C. R. Acad. Sci. Paris, Ser. I 349 (2011) 699–702

Contents lists available at ScienceDirect

C. R. Acad. Sci. Paris, Ser. I

www.sciencedirect.com

Statistics

Unbiased risk estimation and scoring rules

Estimation de risque non biaisée et règles d'évaluation

Werner Ehm

Institute for Frontier Areas of Psychology and Mental Health, Wilhelmstr. 3a, 79098 Freiburg, Germany

ARTICLE INFO

Article history:
Received 15 April 2011
Accepted 27 April 2011
Available online 16 June 2011

Presented by Paul Deheuvels

ABSTRACT

Stein unbiased risk estimation is generalized twice, from the Gaussian shift model to nonparametric families of smooth densities, and from the quadratic risk to more general divergence type distances. The development relies on a connection with local proper scoring rules.

RÉSUMÉ

La méthode d'estimation du risque Steinien est doublement généralisée, d'une part du modèle de translation Gaussien à des familles non paramétriques et d'autre part du risque quadratique à des distances du type divergence plus générales. Cette extension repose sur une relation avec des règles d'évaluation locales propres.

1. Introduction: SURE and the Hyvärinen score

Consider the problem of estimating the parameter θ in the standard Gaussian shift family $P_\theta = \mathcal{N}(\theta, I_d)$, $\theta \in \mathbb{R}^d$, based on an observation $x \in \mathbb{R}^d$. Let T be an estimator of θ of the form $T = x + g(x)$. Using partial integration, Stein [8] showed that under weak conditions about g, the quadratic risk $R(T,\theta) = E_\theta|T-\theta|^2$ of T can be estimated unbiasedly by the expression $\widehat{R}(T) = 2\operatorname{div} g(x) + |g(x)|^2 + d$ called *SURE* (Stein unbiased risk estimate), so that $E_\theta \widehat{R}(T) = R(T,\theta)$ for every $\theta \in \mathbb{R}^d$. Here $|\cdot|$ and $\langle\cdot,\cdot\rangle$ denote the Euclidean norm and inner product on $\mathbb{R}^d$, respectively, and $\operatorname{div} g$ is the divergence of g. If in particular $g = \nabla \log f$ for some function $f > 0$ on $\mathbb{R}^d$, the risk estimate becomes

$$\widehat{R}(T) = 2\Delta \log f(x) + \left|\nabla \log f(x)\right|^2 + d, \tag{1}$$

where as usual, ∇ denotes the gradient and $\Delta = \operatorname{div}\nabla$ the Laplace operator on $\mathbb{R}^d$. This special case occurs if T is the posterior mean with respect to a prior distribution π: then $T = x + \nabla \log f(x)$ where $f(x) = \int p_\theta(x)\,\mathrm{d}\pi(\theta)$ is the corresponding mixture density, so that $g = \nabla \log f$.

The striking similarity between SURE and the *Hyvärinen score* [5],

$$\mathrm{H}(p,x) = 2\frac{\Delta p(x)}{p(x)} - \left|\frac{\nabla p(x)}{p(x)}\right|^2 = 2\Delta \log p(x) + \left|\nabla \log p(x)\right|^2, \tag{2}$$

has been noted in, e.g., [6]. In Eq. (2), p denotes a sufficiently smooth, strictly positive probability density on $\mathbb{R}^d$. Originally, the Hyvärinen score was introduced for score matching, a minimum distance type estimation method. Its formal similarity

E-mail address: ehm@igpp.de.

1631-073X/$ – see front matter
doi:10.1016/j.crma.2011.04.015

to SURE is substantiated on reexpressing the risk of T as a distance between densities. Consider the *Hyvärinen divergence* defined for smooth, positive densities p, q on $\mathbb{R}^d$ as

$$d_{\mathrm{H}}(p,q) = \int \left|\nabla \log p(y) - \nabla \log q(y)\right|^2 q(y)\,\mathrm{d}y. \tag{3}$$

