Estimators for models with constraints involving unknown parameters

Ursula U. Müller
Universität Bremen

Wolfgang Wefelmeyer
Universität Siegen

Abstract

Suppose we have independent observations from a distribution which we know to fulfill a finite-dimensional linear constraint involving an unknown finite-dimensional parameter. We construct efficient estimators for finite-dimensional functionals of the distribution. The estimators are obtained by first constructing an efficient estimator for the functional when the parameter is known, and then replacing the parameter by an efficient estimator. We consider in particular estimation of expectations.

AMS 2000 subject classifications. Primary 62G05, 62G20.

Key words and Phrases. Plug-in-estimator, estimating equation, method of moments, coefficient of variation, empirical likelihood, minimum discriminant information adjustment.

1 Introduction

To begin we recall some results for simpler models. Let $X_1, \ldots, X_n$ be independent observations with unknown distribution $P(dx)$. Suppose we want to estimate the expectation $Pf = Ef(X) = \int f(x)\, P(dx)$ of some real-valued P-square-integrable function $f(x)$. A natural estimator is the empirical estimator

$$\hat{P}f = \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

Here $\hat{P} = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$ stands for the empirical distribution, which assigns weight $1/n$ to each observation. In the nonparametric model, with P unknown, the empirical estimator is efficient. This result is due to Levit [20]; see also Koshevnik and Levit [18]. Efficiency of the empirical distribution function $\hat{F}(t) = \hat{P}(-\infty, t]$, $t \in \mathbf{R}$, considered as infinite-dimensional functional, is proved by Beran [4].

Now suppose that the model consists of all distributions P which fulfill the linear constraint $Pa = \int a(x)P(dx) = 0$, where $a(x)$ is a known k-dimensional vector of P-square-integrable functions. Then the empirical estimator $\hat{P}a$ estimates the zero vector, and we can introduce new estimators $\hat{P}(f - c^\top a)$ for Pf, with c some known k-dimensional vector of real numbers. Such estimators have asymptotic variance $P(f - c^\top a)^2$. If $Paa^\top$ is invertible, then by the Cauchy–Schwarz inequality the asymptotic variance is minimized for

$$c_f = (Paa^\top)^{-1}Paf.$$

The constant c_f depends on the unknown distribution P and must be estimated. A simple estimator is the empirical estimator

$$\hat{c}_f = (\hat{P}aa^\top)^{-1}\hat{P}af = \left(\sum_{i=1}^n a(X_i)a(X_i)^\top\right)^{-1} \sum_{i=1}^n a(X_i)f(X_i).$$

The resulting estimator $\hat{P}(f - \hat{c}_f^\top a)$ has the same asymptotic variance as the "estimator" $\hat{P}(f - c_f^\top a)$ which still depends on the unknown P through c_f. Levit [21] shows that $\hat{P}(f - \hat{c}_f^\top a)$ is efficient. Similar results exist for nonlinear constraints: Koshevnik and Levit [18, Section 6], Hipp [15], and Bickel, Klaassen, Ritov and Wellner [5, Section 3.2, Example 3]. For infinite-dimensional constraints see Koshevnik [17] and Bickel et al. [5, Section 6.2].

In this paper we assume that the model consists of all distributions P which fulfill the linear constraint $Pa_\vartheta = \int a_\vartheta(x)P(dx) = 0$, where a_ϑ is a known k-dimensional vector of P-square-integrable functions which depend on an unknown p-dimensional parameter ϑ. The main aim of the paper is to construct efficient estimators for linear functionals Pf under such a linear constraint. We construct such estimators in two steps. First we keep ϑ fixed and obtain an efficient estimator $\hat{P}(f - \hat{c}_{f,\vartheta}^\top a_\vartheta)$ as above, with $\hat{c}_{f,\vartheta}$ defined as $\hat{c}_f$, now with a_ϑ in place of a. Then we replace ϑ by an efficient estimator $\hat{\vartheta}$ and obtain the estimator $\hat{P}(f - \hat{c}_{f,\hat{\vartheta}}^\top a_{\hat{\vartheta}})$ for Pf. We prove in Theorem 3 that this estimator is efficient.

The heuristic principle behind our construction is the following: If $\hat{t}_\vartheta$ is an efficient estimator for some (real-valued) functional $t(P)$ when ϑ is known, and $\hat{\vartheta}$ is efficient for ϑ, then the "plug-in" estimator $\hat{t}_{\hat{\vartheta}}$ is efficient for $t(P)$ when ϑ is unknown. For semiparametric models $P_{\vartheta,F}$, with F a possibly infinite-dimensional nuisance parameter, Klaassen and Putter [16] give conditions for the plug-in principle to hold; see also Müller, Schick and Wefelmeyer [23] for semiparametric stochastic process models. For the constrained model considered here, we cannot introduce a nuisance parameter, in general. We show in Theorem 1 that nevertheless the plug-in principle continues to hold under appropriate conditions. The conditions can be weakened at the expense of more complicated estimators.

Our construction requires an efficient estimator for ϑ. The constraint $Pa_\vartheta = 0$

suggests estimating equations of the form

$$B_\vartheta \sum_{i=1}^n a_\vartheta(X_i) = 0,$$

where B_ϑ is a $p \times k$-dimensional matrix of weights possibly depending on the unknown parameter ϑ. The weight matrix minimizing the asymptotic variance of the estimator from this equation depends, in general, on the unknown distribution P and must be estimated. We show in Theorem 2 that the resulting estimating equation gives an efficient estimator for ϑ.

