Scandinavian Journal of Statistics Theory and Applications

Scandinavian Journal of Statistics, Vol. 42: 1092–1108, 2015
doi: 10.1111/sjos.12153
© 2015 Board of the Foundation of the Scandinavian Journal of Statistics. Published by Blackwell Publishing Ltd.

Generalized Maximum Spacing Estimation for Multivariate Observations

KRISTI KULJUS
Department of Mathematics and Mathematical Statistics, Umeå University

BO RANNEBY
Department of Forest Economics, Swedish University of Agricultural Sciences

ABSTRACT. In this paper, the maximum spacing method is considered for multivariate observations. Nearest neighbour balls are used as a multidimensional analogue to univariate spacings. A class of information-type measures is used to generalize the concept of maximum spacing estimators. Weak and strong consistency of these generalized maximum spacing estimators are proved both when the assigned model class is correct and when the true density is not a member of the model class. An example of the generalized maximum spacing method in model validation context is discussed.

Key words: divergence measures, maximum spacing estimation, nearest neighbour balls, strong consistency, weak consistency

1. Generalized maximum spacing estimate

1.1. Introduction

For independent and identically distributed univariate observations, a new estimation method, the maximum spacing (MSP) method, was defined in Ranneby (1984) and independently by Cheng & Amin (1983). In Ranneby *et al.* (2005), the MSP method was extended to multivariate observations for the Kullback-Leibler information measure using both nearest neighbour balls and Dirichlet cells. The approach with Dirichlet cells was previously applied in Ranneby (1996) and studied in more detail by Jimenez & Yukich (2002). In this paper, the multivariate maximum spacing estimation method based on nearest neighbour balls is considered for a broader class of information-type measures. We prove both weak and strong consistency of these generalized maximum spacing estimators under general conditions. In the univariate case, such generalized MSP estimators based on different metrics were studied in Ranneby & Ekström (1997), Ekström (2001) and Ghosh & Jammalamadaka (2001). Strong consistency of the MSP estimators in the case of Kullback-Leibler information and for univariate observations was proved in Ekström (1997) and Shao & Hahn (1999). Because estimators based on different information measures have different properties (regarding e.g. robustness, bias and variance), they behave differently in various situations. Which estimator is more suitable in any particular situation depends on these properties. For example, even if many suitable choices of information measure lead to limiting normal distributions of maximum spacing functions, the speed of convergence can be quite different (Penev & Ruderman, 2011). As mentioned and exemplified already in Ranneby (1984), an advantage of the maximum spacing method compared with the maximum likelihood method is the possibility of checking the validity of the assigned model class at the same time with solving the estimation problem. In this article, we will demonstrate that combining information from spacing functions under different divergence measures can provide further insight in the model validation context.

In Section 2 we will prove weak consistency and in Section 3 strong consistency of the generalized MSP estimators. Some issues concerning a suitable choice of information measure in different situations in the multivariate case will be discussed and illustrated in Section 4.

1.2. Notation and definitions

Let $\xi_1, \dots, \xi_n$ be a sequence of independent and identically distributed d-dimensional random vectors with distribution P_0 that is absolutely continuous with respect to Lebesgue measure. Let the corresponding density function be $g(x)$. Define the nearest neighbour distance to the random variable ξ_i as

$$R_n(i) = \min_{j \neq i} |\xi_i - \xi_j| , \qquad i = 1, \dots, n .$$

Let $B(x, r) = \{y : |x - y| \leq r\}$ denote the ball of radius r and centre x. Let NN_i denote the nearest neighbour of ξ_i, and let $B_n(\xi_i)$ denote its nearest neighbour ball, that is, this is a ball with centre ξ_i and radius $R_n(i)$. Suppose, we assign a model with density functions $\{f_\theta(x),\ \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^q$. Define random variables $z_{i,n}(\theta)$ as

$$z_{i,n}(\theta) = nP_\theta(B_n(\xi_i)) , \qquad i = 1, \dots, n .$$

Let $h : (0, \infty) \to (-\infty, 0]$ be a strictly concave function that has its maximum at $x = 1$. The following functions are some examples of such h:

$$h_1(x) = \ln x - x + 1, \quad h_2(x) = (1 - x)\ln x, \quad h_3(x) = -|1 - x^{1/p}|^p,$$

$$h_4(x) = -|1 - x|^p, \quad h_5(x) = \mathrm{sgn}(1 - \alpha)(x^\alpha - \alpha x + \alpha - 1),$$

where $\alpha > 0, \alpha \neq 1$ and $p \geq 1$. Here, h_2 corresponds to Jeffreys' divergence measure, h_3 to the Hellinger distance, h_4 to Vajda's measure of information and h_5 to Rényi's divergence measure. It is natural to generalize the maximum spacing method to the multivariate case and to define the generalized maximum spacing function $S_n(\theta)$ as follows:

$$S_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} h(z_{i,n}(\theta)) .$$

Definition 1. The parameter value that maximizes $S_n(\theta)$ is called the generalized maximum spacing estimate (GMSP estimate) of θ and denoted by $\hat{\theta}_n$. If $\sup_\theta S_n(\theta)$ is not attained for any θ in the admissible set Θ, the GMSP estimate $\hat{\theta}_n$ is defined as any point of Θ that satisfies

$$S_n(\hat{\theta}_n) \geq -c_n + \sup_{\theta \in \Theta} S_n(\theta) ,$$

where $c_n > 0$ is a sequence of constants such that $c_n \to 0$ as $n \to \infty$.

We need some further notation. Let $||B(x, r)||$ denote the volume of the ball $B(x, r)$. Define random variables $\eta_{i,n}$ as

$$\eta_{i,n} = n||B_n(\xi_i)|| , \qquad i = 1, \dots, n.$$

Let

$$z_n(\theta, x, y) = nP_\theta(B(x, r_n)) , \quad \text{where} \quad ||B(x, r_n)|| = y/n .$$

© 2015 Board of the Foundation of the Scandinavian Journal of Statistics.

In Ranneby *et al.* (2005), it was shown that $(\xi_i, \eta_{i,n})$ converges in distribution to (X, Y), where X has density $g(x)$ and $Y|X = x$ is exponentially distributed with the parameter $g(x)$. Denote the distribution function of $(\xi_i, \eta_{i,n})$ by $P_n(x, y)$ and the distribution function of (X, Y) by $P(x, y)$, then the density of (X, Y) is given by $p(x, y) = g^2(x)e^{-yg(x)}$, where $y > 0$. Consider a constant $M > 0$ and define the following functions:

$$t_M(x) = \max\{-M, h(x)\}, \quad T_n(M, \theta) = \frac{1}{n}\sum_{i=1}^{n} t_M(z_{i,n}(\theta)),$$

$$T(M, \theta) = \int t_M(yf_\theta(x))dP(x, y), \quad T(\theta) = \int h(yf_\theta(x))dP(x, y).$$

Observe that $T_n(M, \theta)$ is an approximation to $S_n(\theta)$ in the neighbourhood of $\sup_\theta S_n(\theta)$, and $T(M, \theta)$ is the limiting function of $E\, T_n(M, \theta)$. Further, $T(\theta) = \lim_{M\to\infty} T(M, \theta)$.

1.3. Consistency of the generalized maximum spacing estimate

We will prove both weak consistency and strong consistency of the GMSP estimate in two distinguished cases:

(1) The assigned model class is correct, that is, there exists $\theta_0 \in \Theta$ such that $g(x) = f_{\theta_0}(x)$.
(2) The true density $g(x)$ does not necessarily belong to $\{f_\theta(x),\ \theta \in \Theta\}$.

The idea is to approximate $S_n(\theta)$ with a bounded function $T_n(M, \theta)$. Suppose $S_n(\theta)$ converges to $T(\theta)$ uniformly in θ and that $T(\theta)$ has a unique maximum at θ_0. With convergence, we mean either convergence in probability or almost surely. Then $S_n(\hat{\theta}_n)$ converges to $\sup_\theta T(\theta) = T(\theta_0)$ and $S_n(\hat{\theta}_n) - T(\hat{\theta}_n)$ converges to zero, implying that $T(\hat{\theta}_n)$ converges to $T(\theta_0)$. An identifiability condition then implies the convergence of $\hat{\theta}_n$ to θ_0. However, ensuring uniform convergence of $S_n(\theta)$ is too restrictive. Because we are interested in convergence of $\sup_\theta S_n(\theta)$, it does not matter what happens with small values of $S_n(\theta)$. Therefore, we can consider the approximation $T_n(M, \theta)$ and rely on uniform convergence of $T_n(M, \theta)$. To obtain consistent GMSP estimates, we need conditions that prevent distributions corresponding to neighbouring parameter values from varying too much.