If $p = f$ is a mixture density as above and $q = p_\theta$ is the density of P_θ, we have $\nabla \log f(x) - \nabla \log p_\theta(x) = \nabla \log f(x) + x - \theta = T - \theta$, where again $T = x + \nabla \log f(x)$ is the corresponding posterior mean. Consequently,

$$R(T,\theta) = E_\theta |T-\theta|^2 = \int \left|\nabla \log f(x) - \nabla \log p_\theta(x)\right|^2 p_\theta(x)\,\mathrm{d}x = d_{\mathrm{H}}(f, p_\theta), \tag{4}$$

that is, *the risk $R(T,\theta)$ of the parameter estimate $T = x + \nabla \log f(x)$ equals a distance between densities, $d_{\mathrm{H}}(f, p_\theta)$*. Furthermore, the analogue of SURE in the density scenario is the Hyvärinen score $\mathrm{H}(f, x)$, essentially. In fact, Hyvärinen's idea, reinventing Stein's, was to apply partial integration to (3) which, assuming boundary terms vanish, gives

$$d_{\mathrm{H}}(p,q) = \int \left(2\Delta \log p(y) + \left|\nabla \log p(y)\right|^2\right) q(y)\,\mathrm{d}y + \int \left|\nabla \log q(y)\right|^2 q(y)\,\mathrm{d}y; \tag{5}$$

cf. [1,5]. Since $\int |\nabla \log p_\theta(x)|^2 p_\theta(x)\,\mathrm{d}x = d$ $(\theta \in \mathbb{R}^d)$ in the standard normal case, where $q = p_\theta$, it follows that

$$E_\theta\big(\mathrm{H}(f,x) + d\big) = E_\theta\big(2\Delta \log f(x) + \left|\nabla \log f(x)\right|^2 + d\big) = d_{\mathrm{H}}(f, p_\theta). \tag{6}$$

That is, *the modified Hyvärinen score $H(f,x) + d$ represents an unbiased estimate of the distance $d_{\mathrm{H}}(f, p_\theta)$ of f from the unknown "true" density p_θ, for any density $f > 0$ on $\mathbb{R}^d$ satisfying suitable regularity conditions.*

The purpose of this note is to expand on this aspect of unbiased risk estimation by tying it to scoring rules. Local proper scoring rules are constructed as gradients of concave functionals [3,4], and then shown to generalize SURE in that they furnish unbiased estimates of modified Bregman type distances. The development is related to (parts of) work by Dawid and Lauritzen [1]. See also [2,7].

2. Local proper scoring rules and unbiased risk estimation

We restrict the discussion of scoring rules to the setting relevant for this note, and refer to [3] for general information. Let $\mathcal{P}$ denote the class of all probability densities with respect to the Lebesgue measure on $\mathbb{R}^d$ such that the following conditions hold for every $p \in \mathcal{P}$: (P1) $p \in C^2$; (P2) $p > 0$ everywhere on $\mathbb{R}^d$; (P3) for every $m > 0$ and $i, j \in \{1, \ldots, d\}$

$$\lim_{|x|\to\infty} |x|^m \big(p(x) + \left|\partial_{x_i} p(x)\right| + \left|\partial^2_{x_i x_j} p(x)\right|\big) = 0;$$

(P4) there exists $a = a(p) > 0$ such that for $i, j \in \{1, \ldots, d\}$,

$$\lim_{|x|\to\infty} |x|^{-a} \left(\left|\log p(x)\right| + \left[\frac{\partial_{x_i} p(x)}{p(x)}\right]^2 + \frac{|\partial^2_{x_i x_j} p(x)|}{p(x)}\right) = 0.$$

The class $\mathcal{P}$ is quite large, being convex and comprising, e.g., all normal and logistic distributions.

A *scoring rule* is a mapping $\mathrm{S} : \mathcal{P} \times \mathbb{R}^d \to \mathbb{R}$ assigning a numerical score, $\mathrm{S}(p,x)$, to the density forecast, p, when the observation that materializes is x. We write $\mathrm{S}(p,q) = \int \mathrm{S}(p,x) q(x)\,\mathrm{d}x = E_q \mathrm{S}(p,\cdot)$ for the expected score when the density forecast is p and the probability measure underlying x is $q(x)\,\mathrm{d}x$. The scoring rule S is (*strictly*) *proper* relative to $\mathcal{P}$ if $\mathrm{S}(q,q) \leqslant \mathrm{S}(p,q)$ for all $p, q \in \mathcal{P}$ (with equality only if $p = q$). The scoring rule S is *local* (of order two, for the class $\mathcal{P}$) if there exists a real function s such that

$$\mathrm{S}(p,x) = \mathrm{s}\big(x, \log p(x), \nabla \log p(x), \nabla^2 \log p(x)\big) \quad \big(p \in \mathcal{P}, x \in \mathbb{R}^d\big),$$

$\nabla^2 f(x)$ denoting the Hessian matrix of second-order partial derivatives of a function $f : \mathbb{R}^d \to \mathbb{R}$ at x.