Our estimators are related to the empirical likelihood principle. To describe the connection, let us return for a moment to the constraint $Pa = 0$ not involving a parameter, and to estimation of a linear functional Pf. It can be shown that the improved empirical estimator $\hat{P}(f - \hat{c}_f^\top a)$ is an asymptotic version of the *empirical likelihood estimator* $\hat{P}^{\mathrm{lik}} f$, where $\hat{P}^{\mathrm{lik}} = \sum_{i=1}^n p_i \delta_{X_i}$ is a multinomial distribution on the observations $X_1, \ldots, X_n$, with probabilities p_i chosen such that $\prod_{i=1}^n p_i$ is maximized subject to the constraint $\sum p_i a(X_i) = 0$. $\hat{P}^{\mathrm{lik}}$ may be interpreted as minimum discriminant information adjustment of $\hat{P}$ in the sense of Haberman [10]. See also Sheehy [31]. The empirical likelihood was introduced by Owen [27]. The stochastic equivalence of the two estimators follows from an appropriate stochastic expansion of the probabilities p_i. See Owen [28, relation (2.17)] for constraints of the form $E(X - \vartheta) = 0$, and Zhang [33, Lemma 2.1] for estimators of the distribution function $F(t) = P(-\infty, t]$. A generalization of empirical likelihood to estimation of *nonlinear* functionals, namely M-functionals, is in Zhang [33, 35].

Similarly, in the model with constraint $Pa_\vartheta = 0$ with unknown parameter ϑ, our efficient estimator for ϑ may be shown to be an asymptotic version of the corresponding empirical likelihood estimator $\hat{\vartheta}^{\mathrm{lik}}$ for ϑ, which maximizes $\prod_{i=1}^n p_i(\vartheta)$ in ϑ, where the $p_i(\vartheta)$ again maximize $\prod_{i=1}^n p_i$, now under the constraint $\sum_{i=1}^n p_i a_\vartheta(X_i) = 0$. We write $\hat{P}_\vartheta^{\mathrm{lik}} f = \sum_{i=1}^n p_i(\vartheta) f(X_i)$ for the empirical likelihood estimator of Pf when ϑ is known. Then our efficient plug-in estimator $\hat{P}(f - \hat{c}_{f,\hat{\vartheta}}^\top a_{\hat{\vartheta}})$ for Pf can be shown to be stochastically equivalent to the corresponding empirical likelihood plug-in estimator $\hat{P}_{\hat{\vartheta}^{\mathrm{lik}}}^{\mathrm{lik}} f$. Hence our Theorems 2 and 3 imply efficiency of $\hat{\vartheta}^{\mathrm{lik}}$ and $\hat{P}_{\hat{\vartheta}^{\mathrm{lik}}}^{\mathrm{lik}} f$. We do not give details of these arguments here. For a direct efficiency proof see Qin and Lawless [29, Theorem 3]; they restrict attention to estimating ϑ and $F(t)$. Zhang [34] shows weak convergence of the empirical likelihood estimator for the distribution function F considered as infinite-dimensional functional.

In the model constrained by $Pa_\vartheta = 0$, we must assume $p \le k$, i.e. at most as many unknown parameters as constraints, because otherwise the model would not be well-defined, in general. A degenerate case is $p = 0$. This means that the parameter ϑ is known, or to put it differently, the vector a does not depend on a parameter. Then we are back to the constraint $Pa = 0$ discussed first.

The case $p = k$ is also degenerate because then the unknown distribution is not constrained at all, and the condition $Pa_\vartheta = 0$ simply defines a functional $\vartheta(P)$ on a nonparametric model. See also Ahn and Schmidt [1]. This case is also of interest, for two reasons. One is that we may want to estimate the parameter ϑ. The other is that certain *nonlinear* functionals on nonparametric models may conveniently be expressed using *linear* constraints. Let us illustrate this point with a simple example. Suppose we want to estimate the variance $E(X - EX)^2$ when P is an unknown distribution on the real line. We introduce the one-dimensional parameter $\vartheta = EX$, i.e. the one-dimensional constraint $Pa_\vartheta = 0$ with $a_\vartheta(x) = x - \vartheta$. The advantage is that for known ϑ the nonlinear functional $E(X - EX)^2$ becomes the linear functional $E(X - \vartheta)^2$. An efficient estimator for the variance is then obtained by plugging the empirical estimator $\hat{\vartheta} = \frac{1}{n}\sum_{i=1}^n X_i$ for $\vartheta = EX$ into the empirical estimator $\frac{1}{n}\sum_{i=1}^n (X_i - \vartheta)^2$ for the variance when ϑ is known. The resulting estimator is of course just the sample variance.

Constraints can be written in different ways. A simple example is the model with known coefficient of variation c,

$$EX = c(E(X - EX)^2)^{1/2}. \tag{1.1}$$

This is a nonlinear constraint not involving a parameter. Set $EX = \vartheta$ and rewrite the constraint (1.1) as

$$c^2 E(X - \vartheta)^2 = \vartheta^2$$

to obtain linear constraints as considered in this paper, $Pa_\vartheta = 0$ with $a_{1\vartheta}(x) = x - \vartheta$ and $a_{2\vartheta}(x) = c^2(x - \vartheta)^2 - \vartheta^2$. A third way of writing this example is used in Qin and Lawless [30]: They set $E(X - \vartheta)^2 = \sigma^2$ with constraint $\vartheta = c\sigma$ on the two parameters.

Different ways of writing constraints may be convenient for different purposes. To calculate variance bounds, it is usually not helpful to introduce additional parameters into a given model. For the construction of estimators, this can however be useful. We have already seen this in the simple case of estimating the variance in a nonparametric model. Similarly, the model with known coefficient of variation is conveniently written in terms of two *linear* constraints by introducing the mean as parameter.

The paper is organized as follows. Section 2 recalls a characterization of efficient estimators, based on a version of Hájek's [11] convolution theorem for infinite-dimensional models. Section 3 gives conditions under which the plug-in principle works if efficient estimators for ϑ are available. Section 4 constructs an efficient estimator for ϑ as solution of an estimating equation $\hat{B}_\vartheta \sum_{i=1}^n a_\vartheta(X_i) = 0$, with $\hat{B}_\vartheta$ a random matrix of weights. Section 5 considers estimating expectations Pf. Section 6 describes some examples.

2 Characterization of efficient estimators

In this section we introduce, in the context of constrained models, appropriate definitions of differentiable functional, regular and asymptotically linear estimator, and characterizations of regular and of efficient estimators.