Condition C1 (continuity condition)**.** For each $\varepsilon > 0$ and $\eta > 0$, there exists an integer m, a partition of Θ into disjoint sets $\Theta_1, \ldots, \Theta_m$, compact sets $K_j = A_j \times [b_{j,1}, b_{j,2}] \subset \mathbb{R}^d \times \mathbb{R}^+$ and parameter values $\psi_j \in \Theta_j$, $j = 1, \ldots, m$, such that for each j

(i) $P((X, Y) \in K_j) > 1 - \eta$.
(ii) $\sup_{\theta\in\Theta_j} |z_n(\theta, x, y) - z_n(\psi_j, x, y)| < \varepsilon$ for all $(x, y) \in K_j$ and for all $n \geq n^*(\varepsilon, \eta)$.

Condition C2 (identifiability condition)**.** There exists a point $\theta_0 \in \Theta$ that uniquely maximizes $T(\theta)$. For each $\delta > 0$, there exists a constant $M_1 = M_1(\delta)$ such that

$$\sup_{\theta\in B^c(\theta_0,\delta)} T(M_1, \theta) < T(\theta_0).$$

For comments and examples regarding the continuity condition C1, see Ranneby (1984), Ekström (1998) and Ranneby *et al.* (2005). The identifiability condition C2 is a strong identifiability condition. When weak identifiability conditions are used instead, these are usually combined with other conditions implying that a strong identifiability condition is satisfied.

© 2015 Board of the Foundation of the Scandinavian Journal of Statistics.

Proposition 1. *Under the following assumptions, C2 is satisfied:*

(i) *For almost all x, $f_\theta(x)$ is a continuous function of θ, $\theta \in \Theta$.*
(ii) *For each $\theta \in \bar{\Theta} - \Theta$ and $\theta' \in \Theta$, $\lim_{\theta' \to \theta} f_{\theta'}(x)$ exists.*
(iii) *For $\theta \in \bar{\Theta}$, $\theta \neq \theta_0$ and $\theta' \in \Theta$, $\mu\left(\{x : \lim_{\theta' \to \theta} f_{\theta'}(x) \neq f_{\theta_0}(x)\}\right) > 0$, where μ is Lebesgue measure.*

Proof. When the assigned model class is true, an application of Jensen's inequality conditionally on $Z = Yg(X)$, together with (iii), guarantees that $T(\theta)$ has a unique maximum at θ_0:

$$T(\theta) = E[h(Yf_\theta(X))] = E_Z E[h(Zf_\theta(X)/g(X))|Z] \leq E[h(Yg(X))] = T(\theta_0).$$

In the case of wrong model class, we assume that $T(\theta)$ has a unique maximum at θ_0. For any fixed θ, $T(M, \theta)$ is a sequence of decreasing functions, and according to the monotone convergence theorem, $T(M, \theta) \searrow T(\theta)$ as $M \to \infty$. The Lebesgue dominated convergence theorem, together with (i) and (ii), implies that for every $M > 0$, $T(M, \theta)$ is a continuous function of θ for $\theta \in \bar{\Theta}$. Fatou's lemma gives that $T(\theta)$ is upper semi-continuous for $\theta \in \bar{\Theta}$.

Assume that Θ is bounded, take an arbitrary $\delta > 0$ and consider the following compact subset of $\bar{\Theta}$: $\{\theta : |\theta - \theta_0| \geq \delta\}$. Because $T(\theta)$ is upper semi-continuous, it attains its maximum at $\theta = \theta^*$, say (Royden, 1968, p. 161). Let $T(\theta_0) - T(\theta^*) = a_\delta$. For $\theta \in \{\theta : |\theta - \theta_0| \geq \delta\}$, define $U(M, \theta) = \max\{T(M, \theta), T(\theta^*)\}$. An application of Dini's theorem implies that $\sup_{|\theta - \theta_0| \geq \delta} U(M, \theta) \to T(\theta^*)$. Thus, for $M > M_1(\delta)$,

$$\sup_{|\theta - \theta_0| \geq \delta} T(M, \theta) \leq \sup_{|\theta - \theta_0| \geq \delta} U(M, \theta) \leq T(\theta^*) + \frac{a_\delta}{2} < T(\theta_0).$$

If Θ is not bounded, make one-to-one monotone continuous transformations so that the transformed parameter space

$$\Lambda = \{(\lambda_1, \ldots, \lambda_q) : \ \lambda_i = v_i(\theta_i), \ i = 1, \ldots, q, \ \theta \in \Theta\}$$

is a bounded subset of $\mathbb{R}^q$. □

In Ranneby *et al.* (2005), it was shown that $(\xi_1, \eta_{1,n})$ and $(\xi_2, \eta_{2,n})$ are asymptotically independent. In the proof of weak consistency, it is sufficient with asymptotic independence between $(\xi_1, \eta_{1,n})$ and $(\xi_2, \eta_{2,n})$, implying that the covariances tend to zero. To prove strong consistency, we need to show that the covariances are of order $1/n$, and here, the conditional approach of Schilling (1986) simplifies matters. Because we work only with the nearest neighbours, we can distinguish between the following five mutually exclusive sets for various nearest neighbour geometries of ξ_1 and ξ_2:

$$D_1 = \{\mathrm{NN}_1 = \xi_2, \ \mathrm{NN}_2 = \xi_1\}, \ D_2 = \{\mathrm{NN}_1 = \mathrm{NN}_2\}, \ D_3 = \{\mathrm{NN}_1 = \xi_2, \ \mathrm{NN}_2 \neq \xi_1\},$$

$$D_4 = \{\mathrm{NN}_1 \neq \xi_2, \ \mathrm{NN}_2 = \xi_1\}, \ D_5 = \{\mathrm{NN}_1 \neq \xi_2, \ \mathrm{NN}_2 \neq \xi_1, \ \mathrm{NN}_1 \neq \mathrm{NN}_2\}.$$

Lemma 1. *Let $\xi_1, \ldots, \xi_n$ be independent identically distributed d-vectors with distribution P_0 that is absolutely continuous with respect to Lebesgue measure. Then $P(D_k) = \mathcal{O}(n^{-1})$ for $k = 1, \ldots, 4$, and thus, $\lim_{n \to \infty} P(D_5) = 1$.*

© 2015 Board of the Foundation of the Scandinavian Journal of Statistics.

Proof. We have $P(D_1) < P(\mathrm{NN}_1 = \xi_2) = \mathcal{O}(n^{-1})$. Let $\mathrm{NN}_l^{(-m)}$ denote the nearest neighbour of ξ_l in the set without observation ξ_m. Then

$$P(D_2) = (n-2)P(\mathrm{NN}_1 = \xi_3|\mathrm{NN}_2 = \xi_3)P(\mathrm{NN}_2 = \xi_3)$$

$$< \tfrac{n-2}{n-1}P(\mathrm{NN}_1^{(-2)} = \xi_3|\mathrm{NN}_2 = \xi_3) = \tfrac{n-2}{n-1}P(\mathrm{NN}_1^{(-2)} = \xi_3) = \tfrac{1}{n-1}.$$

Since $P(D_3) = P(D_4) = 1/(n-1) - P(D_1)$, obviously $\lim_{n\to\infty} P(D_5) = 1$. □

Observe that in D_5, ξ_1 and ξ_2 have different neighbours, and they are not each other's neighbours. Therefore, $(\xi_1, \eta_{1,n})$ and $(\xi_2, \eta_{2,n})$ are conditionally independent given D_5. Since $P(D_5) \to 1$ according to lemma 1, asymptotic independence of $(\xi_1, \eta_{1,n})$ and $(\xi_2, \eta_{2,n})$ follows:

Lemma 2. *Let $\xi_1, \dots, \xi_n$ be independent identically distributed d-vectors with density function $g(x)$. Then $(\xi_1, \eta_{1,n})$ and $(\xi_2, \eta_{2,n})$ are asymptotically independent, that is, for any measurable sets $A_1, A_2 \subset \mathbb{R}^d \times \mathbb{R}^+$,*

$$\lim_{n\to\infty} P((\xi_1, \eta_{1,n}) \in A_1, (\xi_2, \eta_{2,n}) \in A_2) = \int_{A_1} dP(x,y) \int_{A_2} dP(x,y).$$

2. Weak consistency

In Ranneby *et al.* (2005), weak consistency of MSP estimates was proved for $h(x) = \ln x$ under the assumption that $g(x)$ belongs to the assigned model class. In this article, we give also conditions needed for consistency to hold when the assigned model class is not necessarily true. Because of the modified definition of the generalized MSP function $S_n(\theta)$, the proof of weak consistency can be simplified. To prove weak consistency of the GMSP estimate, we need an integrability condition that will be called W1 and W2, respectively, for the cases when the assigned model class is true and when it is not necessarily true.

Assumption W1. Suppose $g(x)$ belongs to the assigned model class, that is, there exists $\theta_0 \in \Theta$ such that $g(x) = f_{\theta_0}(x)$. Assume that $\int_0^\infty h^2(u)e^{-u}du < \infty$.

Assumption W2. Suppose the assigned model class is not necessarily true, that is, the density $g(x)$ does not have to belong to $\{f_\theta(x),\ \theta \in \Theta\}$. Let θ_0 be the parameter value that maximizes $T(\theta)$. Assume that

$$E\, h^2(z_{1,n}(\theta_0)) \to \int h^2(yf_{\theta_0}(x))dP(x,y) < \infty\,.$$

We can now state the main theorem of this section.