The classical example of a (strictly) proper local scoring rule is the logarithmic score, $\mathrm{S}(p,x) = -\log p(x)$. Another example is the Hyvärinen score (2). The latter can be regarded as being local of order two, in the obvious sense, and the former as local of order zero. Local scoring rules of any order $m \geqslant 0$ were recently investigated in [7], in the case $d = 1$. Hereafter, "local" always is understood as "local of order two."

The following result lifts the construction of local proper scoring rules in [2] from the one- to the higher-dimensional case $d \geqslant 1$. Let $\mathcal{K}$ denote the class of the *kernels* $\mathrm{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ satisfying the following conditions: (K1) $\mathrm{k} \in C^2$; (K2) there are constants $C, r \in (0,\infty)$ such that whenever k^* stands for the function $\mathrm{k} = \mathrm{k}(x,y)$ or any of its partial derivatives up to order two, then $|\mathrm{k}^*(x,y)| \leqslant C(1 + |x| + |y|)^r$ $(x, y \in \mathbb{R}^d)$. With any $\mathrm{k} \in \mathcal{K}$ we associate a functional $\Phi = \Phi_{\mathrm{k}} : \mathcal{P} \to \mathbb{R}$ defined by

$$\Phi(p) = \int_{\mathbb{R}^d} \mathrm{k}\big(x, \nabla \log p(x)\big) p(x)\,\mathrm{d}x \quad (p \in \mathcal{P}). \tag{7}$$

In view of the growth and decay conditions (K2), (P4), and (P3), the integral in (7) exists and is finite for every $p \in \mathcal{P}$. Let $\nabla_y \mathrm{k}$ denote the partial gradient referring to the argument $y \in \mathbb{R}^d$ of $\mathrm{k} = \mathrm{k}(x, y)$, and recall that $\operatorname{div} g(x)$ stands for the trace of the total derivative at x of a function $x \mapsto g(x)$ mapping $\mathbb{R}^d$ into itself.

Theorem 2.1. *Let $\mathrm{k} \in \mathcal{K}$, and suppose that the associated functional Φ is concave on $\mathcal{P}$. Then*

$$\mathrm{S}(p, x) = \mathrm{k}\big(x, \nabla \log p(x)\big) - \frac{1}{p(x)} \operatorname{div}\big[p(x) \nabla_y \mathrm{k}\big(x, \nabla \log p(x)\big)\big] \tag{8}$$

is a local proper scoring rule relative to $\mathcal{P}$. It is strictly proper if Φ is strictly concave. Furthermore, if $y \mapsto \mathrm{k}(x, y)$ is concave on $\mathbb{R}^d$ for every $x \in \mathbb{R}^d$, then the functional Φ is concave on $\mathcal{P}$.

Proof. The proof follows similar lines as in the case $d = 1$, see [2, Sections 4.1, 5.1]. We only indicate that the tangent construction in [2, Section 4.1] yields the scoring rule (8). To compute the (weak) gradient of Φ at $q \in \mathcal{P}$, let $p_t = (1-t)q + tp$ where $p \in \mathcal{P}$, $t \in [0, 1]$. Formal differentiation ignoring all technicalities gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\big[\Phi(p_t)\big] = \int \frac{\partial}{\partial t}[\mathrm{K}_{p_t} p_t]\,\mathrm{d}x = \int [\mathrm{K}_{p_t}](p - q)\,\mathrm{d}x + \int \left[\frac{\partial}{\partial t}\mathrm{K}_{p_t}\right] p_t\,\mathrm{d}x, \tag{9}$$

wherein we put $\mathrm{K}_{p_t}(x) = \mathrm{k}(x, \nabla \log p_t(x))$ and omitted the argument x of the integrands. For the last integral in (9) we get by the divergence theorem, assuming the boundary integral vanishes,

$$\int \left\langle \nabla_y \mathrm{k}(\cdot, \nabla \log p_t), \nabla\left(\frac{p - q}{p_t}\right)\right\rangle p_t\,\mathrm{d}x = -\int \operatorname{div}\big[p_t \nabla_y \mathrm{k}(\cdot, \nabla \log p_t)\big] \frac{p - q}{p_t}\,\mathrm{d}x. \tag{10}$$