Let $X_1, \ldots, X_n$ be i.i.d. with distribution $P(dx)$. Let $a_\tau(x)$ be a k-dimensional vector of P-square-integrable functions, with p-dimensional parameter τ, and $p \leq k$. Write $Pa_\tau = Ea_\tau(X) = \int a_\tau(x)P(dx)$. The model consists of all distributions P such that $Pa_\tau = 0$ for some τ. Fix P and ϑ with $Pa_\vartheta = 0$. Write

$$L_{2,0}(P) = \{v \in L_2(P) : Pv = 0\}.$$

For $v \in L_{2,0}(P)$ let P_{nv} be Hellinger differentiable with derivative v,

$$P\Big(\Big(\frac{dP_{nv}}{dP}\Big)^{1/2} - 1 - \frac{1}{2}n^{-1/2}v\Big)^2 = o(n^{-1}).$$

Assumption 1. The vector $a_\tau(x)$ is $L_2(P)$-differentiable at $\tau = \vartheta$ with $k \times p$-matrix of partial derivatives $\dot{a}_\vartheta(x)$,

$$P[|a_\tau - a_\vartheta - \dot{a}_\vartheta(\tau - \vartheta)|^2] = o(|\tau - \vartheta|^2),$$

and Lipschitz at $\tau = \vartheta$,

$$|a_\tau(x) - a_\vartheta(x)| \leq z(x)|\tau - \vartheta|$$

for some $z \in L_2(P)$. Also, the $k \times k$-matrix $A_\vartheta = P[a_\vartheta a_\vartheta^\top]$ is nonsingular, and $P\dot{a}_\vartheta$ has full rank p.

Assumption 1 will be in force throughout. From now on we will omit the true parameter ϑ whenever convenient. The perturbed distribution P_{nv} must fulfill a constraint $P_{nv}a_{\vartheta_{nu}} = 0$ for a possibly perturbed parameter $\vartheta_{nu} = \vartheta + n^{-1/2}u$, with $u \in \mathbf{R}^p$,

$$0 = P_{nv}a_{\vartheta_{nu}} = P[(1 + n^{-1/2}v)(a + n^{-1/2}\dot{a}u)] + o(n^{-1/2}).$$

This leads to a constraint $P[av] = -P\dot{a}u$ on the perturbation v. The tangent space V_* of the model at P is the set of all $v \in L_{2,0}(P)$ fulfilling such a constraint. Write

$$V_u = \{v \in L_{2,0}(P) : P[av] = -P\dot{a}u\}.$$

Then V_* is the union of the affine spaces V_u, $u \in \mathbf{R}^p$. The tangent space V_* is linear and closed.

Note that our local models consists of distributions P_{nv} that are absolutely continuous with respect to P. This does not mean that our efficiency results refer only to models consisting of mutually absolutely continuous distributions. In general, we may have chosen our local model too small, in the sense that the corresponding asymptotic variance bounds for estimators are unattainable. In our setting, however, we will exhibit estimators *attaining* the bounds. This implies that the local model was chosen large enough: it contains the least favorable submodel. On the other hand, our efficiency results continue to hold when the underlying distributions fulfill additional (smoothness and moment)

conditions, as long as we find a local model also fulfilling these conditions, and with tangent space equal to V_* or at least dense in V_*.

We recall the following definitions and results from Le Cam's and Hájek's theory of efficient estimation. A reference is Bickel, Klaassen, Ritov and Wellner [5]. A q-dimensional functional $t(P)$ is called *differentiable* at P with *gradient* g if $g \in L_{2,0}(P)^q$ and

$$n^{1/2}(t(P_{nv}) - t(P)) \to P[gv] \quad \text{for } v \in V_*.$$

The *canonical* gradient g_* is the componentwise projection of g onto the tangent space V_*. An estimator $\hat{t}$ for $t(P)$ is called *regular* at P with *limit* L if

$$n^{1/2}(\hat{t} - t(P_{nv})) \Rightarrow L \quad \text{under } P_{nv}^n \quad \text{for } v \in V_*.$$

The convolution theorem says that if $\hat{t}$ is regular for $t(P)$ with limit L, then

$$L = (P[g_* g_*^\top])^{1/2} N_q + M \quad \text{in distribution,}$$

with N_q a q-dimensional standard normal random vector, and M independent of N_q. This justifies calling a regular estimator *efficient* for $t(P)$ if its limit is

$$L = (P[g_* g_*^\top])^{1/2} N_q \quad \text{in distribution.}$$

An estimator $\hat{t}$ for $t(P)$ is called *asymptotically linear* at P with *influence function* b if $b \in L_{2,0}(P)^q$ and

$$n^{1/2}(\hat{t} - t(P)) = n^{-1/2} \sum_{i=1}^{n} b(X_i) + o_{P^n}(1).$$

Result 1. *An asymptotically linear estimator for $t(P)$ is regular if and only if its influence function is a gradient for $t(P)$.*

Result 2. *A regular estimator for $t(P)$ is efficient if and only if it is asymptotically linear with influence function equal to the canonical gradient of $t(P)$.*

3 Plug-in estimators

In this section we describe how to construct an efficient estimator of a finite-dimensional functional $t(P)$. We begin by decomposing the tangent space V_* into tangent space for known ϑ and orthogonal complement.

In Section 2 we have seen that the tangent space V_* consists of the solutions v of *inhomogeneous* equations $P[av] = -P\dot{a}u$ for some $u \in \mathbf{R}^p$. If ϑ is *known*, the tangent space reduces to the space of solutions of the corresponding *homogeneous* equation,

$$V_0 = \{v \in L_{2,0}(P) : P[av] = 0\}.$$

Let $[a]$ denote the linear span of the components $a_1, \ldots, a_k$ of a. We have the orthogonal decomposition

$$L_{2,0}(P) = V_0 \oplus [a]. \tag{3.1}$$

Set $A = P[aa^\top]$ and

$$\ell \;=\; -P\dot{a}^\top A^{-1} a. \tag{3.2}$$

Then

$$P[a\ell^\top] = -P\dot{a}. \tag{3.3}$$

Let e_j denote the p-dimensional unit vector. The j-th component ℓ_j of ℓ is the unique solution of $P[a\ell_j] = -P\dot{a}e_j$ that is *orthogonal* to V_0. Write $[\ell]$ for the linear span of $\ell_1, \ldots, \ell_p$. Then V_* has the orthogonal decomposition

$$V_* = V_0 \oplus [\ell]. \tag{3.4}$$

We have

$$I = P[\ell\ell^\top] = P\dot{a}^\top A^{-1} P\dot{a}. \tag{3.5}$$

By Assumption 1, the $p \times p$-matrix I is nonsingular. It will play the role of *Fisher information* for ϑ.

