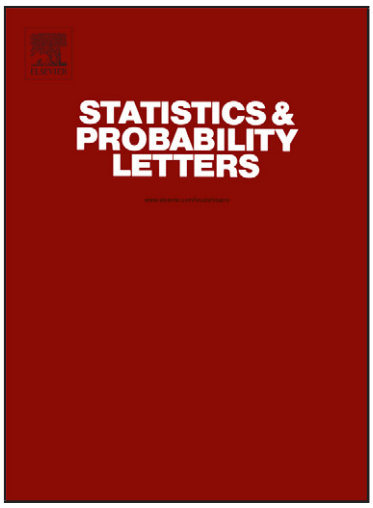
Consistency of minimum divergence estimators based on grouped data

Federico Bassetti, Antonella Bodini, Eugenio Regazzini



www.elsevier.com/locate/stapro

Consistency of minimum divergence estimators based on grouped data

Federico Bassetti, Antonella Bodini, Eugenio Regazzini

PII: S0167-7152(07)00005-3
DOI: doi:10.1016/j.spl.2006.11.021
Reference: STAPRO 4544

To appear in: *Statistics & Probability Letters*

Received date: 11 May 2004
Revised date: 9 September 2006
Accepted date: 26 November 2006

Cite this article as: Federico Bassetti, Antonella Bodini and Eugenio Regazzini, Consistency of minimum divergence estimators based on grouped data, *Statistics & Probability Letters* (2007), doi:10.1016/j.spl.2006.11.021

Consistency of minimum divergence estimators based on grouped data

Federico Bassetti [a] , Antonella Bodini [b] , Eugenio Regazzini [a,*]

[a] *Università degli Studi di Pavia, via Ferrata 1, 27100, Pavia, Italia*

[b] *CNR-IMATI, via Bassini 15, 20133, Milano, Italia*

Abstract

Consistency of minimum divergence estimators, based on grouped data, is studied under conditions which, to our knowledge, are weaker than the ones considered in the existing literature. Comments on the hypotheses and the interpretation of the main results are made, and an illustrative example is given.

Key words: Consistency of point estimators, minimum divergence estimators

1 Introduction.

Estimation methods based on the minimization of discrepancy between the empirical law of grouped observations and the discretization of a parametric model are well-known in literature, and their properties are widely studied.

* Corresponding author.
Email addresses: bassetti@dimat.unipv.it (Federico Bassetti), anto@mi.imati.cnr.it (Antonella Bodini), eugenio@dimat.unipv.it (Eugenio Regazzini).

These methods are largely used in economics and medical applications, for example, where observations are generally grouped, but the main interest lies in the distribution of the underlying continuous data (see, for instance, Victoria–Feser, 2000).

The main object of this work is to provide some consistency criteria for *minimum g–divergence estimates* (MgEs for short), where g–divergence is meant in the sense of Ali and Silvey (1966) and Csiszàr (1967). The resulting class of g–divergences contains several well–known measures of discrepancy such as the Kullback–Leibler divergence, the total variation distance, the Hellinger distance and the χ^2 distance. After noting that maximum likelihood estimators can be thought of as MgEs in a wide sense (see next section), motivation for considering g–divergence estimates different from the maximum likelihood is the efficiency and the robustness of many of them as explained, for example, in Lindsay (1994) and in Bassetti and Regazzini (2005). In the case of groupement defined by sample quantiles, consistency and asymptotic normality of MgEs are analyzed in Morales, Pardo and Vajda (1995), Menéndez, Morales and Pardo (1997), and in Menéndez, Morales, Pardo and Vajda (2001). In these papers the study of consistency is subordinated to the analysis of asymptotic normality and, therefore, it is carried out under conditions that are appropriate for asymptotic normality, but redundant for the validity of consistency. For example, restrictive conditions are usually considered like: monotonicity conditions with respect to the parameter, for the statistical model; regularity conditions such as the fact that the defining function g must be twice continuously differentiable in a neighborhood of 1. Consequently, important forms of g–divergence – such as the total variation distance – are excluded, in spite of the fact that "regular" minimum g–divergence estimates may perform less well

than minimum total variation estimates, from the point of view of robustness. See Section 4 of Bassetti and Regazzini (2005).