Setting $t = 0$ in (9) and (10) and noting that $p_0 = q$ we find that

$$\frac{\mathrm{d}}{\mathrm{d}t}\big[\Phi(p_t)\big]\Big|_{t=0} = \int \left\{\mathrm{k}(\cdot, \nabla \log q) - \frac{1}{q} \operatorname{div}\big[q \nabla_y \mathrm{k}(\cdot, \nabla \log q)\big]\right\}(p - q)\,\mathrm{d}x. \tag{11}$$

Thus, the gradient of Φ at q is given by the expression in curly brackets in (11). The scoring rule resulting from the tangent construction, $\mathrm{S}(q, \cdot)$, differs from this gradient only by a correction term which can be shown to vanish. The negligibility of the boundary integral in (10), and all the technicalities (existence of integrals, exchangeability of differentiation and integration, etc.) can be settled similarly as in [2, Section 4.1], using the assumptions made about the classes $\mathcal{P}$ and $\mathcal{K}$. □

Any convex combination of a scoring rule S as in Theorem 2.1 with the logarithmic score yields a local proper scoring rule. In the case $d = 1$, scoring rules of this form exhaust the class of all local proper scoring rules [2,7]. The complete characterization in the case $d > 1$ remains open.

Examples. Let $\mathrm{k} \in \mathcal{K}$ be a kernel of the form $\mathrm{k}(x, y) = \mathrm{k}(y) = \psi(|y|)$, where ψ is a concave C^2-function on $[0, \infty)$ with $\psi(0) = \psi'(0) = 0$. Then $y \mapsto \mathrm{k}(y)$ is concave on $\mathbb{R}^d$, and the corresponding scoring rule (8) is proper. Explicitly we have

$$\mathrm{S}(p, \cdot) = \psi\big(|\sigma|\big) - \frac{\psi'(|\sigma|)}{|\sigma|}\big(|\sigma|^2 + \Delta \log p\big) - \left[\psi''\big(|\sigma|\big) - \frac{\psi'(|\sigma|)}{|\sigma|}\right]\left\langle \frac{\sigma}{|\sigma|}, \big(\nabla^2 \log p\big)\frac{\sigma}{|\sigma|}\right\rangle$$

where $\sigma = \nabla \log p$. For $\psi(t) = -t^2$ we obtain the Hyvärinen score (2); putting $\psi(t) = -\log\cosh t$ yields another interesting example parallel to [2, Example 5.3].

A local scoring rule S that is proper relative to $\mathcal{P}$ gives rise to a Bregman type divergence measure $d_{\mathrm{S}}(p, q) = \mathrm{S}(p, q) - \mathrm{S}(q, q)$ on $\mathcal{P} \times \mathcal{P}$. The following representation of d_{S} is closely related to [7, Eq. (53)].

Theorem 2.2. *Suppose that* S *is of the form* (8) *for some kernel $\mathrm{k} \in \mathcal{K}$ such that $y \mapsto \mathrm{k}(x, y)$ is concave on $\mathbb{R}^d$ for every $x \in \mathbb{R}^d$. Then the divergence d_{S} admits the representation*

$$d_{\mathrm{S}}(p, q) = E_q\left\{\mathrm{k}(\cdot, \nabla \log p) - \mathrm{k}(\cdot, \nabla \log q) + \left\langle \frac{\nabla q}{q} - \frac{\nabla p}{p}, \nabla_y \mathrm{k}(\cdot, \nabla \log p)\right\rangle\right\} \quad (p, q \in \mathcal{P}). \tag{12}$$

Proof. Let $p, q \in \mathcal{P}$. By the assumptions on $\mathcal{P}$ and $\mathcal{K}$, the divergence theorem applied to the scalar function $u(x) = q(x)/p(x)$ and the vector field $v(x) = p(x)\nabla_y \mathrm{k}(x, \nabla \log p(x))$ gives

$$\lim_{r \to \infty} -\int_{|x| \leqslant r} \frac{q}{p} \operatorname{div}\big[p\ \nabla_y \mathrm{k}(\cdot, \nabla \log p)\big]\,\mathrm{d}x = \lim_{r \to \infty} \int_{|x| \leqslant r} \left\langle \frac{\nabla q}{q} - \frac{\nabla p}{p}, \nabla_y \mathrm{k}(\cdot, \nabla \log p)\right\rangle q\,\mathrm{d}x. \tag{13}$$