Lemma 1. *Let $t(P)$ be a q-dimensional functional which is differentiable at P with gradient $g \in L_{2,0}(P)^q$. The canonical gradients of $t(P)$ for known and unknown ϑ, respectively, are*

$$g_0 = g - g_a, \quad g_* = g - g_a + g_\ell,$$

where g_a and g_ℓ are the projections of g onto $[a]$ and $[\ell]$, respectively. We have

$$\begin{aligned} g_a &= P[ga^\top]A^{-1}a, \\ g_\ell &= P[g\ell^\top]I^{-1}\ell \;=\; P[ga^\top]A^{-1}P\dot{a}(P\dot{a}^\top A^{-1}P\dot{a})^{-1}P\dot{a}^\top A^{-1}a. \end{aligned}$$

A degenerate case is $p = k$. Then $P\dot{a}$ has an inverse, and $g_\ell = g_a$, i.e. $g_* = g$.

Proof. The tangent space for known ϑ is V_0. By (3.1), the orthogonal complement of V_0 in $L_{2,0}(P)$ is $[a]$. The projection of g onto $[a]$ is of the form $g_a = W^\top a$ with W a $k \times q$-matrix determined by

$$P[(g - W^\top a)a^\top] = 0,$$

i.e. $W = A^{-1}P[ag^\top]$. The projection of g onto V_0 is $g_0 = g - g_a$.

By (3.4), the tangent space for unknown ϑ has the orthogonal decomposition $V_* = V_0 \oplus [\ell]$. The projection of g onto $[\ell]$ is of the form $g_\ell = Z^\top \ell$ with Z a $p \times q$-matrix determined by

$$P[(g - Z^\top \ell)\ell^\top] = 0,$$

i.e. $Z = I^{-1}P[\ell g^\top]$. Hence the projection of g onto V_* is $g_* = g_0 + g_\ell = g - g_a + g_\ell$. Use (3.2) to rewrite ℓ in terms of a. □

Consider the parameter τ as a p-dimensional functional of P by setting $\vartheta(P) = \tau$ if $Pa_\tau = 0$.

Lemma 2. *The functional $\vartheta(P)$ is differentiable at P with canonical gradient*

$$I^{-1}\ell = -(P\dot{a}^\top A^{-1} P\dot{a})^{-1} P\dot{a}^\top A^{-1} a.$$

Proof. By (3.2) we have $\ell = -P\dot{a}^\top A^{-1} a$. By definition of V_u, for $v \in V_u$ we have, using (3.5),

$$P[\ell v] = -P\dot{a}^\top A^{-1} P[av] = P\dot{a}^\top A^{-1} P\dot{a} u = Iu.$$

On the other hand, for $v \in V_u$,

$$n^{1/2}(\vartheta(P_{nv}) - \vartheta(P)) = n^{1/2}(\vartheta_{nu} - \vartheta) = u.$$

The last two equations imply that $I^{-1}\ell$ is a gradient of ϑ. Since $I^{-1}\ell$ is in $[a]$ and hence in V_*, the gradient is canonical. Use (3.2) to rewrite ℓ in terms of a. □

Remark 1. Our efficient estimators will involve consistent estimators for expectations Pk_ϑ. Suppose the function $k_\tau(x)$ fulfills a Lipschitz condition at $\tau = \vartheta$ of the form

$$|k_\tau(x) - k_\vartheta(x)| \le z(x)|\tau - \vartheta| \tag{3.6}$$

for a P-integrable function z. Let $\hat{\vartheta}$ be consistent for ϑ. Then a consistent estimator for Pk_ϑ is obtained as $\frac{1}{n}\sum_{i=1}^n k_{\hat{\vartheta}}(X_i)$. This follows immediately from the inequality

$$\Big|\frac{1}{n}\sum_{i=1}^n k_\tau(X_i) - \frac{1}{n}\sum_{i=1}^n k_\vartheta(X_i)\Big| \le \frac{1}{n}\sum_{i=1}^n z(X_i)|\tau - \vartheta|$$

and the law of large numbers. If we have

$$|k_{j\tau}(x) - k_{j\vartheta}(x)| \le z_j(x)|\tau - \vartheta|$$

with P-square-integrable z_j and $k_{j\vartheta}$, then $k_{1\tau}(x)k_{2\tau}(x)$ fulfills a Lipschitz condition (3.6), and $\frac{1}{n}\sum_{i=1}^n k_{1\hat{\vartheta}}(X_i)k_{2\hat{\vartheta}}(X_i)$ is consistent for $P[k_{1\vartheta}k_{2\vartheta}]$.

By Remark 1, the Lipschitz condition of Assumption 1 implies that

$$\hat{A}_{\hat{\vartheta}} = \frac{1}{n}\sum_{i=1}^n a_{\hat{\vartheta}}(X_i)a_{\hat{\vartheta}}(X_i)^\top$$

is consistent for A. The Lipschitz condition also implies that the empirical process $E_{n\tau} = n^{-1/2}\sum_{i=1}^n (a_\tau(X_i) - Pa_\tau)$ is *stochastically equicontinuous* at $\tau = \vartheta$: For each $\varepsilon, \eta > 0$ there is a $\delta > 0$ such that

$$\limsup_n P^n\Big\{ \sup_{|\tau-\vartheta|\le\delta} |E_{n\tau} - E_{n\vartheta}| > \eta \Big\} \le \varepsilon. \tag{3.7}$$

See e.g. Andrews and Pollard [2].

Theorem 1. *Let $t(P)$ be a q-dimensional functional which is differentiable at P with gradient $g \in L_{2,0}(P)^q$. Let $\hat{t}$ be an estimator for $t(P)$ which is asymptotically linear at P with influence function g, and let $\hat{G}$ be consistent for $P[a_\vartheta g^\top]$. If $\hat{\vartheta}$ is regular and efficient for ϑ, then*

$$\hat{t}_* = \hat{t} - \hat{G}^\top \hat{A}_{\hat{\vartheta}}^{-1} \frac{1}{n}\sum_{i=1}^n a_{\hat{\vartheta}}(X_i)$$

is regular and efficient for $t(P)$.