In view of these remarks, the present paper aims at improving and extending some of the existing results, by showing that the domain of validity of consistency of the estimates at issue can be widened considerably.

The paper is organized as follows. Section 2 deals with some preliminary topics, such as finiteness of the g–divergence, existence and measurability of minimum g–divergence estimates. Section 3 includes the main results about consistency. The proofs are omitted and the interested reader is referred to Bassetti, Bodini and Regazzini (2004).

2 Preliminaries.

Let $(\xi_n)_{n\geq 1}$ be a sequence of observations and X be the range of values of each of them. For every value θ of an unknown parameter varying in Θ, let p_θ be a probability distribution for the sequence $(\xi_n)_{n\geq 1}$ which makes the ξ_ns independent and identically distributed (i.i.d.) according to the probability α_θ defined on the σ-algebra $\mathcal{X}$ of subsets of X. It is assumed that the model is identifiable, i.e. $p_{\theta_1} \neq p_{\theta_2}$ whenever θ_1, θ_2 belong to Θ and $\theta_1 \neq \theta_2$. This paper deals with sample values grouped into classes $C_1, \ldots, C_k$ which form a measurable partition of X. This is tantamount to considering a sequence of "discrete" observations $(\xi_n^*)_{n\geq 1}$ defined by $\xi_n^* = j$ if ξ_n belongs to C_j ($j = 1, \ldots, k;\ n \geq 1$). The ξ_n^*s turn out to be i.i.d. with common distribution given by $\alpha_\theta^*(\{j\}) = p_\theta(\{\xi_1 \in C_j\})$, $j = 1, \ldots, k$. Throughout the paper, $\alpha_\theta^*(\{j\})$ is shortened to $\alpha_j(\theta)$ and $\tilde{n}_j$ denotes the number of sample values $\xi_1, \ldots, \xi_n$

in C_j so that the empirical distribution ν_n of the sample $(\xi_1^*, \ldots, \xi_n^*)$ can be defined by $\nu_n(\{j\}) := \tilde{n}_j/n,\ j = 1, \ldots, k$.

At this stage, one is in a position to define the g–divergence between α_θ^* and ν_n as

$$D_g(\alpha_\theta^*, \nu_n) = \sum_{j=1}^{k} \left[\mathbb{I}_{(0,+\infty)}(\tilde{n}_j) \frac{\tilde{n}_j}{n} g\left(n \frac{\alpha_j(\theta)}{\tilde{n}_j} \right) + \mathbb{I}_{\{0\}}(\tilde{n}_j)\, \bar{g}\, \alpha_j(\theta) \right] \tag{1}$$

where g is any real, continuous, convex function on $[0, +\infty)$ such that $\lim_{u \to +\infty} u^{-1} g(u) = \bar{g}$ and $g(1) = 0$, and $\mathbb{I}_A$ stands for the indicator function of the set A. Note that the total variation distance (d_{TV}), the Hellinger distance (d_H^2), the Kullback–Leibler divergence (d_{KL}) and the χ^2–distance are obtained for $g(s) = |s-1|$, $g(s) = (\sqrt{s}-1)^2$, $g(s) = s \log s$ $[0 \log 0 := 0]$ and $g(s) = (s-1)^2$, respectively. The maximum likelihood estimator can be obtained through $g(x) = -\log x$, but it should be observed that this function satisfies the properties of g except for continuity at 0.

Throughout the present paper it is assumed that: (a) X is a complete separable metric space (i.e. a Polish space); (b) Θ is a measurable subset of a metric space with distance function d and (c) whenever $\theta \mapsto D_g(\alpha_\theta^*, \nu_n)$ is finite, $\hat{\theta}_n$ satisfying

$$D_g(\alpha_{\hat{\theta}_n}^*, \nu_n) = \min\{D_g(\alpha_\theta^*, \nu_n) \ : \ \theta \in \Theta\} \tag{2}$$

exists in Θ. Conditions (a)-(c) will be referred to as "General Conditions" (GC for short) and $\hat{\theta}_n$ is just the *minimum g–divergence estimator* of θ dealt with in this paper.