Theorem 1. *Let $\xi_1, \dots, \xi_n$ be a sequence of independent and identically distributed (i.i.d.) vectors in $\mathbb{R}^d$ with absolutely continuous distribution P_0 and density function $g(x)$. Suppose conditions C1 and C2 hold. Suppose in addition that (i) W1 holds and (ii) W2 holds. Under these assumptions, for both (i) and (ii), $\hat\theta_n \xrightarrow{p} \theta_0$.*

To prove weak consistency, we will show that $S_n(\theta_0) \xrightarrow{p} T(\theta_0)$ and that $T_n(M,\theta) \xrightarrow{p} T(M,\theta)$ uniformly in θ. These two results together with the identifiability condition then imply that $\hat\theta_n$ is a consistent estimator.

© 2015 Board of the Foundation of the Scandinavian Journal of Statistics.

Lemma 3. *Let $\xi_1, \dots, \xi_n$ be a sequence of i.i.d. vectors in $\mathbb{R}^d$ with absolutely continuous distribution P_0 and density function $g(x)$.*

(i) Suppose W1 holds. Then

$$S_n(\theta_0) = \frac{1}{n}\sum_{i=1}^{n} h(z_{i,n}(\theta_0)) \xrightarrow{p} \int h(u)e^{-u}du \,.$$

(ii) Suppose W2 holds. Then

$$S_n(\theta_0) \xrightarrow{p} \int h(yf_{\theta_0}(x))dP(x,y) = T(\theta_0)\,.$$

Proof. We prove (i), and the proof of (ii) is analogous. The exchangeability of $z_{i,n}(\theta_0)$ gives that $ES_n(\theta_0) = E\,h(z_{1,n}(\theta_0))$ and

$$\mathrm{Var}(S_n(\theta_0)) = \frac{1}{n}\mathrm{Var}(h(z_{1,n}(\theta_0))) + \frac{n-1}{n}\mathrm{Cov}(h(z_{1,n}(\theta_0)), h(z_{2,n}(\theta_0)))\,.$$

As $z_{1,n}(\theta_0) = nP_0(B_n(\xi_1))$ has the following density:

$$f_{z_{1,n}}(u) = \tfrac{n-1}{n}\left(1 - \tfrac{u}{n}\right)^{n-2}, \quad u \leq n\,,$$

the Lebesgue dominated convergence theorem implies

$$E\,h(z_{1,n}(\theta_0)) \to \int_0^{\infty} h(u)e^{-u}du\,, \quad E\,h^2(z_{1,n}(\theta_0)) \to \int_0^{\infty} h^2(u)e^{-u}du\,.$$

Because $h(z_{1,n}(\theta_0))$ is a function of $(\xi_1, \eta_{1,n})$, and $h(z_{2,n}(\theta_0))$, a function of $(\xi_2, \eta_{2,n})$, it follows from lemma 2 that they are asymptotically independent. This and the convergence of the first and second moment of $h(z_{1,n}(\theta_0))$ give that $\lim_{n\to\infty}\mathrm{Cov}(h(z_{1,n}(\theta_0)), h(z_{2,n}(\theta_0))) = 0$. Thus, the convergence of $S_n(\theta_0)$ in probability follows by Chebyshev's inequality.

For (ii) observe that for every n, $h(z_n(\theta_0, x, y)) < 1 + h^2(z_n(\theta_0, x, y))$ and

$$h(z_n(\theta_0, x_1, y_1))h(z_n(\theta_0, x_2, y_2)) < h^2(z_n(\theta_0, x_1, y_1)) + h^2(z_n(\theta_0, x_2, y_2))\,.$$

Thus, the convergence of $Eh(z_{1,n}(\theta_0))$ and $E[h(z_{1,n}(\theta_0))h(z_{2,n}(\theta_0))]$ follows because of W2 by the generalized Lebesgue dominated convergence theorem. □

Lemma 4. *Let $\xi_1, \dots, \xi_n$ be a sequence of i.i.d. vectors in $\mathbb{R}^d$ with absolutely continuous distribution P_0. Then for any $\theta \in \Theta$, $T_n(M,\theta) \xrightarrow{p} T(M,\theta)$. Under condition C1, the convergence is uniform in θ.*

Proof. The proof is analogous to the proof of lemma 3 in Ranneby *et al.* (2005), and will therefore not be repeated here. □

We can now complete the proof of theorem 1.

Proof. Consider the case (i), the proof is analogous for (ii). To prove the consistency of $\hat{\theta}_n$, we apply lemma 3 and lemma 4. Consider arbitrary $\delta > 0$. Choose $M_1(\delta)$ according to condition C2 and consider any $M > M_1$. Define $\varepsilon > 0$ as follows:

$$\varepsilon = T(\theta_0) - \sup_{\theta\in B^c(\theta_0,\delta)} T(M,\theta)\,. \tag{1}$$

© 2015 Board of the Foundation of the Scandinavian Journal of Statistics.

According to lemma 3, $\exists n_1(\varepsilon,\delta)$ such that $\forall n > n_1$,

$$P(|S_n(\theta_0) - T(\theta_0)| \geq \varepsilon/4) < \delta/2\,.$$

Lemma 4 implies that $\exists n_2(\varepsilon,\delta)$ such that $\forall n > n_2$,

$$P(|T_n(M,\hat{\theta}_n) - T(M,\hat{\theta}_n)| \geq \varepsilon/2) < \delta/2\,.$$

Choose n_3 so that $c_n < \varepsilon/4$ for $n > n_3$. Take $n^* = \max\{n_1, n_2, n_3\}$. Then $\forall n > n^*$, the following sequence of inequalities holds with probability larger than $1-\delta$:

$$T(\theta_0) - \varepsilon/2 \leq S_n(\theta_0) - c_n \leq S_n(\hat{\theta}_n) \leq T_n(M,\hat{\theta}_n) \leq T(M,\hat{\theta}_n) + \varepsilon/2\,. \tag{2}$$

Therefore, $T(M,\hat{\theta}_n) \geq T(\theta_0) - \varepsilon$ and the identifiability condition implies that $\hat{\theta}_n \in B(\theta_0,\delta)$. Thus, $\hat{\theta}_n \overset{p}{\to} \theta_0$ follows. □

3. Strong consistency

The general idea for proving strong consistency of the GMSP estimate is the same as in the case of weak consistency. We are going to show that $S_n(\theta_0) \overset{a.s.}{\to} T(\theta_0)$ and that $T_n(M,\theta) \overset{a.s.}{\to} T(M,\theta)$ uniformly in θ. This implies that there exists a null set N such that $\forall \omega \in \Omega \setminus N$, $S_n(\theta_0) \to T(\theta_0)$ and $T_n(M,\hat{\theta}_n) - T(M,\hat{\theta}_n) \to 0$ as $n \to \infty$. Consider arbitrary $\delta > 0$ and let $\varepsilon > 0$ be defined as in (1). Then $\forall \omega \in \Omega \setminus N$ there exists $n^*(\omega)$ such that $\forall n > n^*$, the sequence of inequalities given in (2) holds. This implies that $\hat{\theta}_n(\omega) \in B(\theta_0,\delta)$ and $\hat{\theta}_n(\omega) \to \theta_0$. Since this holds on a set of measure one, $\hat{\theta}_n \overset{a.s.}{\to} \theta_0$ follows. Therefore, to prove strong consistency, we have to show that

(I) $S_n(\theta_0) = \frac{1}{n}\sum_{i=1}^n h(z_{i,n}(\theta_0)) \overset{a.s.}{\to} T(\theta_0)$.

(II) $T_n(M,\theta) = \frac{1}{n}\sum_{i=1}^n t_M(z_{i,n}(\theta)) \overset{a.s.}{\to} T(M,\theta)$ for any $\theta \in \Theta$.

(III) $\frac{1}{n}\sum_{i=1}^n I((\xi_i,\eta_{i,n}) \in K_l) \overset{a.s.}{\to} P((X,Y) \in K_l)$, with K_l defined in C1.

Observe that (III) is needed for the uniform convergence of $T_n(M,\theta)$, see the proof of lemma 3 in Ranneby *et al.* (2005).

To prove strong consistency of the GMSP estimate, we need an additional integrability condition that we call S1 and S2, respectively, for the cases when the assigned model class is true and when it is not necessarily true.