Proof. By Result 1 and Lemma 2, the estimator $\hat{\vartheta}$ is asymptotically linear with influence function equal to the canonical gradient,

$$n^{1/2}(\hat{\vartheta} - \vartheta) = I^{-1} n^{-1/2} \sum_{i=1}^n \ell(X_i) + o_{P^n}(1).$$

By Assumption 1,

$$Pa_\tau - Pa_\vartheta = P\dot{a}(\tau - \vartheta) + o(|\tau - \vartheta|).$$

With these relations and stochastic equicontinuity (3.7), applied for $\tau = \hat{\vartheta}$,

$$\begin{aligned} n^{-1/2}\sum_{i=1}^n a_{\hat{\vartheta}}(X_i) &= n^{-1/2}\sum_{i=1}^n a(X_i) + n^{1/2}(Pa_{\hat{\vartheta}} - Pa_\vartheta) + o_{P^n}(1) \\ &= n^{-1/2}\sum_{i=1}^n a(X_i) + P\dot{a}n^{1/2}(\hat{\vartheta} - \vartheta) + o_{P^n}(1) \\ &= n^{-1/2}\sum_{i=1}^n a(X_i) + P\dot{a}I^{-1}n^{-1/2}\sum_{i=1}^n \ell(X_i) + o_{P^n}(1). \end{aligned} \tag{3.8}$$

By Remark 1, $\hat{A}_{\hat{\vartheta}}$ is consistent for A. Hence $\hat{G}^\top \hat{A}_{\hat{\vartheta}}^{-1}$ is consistent for $P[g a_\vartheta^\top]A^{-1}$. Taken together, $\hat{t}_*$ is asymptotically linear with influence function

$$g - P[g a_\vartheta^\top]A^{-1}(a + P\dot{a}I^{-1}\ell) = g - g_a + g_\ell = g_*.$$

By Lemma 1, this is the canonical gradient of $t(P)$. Hence $\hat{t}_*$ is regular and efficient for $t(P)$ by Result 2. □

Remark 2. The asymptotic covariance matrix of an efficient estimator of $t(P)$ is $P[g_* g_*^\top]$. Since $[\ell]$ is a subspace of $[a]$, the vector g_ℓ is shorter than g_a. Since $g - g_a$ is orthogonal to g_a, we have

$$\begin{aligned} P[g_* g_*^\top] &= P[gg^\top] - P[g_a g_a^\top] + P[g_\ell g_\ell^\top] \\ &= P[gg^\top] - P[ga^\top] A^{-1} P[ag^\top] \\ &\qquad + P[ga^\top] A^{-1} P\dot{a} (P\dot{a}^\top A^{-1} P\dot{a})^{-1} P\dot{a}^\top A^{-1} P[ag^\top]. \end{aligned}$$

The asymptotic covariance matrix has a straightforward interpretation: The first term is the asymptotic covariance matrix of $\hat{t}$. The second term is the covariance matrix reduction due to the information in the constraint $Pa_\vartheta = 0$. The third term is the covariance matrix increase due to the fact that we do not know the parameter ϑ in the constraint. The third term is never larger than the second term. In the degenerate case $k = p$, we have $g_\ell = g_a$ and hence $P[g_* g_*^\top] = P[gg^\top]$.

4 Estimators for the parameter

In this section we construct an efficient estimator of ϑ as solution of an estimating equation. For a different, recursive, efficient estimator of ϑ see Nevel'son [26].

Theorem 2. *Any consistent solution of*

$$\frac{1}{n} \sum_{i=1}^n \dot{a}_\tau(X_i)^\top \hat{A}_\tau^{-1} n^{-1/2} \sum_{i=1}^n a_\tau(X_i) = o_{P^n}(1) \tag{4.1}$$

is regular and efficient for ϑ.

Proof. By Remark 1, $\frac{1}{n} \sum_{i=1}^n \dot{a}_{\hat{\vartheta}}(X_i)$ and $\hat{A}_{\hat{\vartheta}}$ are consistent for $P\dot{a}$ and A, respectively. Furthermore,

$$Pa_\tau - Pa_\vartheta = P\dot{a}(\tau - \vartheta) + o(|\tau - \vartheta|).$$

Hence any consistent solution of (4.1) fulfills

$$\begin{aligned} o_{P^n}(1) &= P\dot{a}^\top A^{-1} n^{-1/2} \sum_{i=1}^n a_{\hat{\vartheta}}(X_i) \\ &= P\dot{a}^\top A^{-1} \Big(n^{-1/2} \sum_{i=1}^n a_\vartheta(X_i) + P\dot{a} n^{1/2} (\hat{\vartheta} - \vartheta) \Big) + o_{P^n}(1). \end{aligned}$$

We obtain

$$n^{1/2}(\hat{\vartheta} - \vartheta) = -(P\dot{a}^\top A^{-1} P\dot{a})^{-1} P\dot{a}^\top A^{-1} n^{-1/2} \sum_{i=1}^n a(X_i) + o_{P^n}(1).$$

This proves that $\hat{\vartheta}$ is asymptotically linear with influence function equal to the canonical gradient $I^{-1}\ell$ of $\vartheta(P)$ obtained in Lemma 2. Result 2 now implies that $\hat{\vartheta}$ is regular and efficient for ϑ. □

A proof of Theorem 2 via approximation by a multinomial distribution is sketched in Chamberlain [7]. The estimator in Theorem 2 may be interpreted as generalized method of moments (GMM) estimator in the sense of Hansen [12, 13]. For an interpretation of such estimators as maximum likelihood estimators in certain exponential families see Back and Brown [3].