For the sake of notational simplicity, $\nu_n,\ \tilde{n}_j$ for all $j = 1, \ldots, k$ and $\hat{\theta}_n$ will be considered as functions on X^∞ and D_g as a function from $X^\infty \times \Theta$ into $\bar{\mathbb{R}}$.

As for assumption (c), Bassetti (2004) provides conditions which, conjoined

with (a)-(b), are sufficient for the existence of MgEs. As an example, assume that the set $\{\theta \in \Theta : d_{TV}(a^*_\theta, \alpha^*_{\theta_0}) \leq T\}$ is a relatively compact subset of Θ for some T, and that the functions $\theta \mapsto \alpha_j(\theta)$ are continuous for every j. Then, a minimum total variation estimator $\hat{\theta}_n(x)$ exists for p_{θ_0}–almost all x, for all but a finite number of n. It is worth recalling that an analogous statement obtains for MgEs defined through any form of g satisfying $\phi(D_g(\pi_1, \pi_2)) \geq d_{TV}(\pi_1, \pi_2)$ for every probabilities π_1 and π_2, for some suitable strictly increasing continuous function ϕ, with $\phi(0) = 0$. Thus, one can guarantee the existence of minimum Hellinger distance estimators, minimum Kullback–Leibler estimators, minimum χ^2–distance estimators. See Propositions 2.1 and 2.2 of Bassetti (2004). Finally, it should be noted that (a)–(b), conjoined with the continuity of $\theta \mapsto \alpha_j(\theta)$ for every j, are sufficient for the existence of *near minima* of $\theta \mapsto D_g(a^*_\theta, \nu_n)$, $\tilde{\theta}_n$, i.e. $D_g(\alpha^*_{\tilde{\theta}_n}, \nu_n) \leq \inf_\Theta D_g(\alpha^*_\theta, \nu_n) + \epsilon_n$ ($\epsilon_n \to 0$). Moreover, it is easy to check that all the results given in the rest of the paper obtain for any sequence of *near minimum g–estimator*.

In general, even if Θ is a Polish space, $x \mapsto \hat{\theta}_n(x)$ is not necessarily a random variable (with respect to $\mathcal{X}^\infty/\mathcal{B}(\Theta)$) and, therefore, the issue of the measurability of a MgE deserves careful consideration. There are several studies regarding measurability of extrema. For example, Corollary 1 in Brown and Purves (1973) states that the measurability of $\hat{\theta}_n$ is guaranteed if (GC) hold together with: (a) Θ is a σ–compact subset of a Polish space; (b) $(x, \theta) \mapsto D_g(\alpha^*_\theta, \nu_{n,x})$ is measurable with respect to $(X^\infty \times \Theta, \mathcal{X}^\infty \otimes \mathcal{B}(\Theta))$, $\nu_{n,x}$ being the realization of ν_n when $(\xi_n)_{n\geq 1} = x$; (c) $\theta \mapsto D_g(\alpha^*_\theta, \nu_{n,x})$ is a lower semicontinuous function for every x in X^∞. As a matter of fact, a straightforward application of this corollary gives

Proposition 1 *Let* (GC) *be in force and let* Θ *be a* σ*–compact subset of a*

Polish space. If $\theta \mapsto \alpha_j(\theta)$ *is continuous for every* $j = 1, \ldots, k$ *and* $(x, \theta) \mapsto D_g(\alpha^*_\theta, \nu_{n,x})$ *is finite on* $X^\infty \times \Theta$ *for some* n*, then* $\hat{\theta}_n$ *is* $(X^\infty, \mathcal{X}^\infty)/\ (\Theta, \mathcal{B}(\Theta))$*–measurable.*

Proof. See proof of Proposition 2.1 in Bassetti, Bodini and Regazzini (2004).

When $\bar{g}$ is finite then D_g is finite. The following proposition gives a sufficient condition in order that D_g may be finite even if $\bar{g} = +\infty$.