Assumption S1. Suppose $g(x)$ belongs to the assigned model class, that is, there exists $\theta_0 \in \Theta$ such that $g(x) = f_{\theta_0}(x)$. Assume that

$$\int_0^\infty u^2(h'(u))^2 e^{-u}\,du < \infty.$$

Assumption S2. Suppose the assigned model class is not necessarily true, that is, the density $g(x)$ does not have to belong to $\{f_\theta(x),\ \theta \in \Theta\}$. Let θ_0 be the parameter value that maximizes $T(\theta)$, and let $\nu_n(\cdot)$ be the distribution of $z_{1,n}(\theta_0)$ with limit distribution $\nu(\cdot)$. Assume that

$$\int_0^\infty u^2(h'(u))^2\,d\nu_n(u) \to \int_0^\infty u^2(h'(u))^2\,d\nu(u) < \infty\,.$$

Theorem 2. *Let $\xi_1,\ldots,\xi_n$ be a sequence of i.i.d. vectors in $\mathbb{R}^d$ with absolutely continuous distribution P_0 and density function $g(x)$. Suppose conditions C1 and C2 hold. Suppose in addition*

© 2015 Board of the Foundation of the Scandinavian Journal of Statistics.

that (i) W1 and S1 hold and (ii) W2 and S2 hold. Under these assumptions, for both (i) and (ii), $\hat{\theta}_n \overset{a.s.}{\to} \theta_0$.

Proof. Denote every single term in the sums of (I)–(III) by $h_{i,n}$ and let $Eh_{i,n} = \mu_n$. Define S_n, S_n^* and H_n as

$$S_n = \sum_{i=1}^{n} h_{i,n}, \quad S_n^* = \sum_{i=1}^{n} (h_{i,n} - \mu_n), \quad H_n = \frac{S_n^*}{n}.$$

Since μ_n converges for (I)–(III), to prove the almost-sure convergences in (I)–(III), we need to show that $H_n \overset{a.s.}{\to} 0$ in all the three cases. Write H_n as

$$H_n = \frac{m^2}{n}\left(H_{m^2} + \frac{1}{m^2}(S_n^* - S_{m^2}^*)\right), \quad m^2 \le n < (m+1)^2.$$

Let $W_m = \max\{\frac{1}{m^2}|S_n^* - S_{m^2}^*| : m^2 \le n < (m+1)^2\}$. Then $|H_n| \le |H_{m^2}| + W_m$, and it is sufficient to show that

$$H_{m^2} \overset{a.s.}{\to} 0 \quad \text{and} \quad W_m \overset{a.s.}{\to} 0 \quad \text{as} \quad m \to \infty. \tag{3}$$

To prove (3), we use the Borel-Cantelli lemma and show that for every $\varepsilon > 0$,

$$\sum_{m=1}^{\infty} P(|H_{m^2}| > \varepsilon) \le \sum_{m=1}^{\infty} \frac{\operatorname{Var} H_{m^2}}{\varepsilon^2} < \infty,$$
$$\sum_{m=1}^{\infty} P(|W_m| > \varepsilon) \le \sum_{m=1}^{\infty} \frac{EW_m^2}{\varepsilon^2} < \infty.$$

For this to hold, we need that $\operatorname{Var} H_n = \mathcal{O}(n^{-1})$ and $E\,W_n^2 = \mathcal{O}(n^{-2})$. Lemma 5 and theorem 3 in the succeeding texts imply that these conditions are fulfilled under the assumptions of the theorem. □

An upper bound on $\operatorname{Var} H_n$ and $E\,W_n^2$ can be obtained through an upper bound on $\operatorname{Var} S_n$. We use the approach of Evans (2008), where the Efron-Stein jackknife inequality for the variance of symmetric statistics is used to obtain an upper bound on $\operatorname{Var} S_n$. Evans (2008) uses partly the work of Reitzner (2003) on random polytopes, where an upper bound on $\operatorname{Var} S_n$ is received by adding a sample point ξ_{n+1} and considering the difference between S_n and S_{n+1}.

Lemma 5. *Suppose* $E(\sum_{i=1}^{n}(h_{i,n} - h_{i,n+1}))^2 = \mathcal{O}(1)$. *Then* $\operatorname{Var} H_n = \mathcal{O}(n^{-1})$ *and* $E\,W_n^2 = \mathcal{O}(n^{-2})$.

Proof. Since S_n is invariant under permutations of its arguments, applying the Efron-Stein inequality to S_n (Evans, 2008, p. 3180-3181) gives

$$\operatorname{Var} H_n = \frac{1}{n^2}\operatorname{Var} S_n \le \frac{n+1}{n^2}E(S_n - S_{n+1})^2,$$

where $E(S_n - S_{n+1})^2 \le 2E\left(\sum_{i=1}^{n}(h_{i,n} - h_{i,n+1})\right)^2 + 2Eh_{n+1,n+1}^2$. For W_m, we obtain

$$W_m \le \frac{1}{m^2}\sum_{j=m^2+1}^{(m+1)^2-1} |S_j^* - S_{j-1}^*|, \quad \text{thus} \quad EW_m^2 \le \frac{2}{m^3}\sum_{j=m^2+1}^{(m+1)^2-1} E(S_j^* - S_{j-1}^*)^2.$$

For $E(S_n^* - S_{n+1}^*)^2$, we have

$$E(S_n^* - S_{n+1}^*)^2 \le 3E(S_n - S_{n+1})^2 + 3n^2(\mu_n - \mu_{n+1})^2 + 3\mu_{n+1}^2,$$

© 2015 Board of the Foundation of the Scandinavian Journal of Statistics.

where

$$(\mu_n - \mu_{n+1})^2 = \frac{1}{n^2}\left[E\left(\sum_{i=1}^{n}(h_{i,n} - h_{i,n+1})\right)\right]^2 \leq \frac{1}{n^2}E\left(\sum_{i=1}^{n}(h_{i,n} - h_{i,n+1})\right)^2.$$

For (II) and (III), $Eh_{i,n+1}^2 = \mathcal{O}(1)$ is trivial, and for (I), it is implied by W1 and W2. Thus, it follows that if $E\left(\sum_{i=1}^{n}(h_{i,n} - h_{i,n+1})\right)^2$ is bounded for large n, then $\mathrm{Var}\, H_n = \mathcal{O}(n^{-1})$ and $E\,W_n^2 = \mathcal{O}(n^{-2})$. □

To prove that $E\left(\sum_{i=1}^{n}(h_{i,n} - h_{i,n+1})\right)^2 = \mathcal{O}(1)$, we use exchangeability of $d_{i,n} = h_{i,n} - h_{i,n+1}$, $i = 1, \ldots, n$, and conditioning on the events $\{\xi_{n+1} \in B_n(\xi_i)\}$ and $\{\xi_{n+1} \notin B_n(\xi_i)\}$, $i \neq n+1$. If $\xi_{n+1} \notin B_n(\xi_i)$, then $B_{n+1}(\xi_i) = B_n(\xi_i)$. Observe that for (a) and (b) in the succeeding theorem, the variables $d_{i,n}$ are bounded. The proof of theorem 3 is given in the Appendix.

Theorem 3. *Let $\xi_1, \ldots, \xi_{n+1}$ be a sequence of i.i.d. vectors in $\mathbb{R}^d$ with absolutely continuous distribution P_0 and density function $g(x)$. Let $h_{i,n}$ be defined as*

(a) $h_{i,n} = \max\{-M, h(z_{i,n}(\theta))\}$.
(b) $h_{i,n} = I((\xi_i, \eta_{i,n}) \in K_l)$, *where K_l is defined in condition C1.*
(c) $h_{i,n} = h(z_{i,n}(\theta_0)) = h(nP_{\theta_0}(B_n(\xi_i)))$.

Then $E\left(\sum_{i=1}^{n}(h_{i,n} - h_{i,n+1})\right)^2 = \mathcal{O}(1)$. For (c), the result holds under the additional assumptions W1 and S1 when the assigned model class is true, and under W2 and S2 when the assigned model class is not necessarily true.

4. Model validation

In this section, we will show that for checking the validity of the assigned model class, it is useful to study the behaviour of the maximum spacing function under different divergence measures. Since $T(\theta) \leq T(\theta_0)$ (with equality if and only if $f_\theta(x) = g(x)$ a.s.), values of $S_n(\hat{\theta}_n)$ 'much' smaller than $T(\theta_0)$ give rise to doubts about the correctness of the assigned model. The asymptotic distribution of $\sqrt{n}S_n(\theta_0)$ can be used to obtain an indication of the assigned model class being wrong. Since $S_n(\hat{\theta}_n) \geq S_n(\theta_0)$, under $\sqrt{n}(S_n(\theta_0) - T(\theta_0)) \sim \mathcal{A}s\mathcal{N}(0, \sigma_h^2)$ it holds that

$$P\left(S_n(\hat{\theta}_n) \leq T(\theta_0) - z_{1-\alpha}\frac{\sigma_h}{\sqrt{n}}\right) \leq P\left(S_n(\theta_0) \leq T(\theta_0) - z_{1-\alpha}\frac{\sigma_h}{\sqrt{n}}\right) \simeq \alpha,$$

where $z_{1-\alpha}$ denotes the $(1-\alpha)$-quantile of the standard normal distribution. Therefore, values of $S_n(\hat{\theta}_n)$ smaller than $T(\theta_0) - z_{1-\alpha}\sigma_h/\sqrt{n}$ are unlikely to occur for large sample sizes.