5 Estimators for expectations

In this section we construct an efficient estimator for the expectation Pf of a P-square-integrable function $f(x)$. Since P_{nv} has Hellinger derivative v, and $Pv = 0$, we have

$$n^{1/2}(P_{nv}f - Pf) \to P[vf] = P[v(f - Pf)].$$

Hence $f - Pf$ is a gradient of Pf in $L_{2,0}(P)$. The canonical gradient g_f is now obtained from Lemma 1,

$$g_f = f - Pf - g_{af} + g_{\ell f},$$

with

$$\begin{aligned} g_{af} &= P[fa^\top]A^{-1}a, \\ g_{\ell f} &= P[f\ell^\top]I^{-1}\ell \;=\; P[fa^\top]A^{-1}P\dot{a}(P\dot{a}^\top A^{-1}P\dot{a})^{-1}P\dot{a}^\top A^{-1}a. \end{aligned}$$

Theorem 3. *Let $\hat{\vartheta}$ be regular and efficient for ϑ. Set*

$$\hat{F}_\tau = \frac{1}{n}\sum_{i=1}^{n} a_\tau(X_i)f(X_i).$$

Then

$$\hat{t}_{\hat{\vartheta}} = \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \hat{F}_{\hat{\vartheta}}^\top \hat{A}_{\hat{\vartheta}}^{-1}\frac{1}{n}\sum_{i=1}^{n} a_{\hat{\vartheta}}(X_i) \tag{5.1}$$

is regular and efficient for Pf.

Proof. By Remark 1, $\hat{F}_{\hat{\vartheta}}$ is consistent for $P[af]$. Hence $\hat{F}_{\hat{\vartheta}}^\top \hat{A}_{\hat{\vartheta}}^{-1}$ is consistent for $P[fa^\top]A^{-1}$. With expansion (3.8) we now obtain that $\hat{t}_{\hat{\vartheta}}$ is asymptotically linear with influence function

$$f - Pf - P[fa^\top]A^{-1}(a + P\dot{a}I^{-1}\ell) = f - Pf - g_{af} + g_{\ell f} = g_f.$$

Hence $\hat{t}_{\hat{\vartheta}}$ is regular and efficient for Pf by Result 2. □

Estimators of similar form arise when one estimates expectations in semiparametric models; see Brown and Newey [6, Section 4].

6 Examples

Linear constraints $Pa_\vartheta = 0$ involving a known k-dimensional vector of functions $a_\vartheta(x)$ with unknown p-dimensional parameter ϑ arise in many different contexts. Three types of vectors $a_\vartheta(x)$ are of particular interest. Functions $a_\vartheta(x) = r(x) - \gamma(\vartheta)$ arise in the method of moments; functions $a_\vartheta(x) = \Gamma(\vartheta)s(x)$ with $k \times q$-matrix Γ and q-vector s describe relations between expectations of different components of a function $s(x)$; functions $a_\vartheta(x) = r(x - \vartheta)$ define location models with certain symmetries. In the following we briefly discuss these three types and some more specific examples.

1. Let $a_\vartheta(x) = r(x) - \gamma(\vartheta)$. Then the constraint $Pa_\vartheta = 0$ can be written $Pr = \gamma(\vartheta)$, and the matrix of partial derivatives $\dot{a}_\vartheta(x) = -\dot{\gamma}(\vartheta)$ does not depend on x. We have $A = P[aa^\top] = P[rr^\top] - \gamma\gamma^\top$.

The efficient estimating equation (4.1) for ϑ is

$$\dot{\gamma}(\vartheta)^\top \Big(\frac{1}{n} \sum_{i=1}^n r(X_i) r(X_i)^\top - \gamma(\vartheta)\gamma(\vartheta)^\top \Big)^{-1} n^{-1/2} \sum_{i=1}^n (r(X_i) - \gamma(\vartheta)) = o_{P^n}(1).$$

The efficient estimator (5.1) for Pf is

$$\begin{aligned} \hat{t}_{\hat{\vartheta}} \;\; = \;\; & \frac{1}{n} \sum_{i=1}^n f(X_i) - \frac{1}{n} \sum_{i=1}^n f(X_i)(r(X_i) - \gamma(\hat{\vartheta}))^\top \\ & \Big(\frac{1}{n} \sum_{i=1}^n r(X_i) r(X_i)^\top - \gamma(\hat{\vartheta})\gamma(\hat{\vartheta})^\top \Big)^{-1} \frac{1}{n} \sum_{i=1}^n (r(X_i) - \gamma(\hat{\vartheta})). \end{aligned}$$

For real-valued observations, such a constraint arises in the method of moments, $EX^j = \gamma_j(\vartheta)$, $j = 1, \ldots, k$, with $r(x) = (x, \ldots, x^k)^\top$. A particular case is $(E(X - EX)^2)^{1/2} = cEX$, with known coefficient of variation c. Introduce $\vartheta = EX$ to obtain $EX^2 = (c^2 + 1)\vartheta^2$, i.e. $r(x) = (x, x^2)^\top$ and $\gamma(\vartheta) = (\gamma_1(\vartheta), \gamma_2(\vartheta))^\top$ with $\gamma_1(\vartheta) = \vartheta$ and $\gamma_2(\vartheta) = (c^2 + 1)\vartheta^2$.

Let $X_i = (Y_i, Z_i)$ be bivariate i.i.d. observations with $EY = EZ$. Such models are used in survey sampling. For an approach via empirical likelihood see Kuk and Mak [19] and Chen and Qin [8]. We introduce a real-valued parameter ϑ and write the model in terms of a two-dimensional constraint $Pa_\vartheta = 0$ with $a_\vartheta(x) = (y - \vartheta, z - \vartheta)^\top$. Here A_ϑ is the covariance matrix

$$A_\vartheta = \mathrm{cov} X = \begin{pmatrix} \eta & \rho \\ \rho & \zeta \end{pmatrix},$$

say. Let $\hat{\eta}$, $\hat{\zeta}$ and $\hat{\rho}$ be the corresponding empirical estimators. Then the efficient estimating equation (4.1) for ϑ is

$$(\hat{\eta} - \hat{\rho}) \sum_{i=1}^n (Y_i - \vartheta) + (\hat{\zeta} - \hat{\rho}) \sum_{i=1}^n (Z_i - \vartheta) = 0.$$

Hence an efficient estimator for ϑ is

$$\hat{\vartheta} = \left(1 + \frac{\hat{\zeta} - \hat{\rho}}{\hat{\eta} - \hat{\rho}}\right)^{-1} \frac{1}{n} \sum_{i=1}^{n} Y_i + \left(1 + \frac{\hat{\eta} - \hat{\rho}}{\hat{\zeta} - \hat{\rho}}\right)^{-1} \frac{1}{n} \sum_{i=1}^{n} Z_i.$$

Not surprisingly, this is the optimal weighted average of the two sample means. Similarly as for known coefficient of variation, see the Introduction, it would also make sense to introduce a *two*-dimensional parameter μ via $EX = \mu$, with constraint on the parameter, here $\mu_1 = \mu_2$.