Proposition 2 *If* $\alpha_j(\theta_0) > 0$ *for every* $j = 1, \ldots, k$ *and some* θ_0 *in* Θ*, then there is an event* N *in* $\mathcal{X}^\infty$*, of* p_{θ_0}*-probability zero, such that for every* x *in* N^c *there exists a positive integer* $\bar{n} = \bar{n}(x)$ *such that* $D_g(\alpha^*_\theta, \nu_{n,x})$ *is finite for every* $n \geq \bar{n}(x)$ *and for every* θ *in* Θ*.*

Proof. See proof of Proposition 2.2 in Bassetti, Bodini and Regazzini (2004).

3 Consistency criteria for MgEs.

The first criterion, inspired by Corollary 3.2.3 in van der Vaart and Wellner (1996), works under the assumption that θ_0 is a well-separated point of minimum of $\theta \mapsto D_g(\alpha^*_\theta, \alpha^*_{\theta_0})$, in addition to the hypothesis that the probability $\alpha^*_{\theta_0}$ dominates $\{\alpha^*_\theta\ :\ \theta \in \Theta\}$.

Proposition 3 *Let* (GC) *be inforce. Moreover assume that* $\bar{g}$ *is finite and the following conditions are valid:* (i) $\alpha_j(\theta) = 0$ *for every* θ *whenever* $\alpha_j(\theta_0) = 0$*;* (ii) *for every* $\epsilon > 0$, $\inf\{D_g(\alpha^*_\theta, \alpha^*_{\theta_0})\ :\ d(\theta, \theta_0) > \epsilon\} > 0$*. Then there exists a set* N *in* $\mathcal{X}^\infty$ *such that* $p_{\theta_0}(N) = 0$ *and* $\lim_{n \to +\infty} \hat{\theta}_n(x) = \theta_0$ $(x \in\ N^c)$*.*

Proof. See proof of Proposition 3.1 in Bassetti, Bodini and Regazzini (2004).

In view of Proposition 2, the assumption that $\bar{g}$ is finite can be replaced by the rather common condition that $\alpha_j(\theta_0)$ is strictly positive for every j:

Corollary 1 *Under* (GC), *if* $\alpha_j(\theta_0) > 0$ *for every* j *and if conditions* (i)–(ii) *in* Proposition 3 *obtain, then there is a set* N *in* $\mathcal{X}^\infty$ *with* $p_{\theta_0}(N) = 0$ *such that, for every* x *in* N^c, $\hat{\theta}_n(x)$ *exists whenever* $n \geq \bar{n}$, *for some suitable* $\bar{n} = \bar{n}(x)$, *and* $\hat{\theta}_n(x) \to \theta_0$ *as* $n \to +\infty$.

Proof. See proof of Corollary 3.2 in Bassetti, Bodini and Regazzini (2004).

The assumption that $\bar{g}$ is finite excludes many interesting forms of g–divergence such as the Kullback–Leibler divergence and the χ^2–distance.

In contrast to Proposition 3 and Corollary 1, the following propositions do not place any restriction on the support of the elements of $\{\alpha_\theta^* : \theta \in \Theta\}$. Moreover, they encompass the total variation distance, the Hellinger distance, the Kullback–Leibler divergence, the χ^2–distance and other distances with $g(s) = |s-1|^p$.

Proposition 4 *Let* (GC) *be in force and let* g *be one of the following functions defined for* $x \geq 0$: $g(x) = |x-1|$, $g(x) = (\sqrt{x}-1)^2$. *Moreover, let the following conditions be satisfied:* (i) *for every* $\epsilon > 0$ *there exists* $\epsilon' = \epsilon'(\epsilon) > 0$ *such that for every* θ *in* Θ, *with* $d(\theta, \theta_0) \geq \epsilon$, $\max_j\{|\alpha_j(\theta) - \alpha_j(\theta_0)|\} \geq \epsilon'$ *is valid. Then, there exists a subset* N *in* $\mathcal{X}^\infty$ *for which* $p_{\theta_0}(N) = 0$ *and* $\lim_{n\to+\infty} \hat{\theta}_n(x) = \theta_0$ $(x \in N^c)$.

Proof. See proof of Proposition 3.3 in Bassetti, Bodini and Regazzini (2004).