In Zhou & Jammalamadaka (1993), asymptotic normality of $\sqrt{n}S_n(\theta_0)$ is derived through the convergence of the empirical process for the multivariate spacings. The asymptotic variance is given by $\sigma_h^2 = \int_0^\infty \int_0^\infty K(s,t)dh(s)dh(t)$, where

$$K(s,t) = e^{-t} - e^{-s-t}\left[1 - s + st - \int_{W(s,t)}(e^{\beta(s,t,w)} - 1)dw\right], \quad 0 \leq s \leq t \leq \infty,$$

$$W(s,t) = \{w \in \mathbb{R}^d : r_1 \leq |w| \leq r_1 + r_2\}, \quad \beta(s,t,w) = \int_{B(0,r_1)\cap B(w,r_2)} dz,$$

© 2015 Board of the Foundation of the Scandinavian Journal of Statistics.

with r_1 and r_2 corresponding to the volumes t and s of the balls $B(0, r_1)$ and $B(0, r_2)$, respectively.

Model validation example. We are going to demonstrate how the GMSP method can be used to discover that the assigned model class is wrong. Suppose, we believe our data come from a bivariate normal distribution with known covariance matrix I, that is $f_\mu(x) \in \mathcal{N}(\mu, I)$, but the true distribution is actually a normal mixture, so $g(x)$ is the density of

$$(1-\epsilon)\mathcal{N}(0, I) + \epsilon\mathcal{N}(\mu_0, I), \quad \epsilon \in (0, 1).$$

Under the true model,

$$\lim_{n\to\infty} E_g S_n(\mu_0) = T_g(\mu_0) = \int h(u)e^{-u} du\,.$$

Let $\hat{\mu}_n$ denote the GMSP estimate of μ, and let μ^* denote the parameter value that maximizes the limiting function of the expected value of the GMSP function, that is, μ^* maximizes

$$T(\mu) = \int h(y f_\mu(x)) g^2(x) e^{-y g(x)}\, dx\, dy = T_g(\mu_0) - a(\mu).$$

Thus, $a(\mu^*)$ presents the difference in $\lim_{n\to\infty} E S_n(\hat{\mu}_n)$ under the true distribution and the assigned model class. In Table 1, the limits of the expected value of the GMSP function are presented for the information measures h_1, h_2 and h_3 with $p = 2$. The parameter values μ^* that minimize $a(\mu)$ for h_1, h_2 and h_3 are the values that minimize the Kullback-Leibler information measure, Jeffreys' divergence measure and the Hellinger distance between g and f_μ, respectively. For h_1, the minimizing argument can be found analytically: $a(\mu)$ is minimized for $\mu^* = \epsilon\mu_0$. For Jeffreys' divergence and the Hellinger distance, we have minimized $a(\mu)$ numerically.

In Table 2, the values of μ^* and $a(\mu^*)$ are presented for h_1, h_2 and h_3 in the case of five values of μ_0 when $\epsilon = 0.1$. The models corresponding to different μ_0 are ordered according to increasing deviation from the mean of the mixture part with weight $1-\epsilon$.

The conditions of asymptotic normality of Zhou & Jammalamadaka (1993) are satisfied for g and for h_1, h_2 and h_3 of our example. The values of σ_{h_1}, σ_{h_2} and σ_{h_3} are presented in Table 3. We can observe that μ^* is always situated on the line that connects the expected values

Table 1. *Limits of the expected value of the GMSP function for different information measures*

	$h_1(x) = \ln x - x + 1$	$h_2(x) = (1-x)\ln x$	$h_3(x) = -(1-\sqrt{x})^2$
$T(\mu)$	$-\gamma - a(\mu)$	$-1 - a(\mu)$	$-2 + \sqrt{\pi} - a(\mu)$
$a(\mu)$	$\int g \ln(\frac{g}{f_\mu})$	$\int (g - f_\mu)\ln(\frac{g}{f_\mu})$	$\sqrt{\pi}(1 - \int \sqrt{f_\mu g})$

Table 2. *Values of* μ^* *and* $a(\mu^*)$ *for* h_1, h_2 *and* h_3 *when* $\epsilon = 0.1$

	h_1		h_2		h_3	
μ_0	μ^*	$a(\mu^*)$	μ^*	$a(\mu^*)$	μ^*	$a(\mu^*)$
M(1,2)	(0.1, 0.2)	0.058	(0.084, 0.167)	0.092	(0.076, 0.152)	0.018
M(2,3)	(0.2, 0.3)	0.311	(0.126, 0.188)	0.411	(0.062, 0.093)	0.058
M(1,4)	(0.1, 0.4)	0.469	(0.058, 0.231)	0.592	(0.019, 0.075)	0.070
M(4,4)	(0.4, 0.4)	1.119	(0.205, 0.205)	1.299	(0.011, 0.011)	0.088
M(1,8)	(0.1, 0.8)	2.600	(0.050, 0.400)	2.868	(0.0001, 0.0004)	0.091

© 2015 Board of the Foundation of the Scandinavian Journal of Statistics.

of the two mixture parts and that the Hellinger distance is least sensitive to deviations from the assigned model. When the true distribution does not belong to the assigned model, the GMSP estimates for the different divergence measures converge to different values opposite to the situation with a correct model specification. Thus, if more than one divergence measure is used, and there is an apparent difference between the estimates, we have an indication of a misspecification of the model. A more direct indication would be to look at the value of the spacing statistic. For large n, we expect $S_n(\hat{\mu}_n) < T(\mu_0) - 1.64\sigma_h/\sqrt{n}$ to hold with an approximate probability of at most 0.05 under the true model. Thus, if the considered model class is wrong and $a(\mu^*)$ large enough, $a(\mu^*) > 1.64\sigma_h/\sqrt{n}$ at least, then our $S_n(\hat{\mu}_n)$ would probably be smaller than the aforementioned bound. Therefore, the comparison of $1.64\sigma_h/\sqrt{n}$ with $a(\mu^*)$ for different n can indicate what sample sizes are needed to give us a possibility to detect violations from the assumed model class under different information measures. Comparing the values in Table 2 and Table 3, we can see that for h_1 and h_2, the sample size $n = 100$ would be enough to make it possible to detect a misspecification for all the models except M(1,2). For the Hellinger divergence, larger sample sizes are needed. To discover a small deviation like in model M(1,2), even the sample size $n = 400$ would not be enough. The comparisons are based on the asymptotic distributions, although the situation may be different for small and moderate sample sizes.

To illustrate how this theoretical reasoning based on asymptotics works in practice, we performed the following simulation study. For every model in Table 2 and in addition for the true model, we generated one hundred data sets of size $n = 100$. For every data set, we calculated the parameter estimate $\hat{\mu}$ and the value of the spacing function $S_{100}(\hat{\mu})$ under h_1, h_2 and h_3. In Table 4, the mean values of $\hat{\mu}$ for each model are presented. Columns '$< b_i$' give the number of data sets with the value of the spacing function lower than $b_i = T_g(\mu_0) - 0.164\sigma_{h_i}$. Model M(0,0) stands for the true model. The simulation study confirms that for h_1 and h_2 and models M(2,3)–M(1,8), there is a fair chance to detect the misspecification. We can also see that when the model is correct, the GMSP estimates based on the different information measures are close to each other, while the difference between them increases with increasing deviation

Table 3. *Values of* $1.64\sigma_h/\sqrt{n}$ *for information measures* h_1, h_2 *and* h_3 *for different sample sizes* n

n	$\sigma_{h_1} = 1.0977$	$\sigma_{h_2} = 1.6904$	$\sigma_{h_3} = 0.3540$
100	0.180	0.277	0.058
200	0.127	0.196	0.041
400	0.090	0.139	0.029

Table 4. *Results from the simulation study with* $n = 100$. *Here,* $b_1 = -0.757$, $b_2 = -1.277$ *and* $b_3 = -0.286$

	h_1		h_2		h_3	
μ_0	mean($\hat{\mu}$)	$< b_1$	mean($\hat{\mu}$)	$< b_2$	mean($\hat{\mu}$)	$< b_3$
M(0,0)	(0.009, 0.005)	4	(0.008, 0.008)	3	(0.007, 0.009)	2
M(1,2)	(0.086, 0.168)	8	(0.077, 0.143)	8	(0.073, 0.132)	8
M(2,3)	(0.210, 0.281)	61	(0.150, 0.179)	54	(0.085, 0.083)	31
M(1,4)	(0.109, 0.396)	80	(0.079, 0.255)	72	(0.036, 0.095)	39
M(4,4)	(0.421, 0.426)	99	(0.253, 0.254)	95	(0.034, 0.024)	56
M(1,8)	(0.086, 0.879)	100	(0.037, 0.537)	100	($\mathbf{-0.042}$, 0.030)	61

© 2015 Board of the Foundation of the Scandinavian Journal of Statistics.

from the true model. Furthermore, it is obvious that the Hellinger distance is robust against the studied misspecifications.

