497

[31] LUISIER, F. (2010). The SURE-LET approach to image denoising PhD thesis, École polytechnique fédérale de lausanne.

[32] LUISIER, F., BLU, T. and UNSER, M. (2010). SURE-LET for orthonormal wavelet-domain video denoising. *IEEE Trans. on Circuits and Systems for Video Technology* **20** 913–919.

[33] LUISIER, F., BLU, T. and WOLFE, P. J. (2012). A CURE for noisy magnetic resonance images: Chi-square unbiased risk estimation. *IEEE Trans. on Image Processing* **21** 3454–3466. MR2960439

[34] LV, J. and LIU, J. S. (2014). Model selection principles in misspecified models. *J. of the Royal Statistical Society: Series B (Statistical Methodology)* **76** 141–167. MR3153937

[35] MALLOWS, C. L. (1973). Some Comments on Cp. *Technometrics* **15** 661–675.

[36] MORRIS, C. N. (1982). Natural exponential families with quadratic variance functions. *The Annals of Statistics* 65–80. MR0642719

[37] PESQUET, J.-C., BENAZZA-BENYAHIA, A. and CHAUX, C. (2009). A SURE Approach for Digital Signal/Image Deconvolution Problems. *IEEE Trans. on Signal Processing* **57** 4616–4632. MR2722323

[38] RAMANI, S., BLU, T. and UNSER, M. (2008). Monte-Carlo SURE: a black-

box optimization of regularization parameters for general denoising algorithms. *IEEE Trans. Image Process.* **17** 1540–1554. MR2517100

[39] Ramani, S., Liu, Z., Rosen, J., Nielsen, J.-F. and Fessler, J. A. (2012). Regularization parameter selection for nonlinear iterative image restoration and MRI reconstruction using GCV and SURE-based methods. *IEEE Trans. on Image Processing* **21** 3659–3672. MR2960455

[40] Raphan, M. and Simoncelli, E. P. (2007). Learning to be Bayesian without supervision. In *Advances in Neural Inf. Process. Syst. (NIPS)* **19** 1145–1152. MIT Press.

[41] Rigollet, P. (2012). Kullback–Leibler aggregation and misspecified generalized linear models. *The Annals of Statistics* **40** 639–665. MR2933661

[42] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6** 461–464. MR0468014

[43] Stein, C. M. (1981). Estimation of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics* **9** 1135–1151. MR0630098

[44] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. of the Royal Statistical Society. Series B. Methodological* **58** 267–288. MR1379242

[45] Vaiter, S., Deledalle, C.-A., Fadili, J., Peyré, G. and Dossal, C. (2017). The Degrees of Freedom of Partly Smooth Regularizers. *Annals of the Institute of Statistical Mathematics* **69** 791–832. MR3671639

[46] Van De Ville, D. and Kocher, M. (2009). SURE-Based Non-Local Means. *IEEE Signal Process. Lett.* **16** 973–976.

[47] Van De Ville, D. and Kocher, M. (2011). Non-local means with dimensionality reduction and SURE-based parameter selection. *IEEE Trans. Image Process.* **9** 2683–2690. MR2866262

[48] Yanagimoto, T. (1994). The Kullback-Leibler risk of the Stein estimator and the conditional MLE. *Annals of the Institute of Statistical Mathematics* **46** 29–41. MR1272745