Unifying approach to score based statistical inference in physical sciences

Vaclav Kus\textsuperscript{1,2} and Richard Finger\textsuperscript{1}

\textsuperscript{1}Department of Mathematics, FNSPE, Czech Technical University in Prague, Trojanova 13, 12000 Praha 2, Czech Republic
\textsuperscript{2}Department of Adaptive Systems, UTIA, Czech Academy of Sciences, Pod Vodarenskou vezí 4, 182 08 Praha 8, Czech Republic

E-mail: vaclav.kus@fjfi.cvut.cz (corresponding author)

Abstract. In this contribution the statistical inference based on score functions is developed with the aim of future utilization throughout different fields of physics, for example in detector collision data processing or neutrino prongs matching. New score functions between theoretical and empirical probability measures are defined and the corresponding minimum score estimators are presented. We find that consistency of different estimators in various score functions leads to the well-known consistency in commonly used statistical distances or disparity measures between probability distributions. Conditions under which a specific score function pass to $\phi$–divergence are formulated. Conversely, each $\phi$–divergence is a score function. Furthermore, the minimization of arbitrary divergence score function leads to the classical histogram density estimator and thus can be used to alternative interpretation of histogram based calculations in (high energy) physics. The Kolmogorov-Smirnov testing statistics can be achieved through absolute score function under the class of mutually complement interval partitioning of the real line. It means that the most popular statistical methods, such as histogram estimation and Kolmogorov goodness of fit testing used in physics, can be covered by one unifying score based statistical approach. Also, these methods were previously successfully applied to data sets originated from the particular material elasticity testing (nondestructive defectoscopy) within Preisach-Mayergoyz space modeling.

1. Introduction

Many different fields of physical applications, such as material defectoscopy, elasticity, machine learning, or, especially high energy physics (HEP) experiments running on particle accelerators, produce a vast amount of raw data containing not only the crucial signal but also a large number of observations originating from another physical background processes. Prior to analyzing the data of the experiment, we need to separate only the desired useful observations from other physical backgrounds by using some appropriate statistical procedures.

In this paper we deal with statistical inference based on a family of score functions leading to the corresponding minimum score estimators which profits from new and classical statistical distances or disparity measures between probability distributions. For example, each $\phi$–divergence is a score function and the minimization of arbitrary divergence score function leads to the classical histogram density estimator used in (high energy) physics. Furthermore, from the score function setup, the Kolmogorov-Smirnov testing statistics can be achieved. It means that we come to the unifying score based statistical applications.
2. Score functions

Let $\mathbf{X}_n = (X_1, \ldots, X_n)$ denote a vector of i.i.d. observations, $\mathcal{P}$ be a nonempty subset of all probability measures $\mathcal{P}(\mathcal{X}, \mathcal{A})$ on a space $\mathcal{X}$ endowed with a $\sigma$-algebra $\mathcal{A}$, $P_0 \in \mathcal{P}$ be a distribution of the components of $\mathbf{X}_n$, and $P_n \in \mathcal{P}(\mathcal{X}, \mathcal{A})$ is the empirical distribution (measure) defined by $P_n(A) = \sum_{i=1}^{n} I(A_i)/n, A \in \mathcal{A}, n = 1, 2, \ldots$ Let $\mathcal{D}_n = \{A_1, \ldots, A_{n m_n}\} \subset \mathcal{A}$ be a sequence of finite or countable partitions of $\mathcal{X}$ and $P \in \mathcal{P}$. We define score function

$$M_\rho(P, P_n) = \sum_{A \in \mathcal{D}_n} P_n(A) \rho \left( \frac{P_n(A) - P(A)}{P(A)} \right),$$

where score kernel $\rho : (-1, \infty) \to [0, \infty)$ is the function which is nonincreasing on $(-1, 0)$ and nondecreasing on $(0, \infty)$, with $\rho(0) = 0$ and $\max\{\rho(-1), \rho(\infty)\} \geq 0$, under the convention $0 \cdot \infty = 0$ and standard limits $\rho(-1)$, $\rho(\infty)$, and $\rho(-1)/\infty$. Further, the supremum score function $M^*_\rho(P, P_n) = \sup_{D_n \in \mathcal{C}_n} M_\rho(P, P_n)$, where the supremum is taken over a class of partitions $\mathcal{C}_n$ of $\mathcal{X}$. In the case of general discrete or nondiscrete parametric or nonparametric models, we interpret the relative deviations $(P_n(A) - P(A))/P(A), A \in \mathcal{D}_n$, $P(A) \neq 0$, as residuals of data $\mathbf{X}_n$ under the model $P \in \mathcal{P}$. Partitions $\mathcal{D}_n$ may eventually be independent of the sample size $n \in \mathbb{N}$.