It can be concluded that to detect deviations from the assigned model, preferably spacing functions based both on the Kullback-Leibler information and the Hellinger distance should be used, and the values of the spacing functions as well as the parameter estimates should be compared.

Acknowledgements

The authors would like to thank the reviewers for the comments and suggestions that helped to improve the article. The research of K. Kuljus was supported by Swedish University of Agricultural Sciences financing project KON 50210.

References

Cheng, R. C. & Amin, N. A. (1983). Estimating parameters in continuous univariate distributions with a shifted origin. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **45**, (3), 394–403.

Ekström, M. (1997). Strong consistency of the maximum spacing estimate. *Theory Probab. Math. Stat.* **55**, 55–73.

Ekström, M. (1998). On the consistency of the maximum spacing method. *J. Statist. Plann. Inference* **70**, (2), 209–224.

Ekström, M. (2001). Consistency of generalized maximum spacing estimates. *Scand. J. Stat.* **28**, 343–354.

Evans, D. (2008). A law of large numbers for nearest neighbour statistics. *Proc. R. Soc. A* **464**, 3175–3192.

Evans, L. & Gariepy, R. (1992). *Measure theory and fine properties of functions*, CRC Press, Boca Raton.

Ghosh, K. & Jammalamadaka, S. R. (2001). A general estimation method using spacings. *J. Statist. Plann. Inference* **93**, (1-2), 71–82.

Jimenez, R. & Yukich, J. E. (2002). Asymptotics for statistical distances based on Voronoi tessellation. *J. Theor. Probab.* **15**, (2), 503–541.

Penev, S. & Ruderman, A. (2011). On the behaviour of tests based on sample spacings for moderate samples. *J. Statist. Plann. Inference* **141**, (3), 1240–1249.

Ranneby, B. (1984). The maximum spacing method. An estimation method related to the maximum likelihood method. *Scand. J. Stat.* **11**, (2), 93–112.

Ranneby, B. (1996). Spatial and temporal models in contextual classification. In *Spatial accuracy assessment in natural resources and environmental sciences: second international symposium* (eds Mowrer, H. T., Czaplewski, R. L. & Hamre, R. H.), Fort Collins, Colorado; 451–458.

Ranneby, B. & Ekström, M. (1997). Maximum spacing estimates based on different metrics. Research Report, Umeå University.

Ranneby, B., Jammalamadaka, S. R. & Teterukovskiy, A. (2005). The maximum spacing estimation for multivariate observations. *J. Statist. Plann. Inference* **129**, (1-2), 427–446.

Reitzner, M. (2003). Random polytopes and the Efron-Stein jackknife inequality. *Ann. Probab.* **31**, (4), 2136–2166.

Royden, H. L. (1968). *Real analysis*, Macmillan, New York.

Schilling, M. F. (1986). Mutual and shared neighbor probabilities: finite- and infinite-dimensional results. *Adv. Appl. Prob.* **18**, (2), 388–405.

Shao, Y. & Hahn, M. G. (1999). Strong consistency of the maximum product of spacings estimates with applications in nonparametrics and in estimation of unimodal densities. *Ann. Inst. Statist. Math.* **51**, (1), 31–49.

Zhou, S. & Jammalamadaka, S. R. (1993). Goodness of fit in multidimensions based on nearest neighbour distance. *Nonparametric Statistics* **2**, 271–284.

Received March 2014, in final form March 2015

Kristi Kuljus, Department of Mathematics and Mathematical Statistics, UmeåUniversity, 901 87 Umeå, Sweden.
E-mail: kristi.kuljus@math.umu.se

© 2015 Board of the Foundation of the Scandinavian Journal of Statistics.

Appendix

Lemma 6. *The density function of* $z_{1,n+1}(\theta)$ *for arbitrary* $\theta \in \Theta$ *(* $f_\theta \neq g$ *) is for* $0 < u < n+1$ *given by*

$$v_{n+1}(\theta,u) = \int g(x)\left(1-\frac{u\frac{g(x)}{f_\theta(x)+\varepsilon_n(x)}}{n+1}\right)^{n-1}\frac{g(x)}{f_\theta(x)+\varepsilon_n(x)}\frac{n}{n+1}dx\,, \tag{4}$$

where $\varepsilon_n(x) \to 0$ *as* $n \to \infty$.

Proof. We start with deriving the distribution function of $z_{1,n+1}(\theta)$:

$$P(z_{1,n+1}(\theta) > u) = \int g(x)P\left((n+1)P_\theta(B(x,R_{n+1}(1))) > u\right)\,dx\,.$$

Observe that $(n+1)P_\theta(B(x,R_{n+1}(1))) > u$ holds if and only if all the observations $\xi_2,\ldots,\xi_{n+1}$ fall outside the ball $B(x,r_n(\theta))$, where $r_n(\theta)$ satisfies $P_\theta(B(x,r_n(\theta))) = u/(n+1)$. Recall that P_0 denotes the probability measure for the true distribution. The Lebesgue-Besicovitch differentiation theorem (Evans & Gariepy, 1992, p. 43) gives

$$\frac{1}{P_0(B(x,r_n(\theta)))}\int_{B(x,r_n(\theta))}\frac{dP_\theta}{dP_0}(y)dP_0 \to \frac{dP_\theta}{dP_0}(x) = \frac{f_\theta(x)}{g(x)}\,.$$

Thus, we obtain

$$P_0(B(x,r_n(\theta))) = \frac{u}{n+1}\frac{g(x)}{f_\theta(x)+\varepsilon_n(x)},$$

where $\varepsilon_n(x) \to 0$. Therefore,

$$P(z_{1,n+1}(\theta) > u) = \int g(x)\left(1-\frac{u}{n+1}\frac{g(x)}{f_\theta(x)+\varepsilon_n(x)}\right)^n dx\,.$$

Differentiation under the integral sign gives the density function in (4). An application of the Fubini-Tonelli theorem verifies that $v_{n+1}(\theta,u)$ is a density function for $z_{1,n+1}(\theta)$. □

Let $A_l \subset \mathbb{R}^d$ and b, b_1 and b_2 be positive constants.

Lemma 7. *Consider* $K_{l,n} = A_l \times \left[\frac{n}{n+1}b, b\right]$. *Then*

$$P((\xi_1,\eta_{1,n}) \in K_{l,n}) = \mathcal{O}(n^{-1}). \tag{5}$$

Proof. Let $V^{-1}(y/n)$ denote the radius of the ball with volume y/n. Note that the volume of the nearest neighbour ball $B_n(\xi_1)$ for fixed $\xi_1 = x$ exceeds y/n if and only if none of the variables $\xi_2,\ldots,\xi_n$ falls in the ball $B(x,V^{-1}(y/n))$. Thus we obtain:

$$nP((\xi_1,\eta_{1,n}) \in K_{l,n}) = n\int_{x\in A_l^\circ} P(n||B_n(\xi_1)|| \in [\tfrac{n}{n+1}b,b]\,|\,\xi_1 = x)g(x)\,dx$$

$$= n\int_{x\in A_l^\circ}\left[[1-P_0(B(x,V^{-1}(\tfrac{b}{n+1})))]^{n-1} - [1-P_0(B(x,V^{-1}(\tfrac{b}{n})))]^{n-1}\right]g(x)\,dx.$$

Let $z_1 = P_0(B(x,V^{-1}(\frac{b}{n+1})))$ and $z_2 = P_0(B(x,V^{-1}(\frac{b}{n})))$. Denote the radii of the balls with volume $\frac{b}{n+1}$ and $\frac{b}{n}$ with r_n^* and r_n, respectively. Then for some $\tilde{z} \in [z_1,z_2]$,

$$(1-z_1)^{n-1} - (1-z_2)^{n-1} = -(n-1)(1-\tilde{z})^{n-2}(z_1-z_2) \le (n-1)(z_2-z_1).$$

© 2015 Board of the Foundation of the Scandinavian Journal of Statistics.