When $\alpha_j(\theta) > 0$ for every j, this result can be extended to further forms of g; for example, $g(x) = |x-1|^p$ with $p > 1$, or $g(x) = x \log x$. In fact, the proof

of Proposition 4 (see Bassetti, Bodini and Regazzini (2004)) can be adapted to these new cases by resorting to Proposition 2 and to the inequalities (see Dacunha-Castelle and Duflo (1994), Section 6.4) $D_g(\alpha^*_\theta, \nu_n) \geq d_{TV}(\alpha^*_\theta, \nu_n)^p$ if $g(x) = |x-1|^p$, $d_{KL}(\alpha^*_\theta, \nu_n) \geq d^2_H(\alpha^*_\theta, \tilde{\nu}_n)$, to obtain

Corollary 2 *Let* (GC) *be in force with* $\alpha_j(\theta_0) > 0$ *for every* j*, and let* g *be one of the functions considered in* Proposition 4 *or one of the following ones:* $g(x) = |x-1|^p$ *with* $p > 1$, $g(x) = x \log x$*. Moreover, let condition* (i) *of* Proposition 4 *be satisfied. Then, there is a set* N *in* $\mathcal{X}^\infty$ *with* $p_{\theta_0}(N) = 0$ *such that, for every* x *in* N^c, $\hat{\theta}_n(x)$ *exists whenever* $n \geq \bar{n}(x)$*, for some* $\bar{n}(x)$*, and* $\hat{\theta}_n(x) \to \theta_0$ *as* $n \to +\infty$.

Proof. See proof of Corollary 3.4 in Bassetti, Bodini and Regazzini (2004).

Remark 1. A quick look at the proof of Proposition 4 leads to the conclusion that the theses of Proposition 4 and Corollary 2 can be extended, if $\alpha_j(\theta_0)$ is strictly positive for every $j = 1, \ldots, k$, to functions g, which do not need to generate divergences, but are majorant of some of the specific gs considered therein. Hence, since $-\log x \geq -2(\sqrt{x}-1)$ and $g(x) = -2(\sqrt{x}-1)$ generates the Hellinger distance, Corollary 2 yields the strong consistency of maximum likelihood estimators.

Remark 2. All the previous propositions state that $\hat{\theta}_n(x) \to \theta_0$, as $n \to +\infty$, for every x in the complement of a $p_{\theta_0}-$null subset of X^∞. In other words: $(\hat{\theta}_n)_{n\geq 1}$ converges to θ_0 *almost surely with respect to the inner probability* associated to p_{θ_0}. See, for example, Sections 1.2 and 1.9 of van der Vaart and Wellner (1996). It should be mentioned that if $\hat{\theta}_{n'}$ is nonmeasurable for any term of a strictly increasing subsequence (n') of integers, then the above almost

sure convergence does not imply, for instance, convergence in outer probability. On the other hand, if $\hat{\theta}_n$ is measurable for all but a finite number of indices n, which hold true under mild conditions as shown in Proposition 1, then what has been proved in the previous propositions and corollaries is equivalent to $\lim_{n_1 \to +\infty} p_{\theta_0}\{d(\hat{\theta}_n, \theta_0) \leq \epsilon \quad \text{for every} \quad n \geq n_1\} = 1$ $(\forall\ \epsilon > 0)$.