For discrete models we put $A = \{x\}, x \in \mathcal{X}$ and denote the empirical probabilities by $P_n(x) = (1/n) \sum_i I(X_i = x)$ and probabilities under the model $P_0, \theta \in \Theta$, by $P_0(x)$. Now, differentiating the score function $M_\rho(P_0, P_n)$ by $\theta$ we obtain the estimating equation in the form

$$\sum_x \rho_A(\delta(x)) \frac{\partial}{\partial \theta} P_0(x) = 0, \quad \delta(x) = \frac{P_n(x) - P_0(x)}{P_0(x)} \in [-1, \infty],$$

where $\rho_A$ is a function depending only on $\delta(x)$. Furthermore, if $\rho_A(\delta)$ is supposed to be increasing and twice differentiable on $(-1, \infty)$ with $\rho_A(0) = 0$ and $\rho'_A(0) = 1$ then we have obtained exactly the same estimation equation of [1] developed for the case of discrete models. The second order efficiency of estimators $\hat{\theta}$ based on the estimating equation, depends on the behaviour of $\rho_A(\delta)$ in the neighbourhood of $\delta = 0$ and robustness of $\hat{\theta}$ is affected by the asymptotic properties of $\rho_A(\delta)$ for $\delta \to \infty$.

For a given score function $M_\rho$, the score estimator $\hat{P}_n \in \mathcal{P}$ of $P_0$ ($M_\rho$-estimator) is defined by condition

$$M_\rho(\hat{P}_n, P_n) = \inf_{P \in \mathcal{P}} M_\rho(P, P_n) \quad a.s.$$

and it is approximate score estimator if the last equation holds up to the term $o(n^{-1/2})$. An estimator $\hat{P}_n$ is said to be consistent in score function $M_\rho$ if $M_\rho(P_0, \hat{P}_n) = o_p(1)$ as $n \to \infty$. The same definitions can be stated for (approximate) minimum supremum score estimator by means of score function $M^*_\rho$.

In the class of estimators $\hat{P}_n$ consistent in $M_\rho$ (or $M^*_\rho$), we seek for the asymptotic distribution of appropriately scaled deviations from their mean values and if $\mathbb{E}M_\rho(P_0, \hat{P}_n) = d_n + o(c_n^{-1})$ then $T_n = c_n(M_\rho(P_0, \hat{P}_n) - d_n)$ can be used as a goodness of fit test statistics. For parameterized families $\mathcal{P} = \{P_0 : \theta \in \Theta\}, \Theta \subset \mathbb{R}^d$, we replace everywhere $P_0$ by $P_{\theta_0}, \theta$ by $\hat{\theta}$, $\hat{P}_n$ by $\hat{P}_{\hat{\theta}_n}$, and inf$_{P \in \mathcal{P}}$ by inf$_{\theta \in \Theta}$, where $\theta_0$ denotes the true parameter.

3. Statistical inference based on score functions

From now on, consider $\mathcal{D}_n$ to be a partition of $\mathcal{X}$, $\mathcal{A}_n \subset \mathcal{A}$ the sub $\sigma$-algebra generated by partitions $\mathcal{D}_n$, and $P^{(n)}, P_n^{(n)}$ be restrictions of $P, P_n$ on $\mathcal{A}_n$. For a score kernel $\rho$, if the function $\phi(t) = \rho((1-t)/t), t \in (0, \infty)$, is convex on $(0, \infty)$ and strictly convex at $t = 1$
then $M_\rho(P, P_n) = D_\phi(P_n^{(\mu)}, P_n^{(\mu)})$. If $\mathcal{C}_n$ contains all $\mathcal{A}$–measurable finite partitions of $\mathcal{X}$ then $M_\rho^*(P, P_n) = D_\phi(P, P_n)$, where $D_\phi$ denotes the classical $\phi$–divergence [2, 3]. On the contrary, for a $\phi$–divergence, if we set

$$\rho(y) = \phi \left( \frac{1}{1 + y} \right) + \phi'_+ (1) \left( \frac{y}{1 + y} \right), \quad y \in (-1, \infty)$$

then $\rho$ is a score kernel, $D_\phi(P_n^{(\mu)}, P_n^{(\mu)}) = M_\rho(P, P_n)$, and $D_\phi(P, P_n) = M_\rho^*(P, P_n)$ if $\mathcal{C}_n$ contains all $\mathcal{A}$–measurable finite partitions of $\mathcal{X}$. These divergence score functions results in usual minimum (restricted) $\phi$–divergence estimator $\hat{P}_n$, or $\theta_n$, and its consistency in $\phi$–divergence. Thus, one can employ a quite deeply developed theory of $\phi$–divergences. Moreover, if a score function represents a metric distance then we obtain the minimum distance estimators (MDE) and tests.