The relation (12) follows on writing $d_S(p,q) = E_q\{S(p,\cdot) - S(q,\cdot)\}$, substituting (8) and using (13), and observing that $\int q^{-1} \operatorname{div}(q\nabla_y k(\cdot, \nabla \log q))q\,dx = 0$. □

Note that the expression in curly brackets in (12) is nonnegative because for a concave function f on $\mathbb{R}^d$ one has $f(y_1) - f(y_2) \geqslant \langle y_1 - y_2, \nabla f(y_1)\rangle$ $(y_1, y_2 \in \mathbb{R}^d)$. For the Hyvärinen score, where $k(x,y) = -|y|^2$, that expression becomes $|\nabla \log p - \nabla \log q|^2$, and d_S becomes the Hyvärinen divergence (3).

To clarify the connection with SURE we note that the partial integration in (13) was used conversely by Stein and Hyvärinen, to pass from the risk representation (12) to an expression of the form $E_q\{S(p,\cdot) - S(q,\cdot)\}$. In the latter, the scoring rule $S(p,\cdot)$ may serve as an unbiased estimate of $E_q S(p,\cdot)$, while the term $E_q S(q,\cdot)$ is the same for all candidates p, hence can be ignored if the focus is on risk *comparison*. In nonparametric density estimation, e.g., risk comparison of competing estimates is applied for bandwidth selection, using cross-validation. Briefly, if $\hat{p}_n = \hat{p}_n(\cdot|x_1, \ldots, x_n)$ is an estimate of the unknown density $q \in \mathcal{P}$ underlying the i.i.d. observations $x_1, \ldots, x_n$ that is symmetric in the x_i, the cross-validated expression $\widehat{R}_n(\hat{p}_{n-1}) = n^{-1} \sum_{i=1}^n S(\hat{p}_{n,-i}, x_i)$ is an unbiased estimate of $R_{n-1}(\hat{p}_{n-1}, q)$, where $R_n(\hat{p}_n, q) = E_q S(\hat{p}_n, q)$ denotes the modified risk ignoring the term $E_q S(q,\cdot) = S(q,q)$, which depends only on q.

The possibility of risk estimation is of course not confined to the local scoring rules considered here, as any proper scoring rule S, whether local or not, gives rise to a divergence measure d_S. Therefore, cross-validatory estimation of the (modified) risk generally is feasible, although exact unbiasedness as with the local scoring rules may not be achievable when global terms are involved. For example, unbiased estimation of the term $\int p(x)^2\,dx$ entering the quadratic score [3] does not seem possible.

The particular interest of the scoring rules of the form (8) ensues from the fact that they do not require the knowledge of the normalizing constants of the probability densities, which may be unknown or hard to obtain in complex settings [5,7]. This advantage can be combined with other desirable features such as improved robustness by working, for instance, with the log cosh scoring rule mentioned above.

Acknowledgements

The author thanks Tilmann Gneiting, Steffen Lauritzen, and Matthew Parry for helpful comments on earlier drafts of the paper.

References

[1] A.P. Dawid, S.L. Lauritzen, The geometry of decision theory, in: Proc. 2nd Int. Symp. Inf. Geom. Appl., Univ. Tokyo, 2005, pp. 22–28.
[2] W. Ehm, T. Gneiting, Local proper scoring rules of order two, Preprint, arXiv:1102.5031v1, 2011.
[3] T. Gneiting, A.E. Raftery, Strictly proper scoring rules, prediction, and estimation, J. Amer. Statist. Assoc. 102 (2007) 359–378.
[4] A.D. Hendrickson, R.J. Buehler, Proper scores for probability forecasters, Ann. Math. Statist. 42 (1971) 1916–1921.
[5] A. Hyvärinen, Estimation of non-normalized statistical models using score matching, J. Mach. Learn. Res. 6 (2005) 695–709.
[6] A. Hyvärinen, Optimal approximation of signal priors, Neural Computation 20 (2008) 3087–3110.
[7] M. Parry, A.P. Dawid, S. Lauritzen, Proper local scoring rules, Preprint, arXiv:1101.5011v1, 2011.
[8] C.M. Stein, Estimation of the mean of a multivariate normal distribution, Ann. Statist. 9 (1981) 1135–1151.