Example. Let X coincide with the real axis $\mathbb{R}$, let Θ stand for $\mathbb{R} \times (0, +\infty)$ and denote the Euclidean norm by $\|\cdot\|_2$. Consider the nondegenerate intervals: $C_j = (x_{j-1}, x_j]$ with $j = 1, \ldots, k-1$ and $C_k = (x_{k-1}, x_k)$, where $x_0 = -\infty$ and $x_k = +\infty$. Let A be a strictly increasing continuous probability distribution function on $\mathbb{R}$. For every $\theta := (\mu, \sigma)$ in Θ, define α_θ to be the probability measure having $x \mapsto A((x-\mu)/\sigma)$ as probability distribution function. Then, $\{\alpha_\theta : \theta \in \Theta\}$ forms a location-scale family. Now, observe that $\theta \mapsto \alpha_j(\theta)$ is continuous for each j. At this stage, note that θ must coincide with θ_0 if $k \geq 3$ and $\alpha_j(\theta) = \alpha_j(\theta_0)$ for every j. Indeed, if these equalities were valid for $j = 1, \ldots, k$, then $A\Big(\frac{x_j - \mu}{\sigma}\Big) = A\Big(\frac{x_j - \mu_0}{\sigma_0}\Big)$ for $j = 1, \ldots, k-1$. This fact yields $\sigma^{-1}(x_j - \mu) = \sigma_0^{-1}(x_j - \mu_0)$, i.e. $x_j(\sigma_0 - \sigma) = \mu\sigma_0 - \mu_0\sigma$ for $j = 1, k-1$, that is to say: $\sigma = \sigma_0$ and $\mu = \mu_0$. At this stage it can be shown that $\inf\{d_{TV}(\alpha^*_\theta, \alpha^*_{\theta_0}) : \theta \in K_q\}$ is strictly positive for every q, $(K_q)_{q \geq 1}$ being a sequence of compact subsets of Θ such that $\cup K_q = \Theta \setminus \{\theta : \|\theta - \theta_0\|_2 < \epsilon\}$. Indeed, if there is a $\bar{q}$ for which $\inf\{d_{TV}(\alpha^*_\theta, \alpha^*_{\theta_0}) : \theta \in K_{\bar{q}}\} = 0$, then, by compactness, $d_{TV}(\alpha^*_{\bar{\theta}}, \alpha^*_{\theta_0}) = 0$ for some $\bar{\theta}$ in $K_{\bar{q}}$, which is a flagrant contradiction. So, in order to prove that (i) in Proposition 4 holds, it suffices to show that for any sequence $(\theta_q)_{q \geq 1}$ with θ_q in K_q, for every q, there is j for which $\alpha_j(\theta_q) \not\to \alpha_j(\theta_0)$, because in this case $\epsilon' := \inf\{d_{TV}(\alpha^*_\theta, \alpha^*_{\theta_0}) : \theta \in \cup_q K_q\}$ is strictly positive. In the case under discussion, it is easy to see that there must be j for which $\alpha_j(\theta_q) \to 0$.

References

Ali, S. M., Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another, *J. R. Stat. Soc. B* **28**, 131–142.

Bassetti, F. (2004). Asymptotic properties of some minimum discrepancy estimators. *Phd. Thesis.* Dipartimento di Matematica "F. Casorati", Università degli Studi di Pavia.

Bassetti, F., Bodini, A., Regazzini, E. (2004). Consistency of minimum divergence estimators based on grouped data. Technical Report MI/04-01, CNR–IMATI, Milano. http://www.mi.imati.cnr.it/iami/abstracts/04-01.html

Bassetti, F., Regazzini, E. (2005). Asymptotic distribution and robustness of minimum total variation distance estimators, *Metron.* **1**, 55-80.

Brown, L. D., Purves, R. (1973). Measurable selections of extrema, *Ann. Statist.* **1**, 902-912.

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations., *Studia Sci. Math. Hungar.* **2**, 299-318.

Dacunha-Castelle, D., Duflo, M. (1994). *Probabilités et Statistiques. 1: Problèms à Temps Fixe.* (Masson, Paris, 2nd ed.).

Lindsay, B. G., (1994). Efficiency versus robustness: the case for minimum Hellinger distance and related methods, *Ann. Statist.* **22**, 1081–1114.

Menéndez, M., Morales, D., Pardo, L. (1997). Maximum entropy principle and statistical inference on condensed ordered data, *Statist. Probab. Lett.* **34**, 85–93.

Menéndez, M., Morales, D., Pardo, L., Vajda, I. (2001). Minimum divergence estimators based on grouped data. *Ann. Inst. Statist. Math.* **53**, 277–288.

Morales, D., Pardo, L., Vajda, I. (1995). Asymptotic divergence of estimates of discrete distributions. *J. Statist. Plann. Inference* **48**, 347–369.

van der Vaart, A. W., Wellner, J. A. (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics* (Springer-Verlag, New York).

Victoria–Feser, M. P. (2000). Robust methods fo the analysis of income distribution, inequality and poverty. *International Statistical Review* **68**, 277-293.