Various specifications of $\rho(t)$ and collections $\mathcal{C}_n$ of partitions leads to a great variety of statistical methods. For example, if $\phi(t) = |t - 1|$, $(\mathcal{X}, \mathcal{A}) = (\mathbb{R}, \mathcal{B})$, and the class of considered partitions of $\mathbb{R}$ is $\mathcal{C}_n = \{(\{-\infty, x\}, [x, \infty]) : x \in \mathbb{R}\}$ then the score function $M_\rho^*$ is proportional to the Kolmogorov distance $K(P, P_n)$, i.e. the minimum $M_\rho^*$–estimator is identical with the minimum Kolmogorov estimator. Minimum Kolmogorov distance estimators have been studied in the nonparametric case in [4, 5]. However, under $\phi(t) = |t - 1|$ and the collection of all measurable binary partitions $\mathcal{C}_n = \{\{A, A - A\} : A \in \mathcal{A}\}$, the $M_\rho^*(P, P_n) = V(P, P_n)$, i.e. minimum $M_\rho^*$–estimator leads to the minimum total variation estimator $(L_1$–estimator). Moreover, if $\mathcal{C}_n$ contains all measurable finite partitions of $\mathcal{X}$ then $M_\rho^*(P, P_n)$ also coincides with total variation $V(P, P_n)$.

Further, if $\phi(t)$ is an arbitrary divergence function, $(\mathcal{X}, \mathcal{A}) = (\mathbb{R}, \mathcal{B})$, $\mathcal{P} \ll \mu$ with the Lebesque measure $\mu$ on $\mathbb{R}$, $X_1, \ldots, X_n$ are i.i.d. by $P_0 \in \mathcal{P}$, and if $D_n$ is a finite or infinite partition of $\mathbb{R}$ into intervals $\mathcal{D}_n = \{A_{n1}, \ldots, A_{nm_n}\}$ with $\mu(A_{nj}) = h_{nj} > 0$ for all $j$, then the minimum score $M_\rho^*$–estimator is the histogram density estimator $\hat{f}_n$ of the true density $f_0 = dP_0/d\mu$, defined for all $j$ by

$$\hat{f}_n(x) = \frac{P_n(A_{nj})}{h_{nj}} = \frac{1}{n h_{nj}} \sum_{i=1}^n I_{A_{nj}}(X_i), \quad x \in A_{nj}.$$ 

The consistency of histogram in the total variation score function $M_\rho^*$ ($L_1$–consistency) has been studied in a number of papers and the asymptotic normality of statistic $c_n(M_\rho^*(P_0, \hat{P}_n) - EM_\rho^*(P_0, \hat{P}_n))$ for classical histograms with uniform partitions was obtained in [6].

4. Examples of score functions

We give a brief summary of score functions widely used in the recent literature under the assumption that $\mathcal{C}_n$ contains all $\mathcal{A}$–measurable finite partitions of $\mathcal{X}$.

- Shannon information score function, Kullback–Leibler

$$M_\rho(P, Q) = \int p \ln \frac{p}{q} d\mu =: I(P, Q)$$

$$\rho(y) = \frac{y - \ln (1 + y)}{1 + y}, \quad y \in (-1, \infty).$$

This reflexive, nonsymmetric Shannon score function leads to the maximum entropy tests and estimates [7].
• Reversed information score, Reversed Kullback-Leibler

\[ M_\rho(P, Q) = \int q \ln \frac{q}{p} \, d\mu =: I(Q, P) \]

\[ \rho(y) = \ln(1 + y) + \frac{1}{1 + y} - 1, \quad y \in (-1, \infty). \]

It is well-known that this reversed Kullback logarithmic score function leads to the maximum likelihood estimates and generalized ratio tests.

• Total variation, L1–norm

\[ M_\rho(P, Q) = \int |p - q| \, d\mu = 2 \sup_{A \in \mathcal{A}} |P(A) - Q(A)| =: V(P, Q) \]

\[ \rho(y) = \frac{|y|}{1 + y}, \quad y \in (-1, \infty). \]

• Pearson’s \( \chi^2 \)–score, \( \chi^a \)–scores

\[ M_\rho(P, Q) = \int q^{1-a}|p - q|^a \, d\mu =: \chi^a(P, Q) \]

\[ \rho(y) = \left( \frac{|y|}{1 + y} \right)^a, \quad a \geq 1, \quad y \in (-1, \infty) \]

For \( a = 2 \), the Pearson’s \( \chi^2 \) score function, as well as Neyman’s reversed \( \chi^2 \) score function, lead to the Pearson and Neyman test statistic. The Cramér-von Mises and Anderson-Darling distances can be obtained through the concept of \( \chi^2 \)–divergences.