2. Let $a_\vartheta(x) = \Gamma(\vartheta)s(x)$ with $k \times q$-matrix Γ and q-vector s. Then the constraint $Pa_\vartheta = \Gamma(\vartheta)Ps = 0$ describes linear relations between expectations of different components of s, with coefficients depending on the parameter ϑ. We have $A = P[aa^\top] = \Gamma P[ss^\top]\Gamma^\top$ and

$$\dot{a}_\vartheta(x) = \dot{\Gamma}(\vartheta)s(x) = (\Gamma^{(1)}(\vartheta)s(x), \ldots, \Gamma^{(p)}(\vartheta)s(x)),$$

where $\Gamma^{(j)}(\vartheta) = \partial_{\vartheta_j}\Gamma(\vartheta)$ is the matrix of partial derivatives of $\Gamma(\vartheta)$ with respect to the j-th component of ϑ.

The efficient estimating equation (4.1) for ϑ is

$$\sum_{i=1}^{n} s(X_i)^\top \dot{\Gamma}(\vartheta)^\top \Big(\Gamma(\vartheta) \sum_{i=1}^{n} s(X_i)s(X_i)^\top \Gamma(\vartheta)^\top\Big)^{-1} \Gamma(\vartheta) n^{-1/2} \sum_{i=1}^{n} s(X_i) = o_{P^n}(1).$$

The efficient estimator (5.1) for Pf is

$$\begin{aligned} \hat{t}_{\hat{\vartheta}} &= \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \sum_{i=1}^{n} f(X_i)s(X_i)^\top \Gamma(\hat{\vartheta})^\top \\ &\qquad \Big(\Gamma(\hat{\vartheta}) \sum_{i=1}^{n} s(X_i)s(X_i)^\top \Gamma(\hat{\vartheta})^\top\Big)^{-1} \Gamma(\hat{\vartheta}) \frac{1}{n} \sum_{i=1}^{n} s(X_i). \end{aligned}$$

Suppose in particular that we have real-valued observations, and dependencies between *centered* moments, say $E(X - EX)^j = \gamma_j(\vartheta_2, \ldots, \vartheta_p)$ for $j = 2, \ldots, k$. If we introduce a new parameter $\vartheta_1 = EX$ and express the centered moments in terms of the moments, we can write the dependencies as a constraint of the form $Pa_\vartheta = \Gamma(\vartheta)Ps = 0$ with $s(x) = (x, \ldots, x^k)^\top$ and Γ a k-dimensional row vector. Dependencies between cumulants rather than centered moments can be written in a similar way.

An important example of a model described by relations between centered moments is the quasi-likelihood model, which assumes that the variance is a function of the mean, $\mathrm{var}X = v(EX)$ with known function $v(x)$. It is conveniently written introducing a real parameter ϑ for the mean,

$$EX = \vartheta, \quad E(X - \vartheta)^2 = v(\vartheta).$$

This is a linear constraint $Pa_\vartheta = 0$ with $a_\vartheta(x) = (x-\vartheta, (x-\vartheta)^2 - v(\vartheta))^\top$. The sample mean would be optimal in the sense of the theory of estimating functions, see Godambe and Heyde [9]. A better, efficient, estimator for ϑ is obtained from estimating equation (4.1). Generalized linear models are extensions to non-identically distributed observations; see McCullagh and Nelder [22, Chapter 10]. Quasi-likelihood models also have versions for regression models and autoregressive models. We refer to Heyde [14] for an approach via estimating functions, and to Wefelmeyer [32] for efficient estimators. General constraints in regression models will be treated in Müller and Wefelmeyer [24].

3. Let the observations be p-dimensional and $a_\vartheta(x) = r(x-\vartheta)$, with r of dimension $k \geq p$. Then $\dot{a}_\vartheta(x) = -r'(x-\vartheta)$ and $A_\vartheta = P[a_\vartheta a_\vartheta^\top] = \int r(x-\vartheta) r(x-\vartheta)^\top P(dx)$.

The efficient estimating equation (4.1) for ϑ is

$$\sum_{i=1}^n r'(X_i-\vartheta)^\top \Big(\sum_{i=1}^n r(X_i-\vartheta) r(X_i-\vartheta)^\top \Big)^{-1} n^{-1/2} \sum_{i=1}^n r(X_i-\vartheta) = o_{P^n}(1).$$

The efficient estimator (5.1) for Pf is

$$\begin{aligned} \hat{t}_{\hat{\vartheta}} &= \frac{1}{n} \sum_{i=1}^n f(X_i) - \sum_{i=1}^n f(X_i) r(X_i-\hat{\vartheta})^\top \\ &\qquad \Big(\sum_{i=1}^n r(X_i-\hat{\vartheta}) r(X_i-\hat{\vartheta})^\top \Big)^{-1} \frac{1}{n} \sum_{i=1}^n r(X_i-\hat{\vartheta}). \end{aligned}$$

For $k = p$ this gives a nonparametric model, and the constraint $\int r(x-\vartheta) P(dx) = 0$ defines a location parameter. For $k > p$ the constraint describes certain symmetries of P around ϑ. Nevel'son [25] writes the model as a semiparametric model $P_{\vartheta F}(dx) = dF(x-\vartheta)$ with distribution function F fulfilling the constraint $\int r dF = 0$, which does not involve the parameter. He determines a variance bound for estimators of ϑ.

Acknowledgment. We thank the referees for suggestions that improved the exposition, and Erich Häusler for pointing out that our estimators are related to empirical likelihood estimators.

References

[1] S.C. Ahn and P. Schmidt, *A separability result for GMM estimation, with applications to GLS prediction and conditional moment tests*, Econometric Rev., 14 (1995), pp. 19–34.