According to Lusin's theorem (see e.g. p. 15 in Evans & Gariepy, 1992), we can choose the set A_l so that it is compact and g is continuous on A_l. Thus, $g(x) \le c$ on A_l for some $c > 0$, and we have for n large enough that

$$z_2 - z_1 = \int_{B(x,r_n)} g(y)dy - \int_{B(x,r_n^*)} g(y)dy = \int_{B(x,r_n)\cap B^c(x,r_n^*)} g(y)dy$$

$$\le c||B(x,r_n) \cap B^c(x,r_n^*)|| = c\left(\tfrac{b}{n} - \tfrac{b}{n+1}\right) = \frac{bc}{n(n+1)}.$$

Thus, it follows that $nP((\xi_1, \eta_{1,n}) \in K_{l,n}) \le bc \int_{x\in A_l^\circ} g(x)dx \le bc$. □

Lemma 8. *Let* $K_{l,n} = A_l \times \left[\tfrac{n}{n+1}b_1, b_1\right] \cup \left[\tfrac{n}{n+1}b_2, b_2\right]$. *Then*

$$P((\xi_1, \eta_{1,n}) \in K_{l,n}, (\xi_2, \eta_{2,n}) \in K_{l,n}) = \mathcal{O}(n^{-2}). \tag{6}$$

Proof. Consider the five mutually exclusive nearest neighbour relationships $D_1, \dots, D_5$ on p. 4. The probability in (6) can then be written as

$$\sum_{k=1}^{5} P((\xi_1, \eta_{1,n}) \in K_{l,n}, (\xi_2, \eta_{2,n}) \in K_{l,n} \mid D_k)P(D_k). \tag{7}$$

Note that each conditional probability in (7) denotes the common value of the respective probability in the group. Consider the conditional probability for D_1, the cases D_2, D_3, D_4 are analogous. Since $\xi_1, \dots, \xi_n$ are exchangeable,

$$P((\xi_1, \eta_{1,n}) \in K_{l,n}, (\xi_2, \eta_{2,n}) \in K_{l,n} \mid D_1) \le P((\xi_1, \eta_{1,n}) \in K_{l,n} \mid D_1)$$

$$= P((\xi_1, \eta_{1,n}) \in K_{l,n} \mid \mathrm{NN}_1 = \xi_2) = P((\xi_1, \eta_{1,n}) \in K_{l,n}).$$

For D_5, the conditional independence of $(\xi_1, \eta_{1,n})$ and $(\xi_2, \eta_{2,n})$ and exchangeability imply

$$P((\xi_1, \eta_{1,n}) \in K_{l,n}, (\xi_2, \eta_{2,n}) \in K_{l,n} \mid D_5) = P((\xi_1, \eta_{1,n}) \in K_{l,n})P((\xi_2, \eta_{2,n}) \in K_{l,n}).$$

According to lemma 1, $P(D_k) = \mathcal{O}(n^{-1})$, $k = 1, \dots, 4$. This and lemma 7 together imply (6). □

Lemma 9. *Let* $\xi_1, \dots, \xi_{n+1}$ *be a sequence of i.i.d. vectors in* $\mathbb{R}^d$ *with density* $g(x)$. *Suppose that (i) S1 holds and (ii) S2 holds. Then for both (i) and (ii),*

$$E[h(\tfrac{n}{n+1}z_{i,n+1}(\theta_0)) - h(z_{i,n+1}(\theta_0))]^2 = \mathcal{O}(n^{-2}).$$

Proof. We prove the statement for (i), the idea of the proof is the same for (ii). Let $b_i = h(z_{i,n+1}(\theta_0)) - h(\tfrac{n}{n+1}z_{i,n+1}(\theta_0))$. Because the density of $z_{i,n+1}(\theta_0)$ is under the true model given by

$$f_{z_{i,n+1}}(u) = \tfrac{n}{n+1}\left(1 - \tfrac{u}{n+1}\right)^{n-1}, \quad u \le n+1,$$

we obtain

$$Eb_i^2 = \tfrac{n}{n+1}\int_0^{n+1}\left(h(u) - h\left(\tfrac{n}{n+1}u\right)\right)^2 (1 - \tfrac{u}{n+1})^{n-1}du = \tfrac{n}{n+1}I.$$

© 2015 Board of the Foundation of the Scandinavian Journal of Statistics.

Let h'_- and h'_+ denote the left and right derivative of h, respectively. Because h is concave, the following sequence of inequalities holds $\forall u \in (0, n+1)$:

$$h'_+(u) \le h'_-(u) \le \frac{h(u) - h(\frac{n}{n+1}u)}{\frac{u}{n+1}} \le h'_+(\tfrac{n}{n+1}u) \le h'_-(\tfrac{n}{n+1}u).$$

Because h has its maximum at $u = 1$, h'_- (and h'_+) is positive when $u < 1$ and negative when $u > 1$. Thus, the following relationships hold:

$$\frac{(h(u) - h(\frac{n}{n+1}u))^2}{\frac{u^2}{(n+1)^2}} \le \begin{cases} (h'_-(\frac{n}{n+1}u))^2, & \text{if} \quad u < 1\,, \\ (h'_-(\frac{n}{n+1}u))^2 + (h'_-(u))^2, & \text{if} \quad 1 \le u < \frac{n+1}{n}\,, \\ (h'_-(u))^2, & \text{if} \quad u \ge \frac{n+1}{n}. \end{cases}$$

Therefore,

$$\begin{aligned} I &\le \int_0^{(n+1)/n} \tfrac{u^2}{(n+1)^2} (h'_-(\tfrac{n}{n+1}u))^2 (1 - \tfrac{u}{n+1})^{n-1} du + \int_1^{n+1} \tfrac{u^2}{(n+1)^2} (h'_-(u))^2 \left(1 - \tfrac{u}{n+1}\right)^{n-1} du \\ &= \int_0^1 \tfrac{z^2}{n^2} (h'_-(z))^2 (1 - \tfrac{z}{n})^{n-1} \tfrac{n+1}{n}\, dz + \int_1^{n+1} \tfrac{u^2}{(n+1)^2} (h'_-(u))^2 (1 - \tfrac{u}{n+1})^{n-1}\, du. \end{aligned}$$

Because a concave function is differentiable almost everywhere, we obtain

$$\tfrac{n}{n+1} I \le \tfrac{1}{n^2} \int_0^{n+1} u^2 (h'(u))^2 (1 - \tfrac{u}{n+1})^{n-1} du. \tag{8}$$

The last integral converges by the Lebesgue dominated convergence theorem to $\int_0^\infty u^2 (h'(u))^2 e^{-u} du$ as $n \to \infty$. Thus, it follows that $Eb_i^2 = \mathcal{O}(n^{-2})$.

For (ii), the density of $z_{i,n+1}(\theta_0)$ is given by $v_{n+1}(\theta_0, u)$ in lemma 6. The required convergence in (8) follows because of the generalized Lebesgue dominated convergence theorem and assumption S2. □

Proof of theorem 3.

Proof.

(a) Let $d_{i,n} = h_{i,n} - h_{i,n+1}$ and $d^*_{i,n} = h_{i,n} - \max\{-M, h(\frac{n+1}{n} z_{i,n}(\theta))\}$. By exchangeability, $E(\sum_{i=1}^n d_{i,n})^2 = nEd^2_{1,n} + n(n-1)E(d_{1,n}d_{2,n})$. Therefore, we need to show that $Ed^2_{1,n} = \mathcal{O}(n^{-1})$ and $E(d_{1,n}d_{2,n}) = \mathcal{O}(n^{-2})$. Observe that $Ed^2_{1,n} \le 2ME|d_{1,n}|$. Denote the event $\{\xi_{n+1} \in B_n(\xi_1)\}$ by $D_{n+1}(\xi_1)$, then

$$E|d_{1,n}| = E[|d_{1,n}|\,|D_{n+1}(\xi_1)]P(D_{n+1}(\xi_1)) + E[|d_{1,n}|\,|\,D^c_{n+1}(\xi_1)]P(D^c_{n+1}(\xi_1))$$

$$\le 2M/n + E|d^*_{1,n}|.$$

Let $\mu_n(x)$ denote the probability measure of $z_{1,n}(\theta)$. Because h is concave, there exist $M_1, M_2 > 0$, $M_1 < M_2$, such that $h^{-1}(-M) = M_1$ and $h^{-1}(-M) = M_2$. Let L_M denote the Lipschitz constant of $\max\{-M, h(x)\}$. Then

$$E|d^*_{1,n}| = \int_0^{M_2} |\max\{-M, h(x)\} - \max\{-M, h(\tfrac{n+1}{n}x)\}| d\mu_n(x) \le \frac{L_M M_2}{n}. \tag{9}$$

Thus, $Ed^2_{1,n} = \mathcal{O}(n^{-1})$ follows.

To calculate $E|d_{1,n}d_{2,n}|$, consider the following mutually exclusive events:

$$J_1 = \{\xi_{n+1} \in B_n(\xi_1), \xi_{n+1} \in B_n(\xi_2)\}, \quad J_2 = \{\xi_{n+1} \notin B_n(\xi_1), \xi_{n+1} \in B_n(\xi_2)\},$$

$$J_3 = \{\xi_{n+1} \in B_n(\xi_1), \xi_{n+1} \notin B_n(\xi_2)\}, \quad J_4 = \{\xi_{n+1} \notin B_n(\xi_1), \xi_{n+1} \notin B_n(\xi_2)\}.$$

© 2015 Board of the Foundation of the Scandinavian Journal of Statistics.