• Le Cam score function, Squared Le Cam distance

\[ M_\rho(P, Q) = \int \frac{(p - q)^2}{p + q} \, d\mu = 2 \left( 1 - 2 \int \frac{pq}{p + q} \, d\mu \right) =: LC^2(P, Q) \]

\[ \rho(y) = \frac{y^2}{(1 + y)(2 + y)}, \quad y \in (-1, \infty). \]

\( LC(P, Q) \) is proved to be a metric satisfying the triangle inequality.

• Hellinger score function, Squared Hellinger distance

\[ M_\rho(P, Q) = \int (\sqrt{p} - \sqrt{q})^2 \, d\mu = 2 \left( 1 - \int \sqrt{pq} \, d\mu \right) =: H^2(P, Q) \]

\[ \rho(y) = \left( \frac{1}{1 + y} - 1 \right)^2, \quad y \in (-1, \infty). \]

\( H(P, Q) \) is the metric. By means of Hellinger score function \( H^2 \) one can obtain the well known robust estimators of multivariate location and covariance proposed in [8].

• Matusita distances of orders \( a \)

\[ M_\rho(P, Q) = \int |p^a - q^a|^\frac{1}{a} \, d\mu =: M_a(P, Q) \]

\[ \rho(y) = \left| \left( \frac{1}{1 + y} \right)^a - 1 \right|^\frac{1}{a}, \quad 0 < a \leq 1, \quad y \in (-1, \infty) \]
All powers \((M_a(P, Q))^a, 0 < a \leq 1\), of the Matusita distances are metric distances satisfying the triangle inequality.

- **Power \(I_a\)-scores**

\[
M_a(P, Q) = \frac{1}{a(a - 1)} \left( \int p^a q^{1-a} \, d\mu - 1 \right) =: I_a(P, Q)
\]

\[
\rho(y) = \frac{1}{a(a - 1)} \left[ \left( \frac{1}{1 + y} \right)^a + a \frac{y}{1 + y} - 1 \right], \quad a \neq 0, 1, \quad y \in (-1, \infty).
\]

Note that \(I_{\frac{1}{2}}(P, Q) = 2M_{\frac{1}{2}}(P, Q) = 2H^2(P, Q), I_2(P, Q) = \chi^2(P, Q)/2, I_{-1}(P, Q) = \chi^2(Q, P)/2\) and the limits of \(I_a\) at \(a = 1\) and \(a = 0\) provided us with the Shannon information divergence \(I_0(P, Q) = I(P, Q)\) and the reversed Kullback-Leibler divergence \(I_{1}(P, Q) = I(Q, P)\).

### 5. Conclusion

We have studied the statistical inference based on score functions with the aim of future utilization throughout different fields of physics, for example in detector collision data processing or neutrino prongs matching. We found that different minimum score estimators and statistical tests leads to the well-known and commonly used statistical distances or disparity measures. It means that the most popular statistical methods, such as histogram estimation and Kolmogorov goodness of fit testing used in physics, can be covered by one unifying score based statistical approach. Also, these methods were previously successfully applied to data sets originated from the particular material elasticity testing (nondestructive defectoscropy) within Preisach-Mayergoyz space modeling. Also, the score functions are prepared for using by the machine learning algorithms (e.g. [9]) developed for the signal versus background separations in high energy physics decay channels or for the robust image signal reconstructions in neutrino experiments NOvA and DUNE at FNAL.

### Acknowledgments

These results were supported by the research grants SGS18/188/OHK4/3T/14 (MEYS), GA16-09848S, and LTT180001 (MEYS).

### References

[1] Lindsay BG 1994 Efficiency versus robustness. The case for minimum Hellinger distance and other methods
*Annals of Statistics* 22 1081–1114

[2] Pardo L 2006 *Statistical Inference Based on Divergence Measures* Chapman and Hall/CRC

[3] Kúš V, Morales D and Vajda I 2008 Extensions of the parametric families of divergences used in statistical
inference *Kybernetika* 44 95-112

[4] Kúš V 2004 Nonparametric Density Estimates Consistent of the Order of \(n^{-1/2}\) in the \(L_1\)-norm
*Metrca* 60 1–14

[5] Hrabakova J, Kus V 2017 Notes on Consistency of Some Minimum Distance Estimators with Simulation
Results *Metrca* 80 243–257

[6] Berlinet A, Devroye L and Győrfi L 1995 Asymptotic normality of \(L_1\)-error in density estimation *Statistics,* 26 329–343

[7] Csiszár I 1991 Why least squares and maximum entropy? An axiomatic approach to inference for nonlinear
inverse problems *Annals of Statistics* 19 2031–2066

[8] Tamura RN and Boos DB 1986 Minimum Hellinger distance estimation for multivariate location and covariance
*Journal of American Statistical Association* 81 223–229

[9] Bour P, Kus V and Franc J 2016 Statistical classification techniques in high energy physics (SDDT algorithm)
*J. Phys.: Conf. Series* 738 012034