[2] D.W.K. Andrews and D. Pollard, *An introduction to functional central limit theorems for dependent stochastic processes*, Internat. Statist. Rev., 62 (1994), pp. 119–132.

[3] K. Back and D.P. Brown, *GMM, maximum likelihood, and nonparametric efficiency*, Econom. Lett., 39 (1992), pp. 23–28.

[4] R. Beran, *Estimating a distribution function*, Ann. Statist., 5 (1977), pp. 400–404.

[5] P.J. Bickel, C.A.J. Klaassen, Y. Ritov and J.A. Wellner, *Efficient and Adaptive Estimation for Semiparametric Models*, Springer, New York, 1998.

[6] B.W. Brown and W.K. Newey, *Efficient semiparametric estimation of expectations*, Econometrica, 66 (1998), pp. 453–464.

[7] G. Chamberlain, *Asymptotic efficiency in estimation with conditional moment restrictions*, J. Econometrics, 34 (1987), pp. 305–334.

[8] J. Chen and J. Qin, *Empirical likelihood estimation for finite populations and the effective usage of auxiliary information*, Biometrika, 80 (1993), pp. 107–116.

[9] V.P. Godambe and C.C. Heyde, *Quasi-likelihood and optimal estimation*, Internat. Statist. Rev., 55 (1987), pp. 231–244.

[10] S.J. Haberman, *Adjustment by minimum discriminant information*, Ann. Statist., 12 (1984), pp. 971–988.

[11] J. Hájek, *A characterization of limiting distributions of regular estimates*, Z. Wahrsch. Verw. Gebiete, 14 (1970), pp. 323–330.

[12] L.P. Hansen, *Large sample properties of generalized method of moments estimators*, Econometrica, 50 (1982), pp. 1029–1054.

[13] L.P. Hansen, *A method for calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators*, J. Econometrics, 30 (1985), pp. 203–238.

[14] C.C. Heyde, *Quasi-Likelihood And Its Application. A General Approach to Optimal Parameter Estimation*, Springer Series in Statistics, Springer, New York, 1997.

[15] C. Hipp, *Efficient estimation under constraints*, Kybernetika, 2 (1991), pp. 100–113.

[16] C.A.J. Klaassen and H. Putter, *Efficient estimation of Banach parameters in semiparametric models*, Technical Report, Department of Mathematics, University of Amsterdam, 1999.

[17] Y.A. Koshevnik, *Efficient estimation for restricted nonparametric models*, Technical Report, Department of Statistical Science, Southern Methodist University, 1992.

[18] Y.A. Koshevnik and B.Y. Levit, *On a non-parametric analogue of the information matrix*, Theory Probab. Appl., 21 (1976), pp. 738–753.

[19] A.Y.C. Kuk and T.K. Mak, *Median estimation in the presence of auxiliary information*, J. Roy. Statist. Soc. Ser. B, 51 (1989), pp. 261–269.

[20] B.Y. Levit, *On optimality of some statistical estimates*, In: Proceedings of the Prague Symposium on Asymptotic Statistics, (J. Hájek, ed.), Vol. 2, pp. 215–238, Charles University, Prague, 1974.

[21] B.Y. Levit, *Conditional estimation of linear functionals*, Problems Inform. Transmission, 11 (1975), pp. 39–54.

[22] P. McCullagh and J.A. Nelder, *Generalized Linear Models*, 2nd ed., Monographs on Statistics and Applied Probability 37, Chapman and Hall, London, 1989.

[23] U.U. Müller, A. Schick and W. Wefelmeyer, *Plug-in estimators in semiparametric stochastic process models*, To appear in: Selected Proceedings of the Symposium on Inference for Stochastic Processes (I.V. Basawa, C.C. Heyde and R.L. Taylor, eds.), IMS Lecture Notes-Monograph Series, Institute of Mathematical Statistics, Hayward, California, 2001.

[24] U.U. Müller and W. Wefelmeyer, *Regression type models and optimal estimators*, In preparation.

[25] M.B. Nevel'son, *One informational lower bound*, Problems Inform. Transmission, 13 (1977), pp. 181–185.

[26] M.B. Nevel'son, *Asymptotic optimality of recursive estimates*, Problems Inform. Transmission, 14 (1978), pp. 35–49.

[27] A.B. Owen, *Empirical likelihood ratio confidence intervals for a single functional*, Biometrika, 75 (1988), pp. 237–249.

[28] A.B. Owen, *Empirical likelihood ratio confidence regions*, Ann. Statist., 18 (1990), pp. 90–120.

[29] J. Qin and J. Lawless, *Empirical likelihood and general estimating equations*, Ann. Statist., 22 (1994), pp. 300–325.

[30] J. Qin and J. Lawless, *Estimating equations, empirical likelihood and constraints on parameters*, Canad. J. Statist., 23 (1995), pp. 145–159.

[31] A. Sheehy, *Kullback–Leibler constrained estimation of probability measures* Technical Report 137, Department of Statistics, University of Washington, Seattle, 1988.

[32] W. Wefelmeyer, *Quasi-likelihood models and optimal inference*, Ann. Statist., 24 (1996), pp. 405–422.

[33] B. Zhang, *M-estimation and quantile estimation in the presence of auxiliary information*, J. Statist. Plann. Inference, 44 (1995), pp. 77–94.

[34] B. Zhang, *Estimating a distribution function in the presence of auxiliary information*, Metrika, 46 (1997a), pp. 221–244.

[35] B. Zhang, *Empirical likelihood confidence intervals for M-functionals in the presence of auxiliary information*, Statist. Probab. Lett., 32 (1997b), pp. 87–97.

Ursula U. Müller
Fachbereich 3: Mathematik und Informatik
Universität Bremen
Postfach 330 440
28334 Bremen
Germany
uschi@math.uni-bremen.de
http://www.math.uni-bremen.de/∼uschi/

Wolfgang Wefelmeyer
Fachbereich 6 Mathematik
Universität Siegen
Walter-Flex-Str. 3
57068 Siegen
Germany
wefelmeyer@mathematik.uni-siegen.de
http://www.math.uni-siegen.de/statistik/wefelmeyer.html