Then $E|d_{1,n}d_{2,n}| = \sum_{k=1}^{4} E[|d_{1,n}d_{2,n}| \,|\, J_k]P(J_k)$. Because $P(\mathrm{NN}_1 = \mathrm{NN}_2) = \mathcal{O}(n^{-1})$, we have $P(J_1) = \mathcal{O}(n^{-2})$. Because $P(\xi_{n+1} \in B_n(\xi_1)) = 1/n$, for $k = 2, 3$, it holds that $P(J_k) \leq 1/n$. Observe that

$$E[|d_{1,n}d_{2,n}| \,|\, J_2] \leq 2ME[|d_{1,n}| \,|\, J_2] = 2ME|d_{1,n}^*| = \mathcal{O}(n^{-1}).$$

Analogously, $E[|d_{1,n}d_{2,n}| \,|\, J_3] = \mathcal{O}(n^{-1})$. For J_4, $E[|d_{1,n}d_{2,n}| \,|\, J_4] = E|d_{1,n}^* d_{2,n}^*|$. Let $\mu_n(x, y)$ denote the probability measure of $(z_{1,n}(\theta), z_{2,n}(\theta))$. Then analogously to (9), we obtain $E|d_{1,n}^* d_{2,n}^*| \leq L_M^2 M_2^2/n^2$. Thus, $E|d_{1,n}d_{2,n}| = \mathcal{O}(n^{-2})$ follows.

(b) Here $h_{i,n}$ is the indicator function $I((\xi_i, \eta_{i,n}) \in K_l)$, where $\eta_{i,n} = n||B_n(\xi_i)||$ and $K_l = A_l \times [b_{l,1}, b_{l,2}]$. Let again $d_{i,n} = h_{i,n} - h_{i,n+1}$. Because of the exchangeability property, we again need to study only $Ed_{1,n}^2$ and $E|d_{1,n}d_{2,n}|$. Observe that $|d_{1,n}| = 1$ only in the following cases:

$$(1)\ h_{1,n} = 1,\ h_{1,n+1} = 0 \iff \begin{cases} (1a)\ \xi_1 \in A_l,\ \eta_{1,n} \in [b_{l,1}, b_{l,2}],\ \eta_{1,n+1} < b_{l,1}, \\ (1b)\ \xi_1 \in A_l,\ \eta_{1,n} \in [b_{l,1}, b_{l,2}],\ \eta_{1,n+1} > b_{l,2}; \end{cases}$$

$$(2)\ h_{1,n} = 0,\ h_{1,n+1} = 1 \iff \begin{cases} (2a)\ \xi_1 \in A_l,\ \eta_{1,n} < b_{l,1},\ \eta_{1,n+1} \in [b_{l,1}, b_{l,2}], \\ (2b)\ \xi_1 \in A_l,\ \eta_{1,n} > b_{l,2},\ \eta_{1,n+1} \in [b_{l,1}, b_{l,2}]. \end{cases}$$

If $\xi_{n+1} \notin B_n(\xi_1)$, then $B_n(\xi_1) = B_{n+1}(\xi_1)$ and $\eta_{1,n} < \eta_{1,n+1}$. Thus, in this case, events (1a) and (2b) are impossible. Observe that under the condition $\xi_{n+1} \notin B_n(\xi_1)$, (1b) and (2a) can be jointly written as

$$(\xi_1, \eta_{1,n}) \in K_{l,n}, \quad \text{where} \quad K_{l,n} = A_l \times [\tfrac{n}{n+1}b_{l,1}, b_{l,1}] \cup [\tfrac{n}{n+1}b_{l,2}, b_{l,2}].$$

As in (a), we can condition on $D_{n+1}(\xi_1)$. Then

$$Ed_{1,n}^2 = E[d_{1,n}^2 \,|\, D_{n+1}(\xi_1)]P(D_{n+1}(\xi_1)) + E[d_{1,n}^2 \,|\, D_{n+1}^c(\xi_1)]P(D_{n+1}^c(\xi_1))$$

$$< 1/n + P((\xi_1, \eta_{1,n}) \in K_{l,n} \,|\, D_{n+1}^c(\xi_1)) = 1/n + P((\xi_1, \eta_{1,n}) \in K_{l,n}) \overset{(5)}{=} \mathcal{O}(n^{-1}).$$

To calculate $E|d_{1,n}d_{2,n}|$, condition again on $J_1, \ldots, J_4$. Because $P(J_1) = \mathcal{O}(n^{-2})$ and $|d_{1,n}d_{2,n}| \leq 1$, we need to consider $E[|d_{1,n}d_{2,n}| \,|\, J_k]$ for $k = 2, 3, 4$, where the cases $k = 2$ and $k = 3$ are analogous. We have

$$\begin{aligned} E[|d_{1,n}d_{2,n}| \,|\, J_2] \leq E[|d_{1,n}| \,|\, J_2] &= P((\xi_1, \eta_{1,n}) \in K_{l,n} \,|\, J_2) \\ &= P((\xi_1, \eta_{1,n}) \in K_{l,n}) \overset{(5)}{=} \mathcal{O}(n^{-1}), \end{aligned}$$

$$\begin{aligned} E[|d_{1,n}d_{2,n}| \,|\, J_4] &= P((\xi_1, \eta_{1,n}) \in K_{l,n}, (\xi_2, \eta_{2,n}) \in K_{l,n} \,|\, J_4) \\ &= P((\xi_1, \eta_{1,n}) \in K_{l,n}, (\xi_2, \eta_{2,n}) \in K_{l,n}) \overset{(6)}{=} \mathcal{O}(n^{-2}). \end{aligned}$$

Because $P(J_2) \leq 1/n$ and $P(J_3) \leq 1/n$, $E|d_{1,n}d_{2,n}| = \mathcal{O}(n^{-2})$ follows.

(c) Here, $h_{i,n} = h(z_{i,n}(\theta_0)) = h(nP_{\theta_0}(B_n(\xi_i)))$. Write $h_{i,n} - h_{i,n+1}$ as

$$h_{i,n} - h_{i,n+1} = (h_{i,n} - \tilde{h}_{i,n+1}) + (\tilde{h}_{i,n+1} - h_{i,n+1}) = a_i + b_i,$$

where $\tilde{h}_{i,n+1} = h(\frac{n}{n+1}z_{i,n+1}(\theta_0)) = h(nP_{\theta_0}(B_{n+1}(\xi_i)))$. Then

© 2015 Board of the Foundation of the Scandinavian Journal of Statistics.

$$\left(\sum_{i=1}^{n}(h_{i,n}-h_{i,n+1})\right)^2 \leq 2\left(\sum_{i=1}^{n} a_i\right)^2 + 2\left(\sum_{i=1}^{n} b_i\right)^2 .$$

We are going to show that the expectation of both terms in the previous sum is bounded above for large n. Consider the situation when adding a sample point. If $\xi_{n+1} \notin B_n(\xi_i)$, then $B_n(\xi_i) = B_{n+1}(\xi_i)$ and $a_i = 0$. Let M_{n+1} be the stochastic set containing the indices of the sample points, which have the new point ξ_{n+1} as the nearest neighbour: $M_{n+1} = \{i : \xi_{n+1} \in B_n(\xi_i)\}$. According to lemma 4.2 in Evans (2008), ξ_{n+1} can be the nearest neighbour of at most $\beta = \lfloor 2\pi^{d/2}/\,\Gamma(d/2)\rfloor$ points of the set. Thus, the Cauchy-Schwarz inequality implies that for a fixed $\omega \in \Omega$,

$$\left(\sum_{i=1}^{n} a_i\right)^2 = \left(\sum_{i\in M_{n+1}(\omega)} a_i\right)^2 \leq \sum_{i\in M_{n+1}(\omega)} 1 \sum_{i\in M_{n+1}(\omega)} a_i^2 \leq \beta \sum_{i=1}^{n} a_i^2.$$

Since this holds for every ω, we have $E\left(\sum_{i=1}^{n} a_i\right)^2 \leq \beta \sum_{i=1}^{n} Ea_i^2$. But

$$Ea_i^2 \leq \frac{2}{n}\left[E(h_{i,n}^2|\xi_{n+1} \in B_n(\xi_i)) + E(\tilde{h}_{i,n+1}^2|\xi_{n+1} \in B_n(\xi_i))\right]$$

$$= \frac{2}{n}[E(h_{i,n}^2) + E(\tilde{h}_{i,n+1})^2],$$

where the last equality holds because $h_{i,n}^2$ is determined by $\xi_1, \dots, \xi_n$ and is thus independent of ξ_{n+1}, and because of exchangeability of $z_{i,n+1}(\theta_0)$, $i = 1, \dots, n+1$. Under the assumptions of the theorem, $\lim_{n\to\infty} Eh_{i,n}^2 < \infty$. Since $\lim_{n\to\infty} Eh_{i,n}^2 = \lim_{n\to\infty} E\tilde{h}_{i,n+1}^2$ (follows because of the expression of the density for $z_{i,n}(\theta_0)$, see the proof of lemma 9), $E\left(\sum_{i=1}^{n} a_i\right)^2 = \mathcal{O}(1)$ follows. For $(\sum_{i=1}^{n} b_i)^2$, applying the Cauchy-Schwarz inequality and lemma 9 gives

$$E\left(\sum_{i=1}^{n} b_i\right)^2 \leq nEb_1^2 + n(n-1)(Eb_1^2)^{1/2}(Eb_2^2)^{1/2} = \mathcal{O}(1).$$

□

© 2015 Board of the Foundation of the Scandinavian Journal of Statistics.