Stochastics and Stochastics Reports, Vol. 43, pp. 161–178
Reprints available directly from the publisher
Photocopying permitted by license only

ON OPTIMALITY OF REGULAR PROJECTIVE ESTIMATORS IN SEMIMARTINGALE MODELS

KACHA DZHAPARIDZE

Centre for Mathematics and Computer Science,
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

and

PETER SPREIJ

Department of Econometrics, Free University,
De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands

(Received 4 June 1992; in final form 10 August 1992)

In the partially specified statistical models the class of regular estimators having a martingale representation is defined, and the best among them is sought in the sense of asymptotic second order characteristics.

KEY WORDS Filtered measurable space, special semimartingale, square integrable martingale, compensator, quadratic variation, regular estimator, projective estimator, asymptotically equivalent estimators, asymptotic differentiability, admissibility, scoring function, spread of estimator.

1 INTRODUCTION

1.1 In order to treat problems of drawing statistical inference in the setting of the general theory of stochastic processes (as presented e.g. in Jacod and Shiryaev (1987) or Liptser and Shiryayev (1989)), the experiment in question is supposed to be a filtered probability space with a family of probability measures, and an observed object is supposed to be a semimartingale with respect to all these measures. A solution then, sought in terms of the predictable characteristics of the observed semimartingale, is applicable to various statistical models in discrete or continuous time such as, for instance, the classical independent observations scheme, or those risen in regression, time series and survival analysis, where the models are only partially specified in terms, e.g., of the first or second order characteristics (regression analysis or spectral analysis in time series), or the intensity of a counting process (in survival analysis). We consider here the asymptotic setting of the problem with the observation time (sample size) increasing unboundedly, though adequate considerations can be carried out for sequences of experiments.

1.2 In the present paper we restrict our attention to the common situation in which the model under consideration admits a finite dimensional parametrization, reducing the model identification problem to the statistical estimation of a parameter.

Specifically, the following problem of estimation will be treated: deriving in the present general setting the Cramer–Rao type lower bound for a class of so-called *regular* estimators, and indicating particular estimators which attain this bound and therefore are *optimal*. Of course, those are nothing but the maximum likelihood estimators (rather a class of estimators asymptotically equivalent to MLE) whenever the model is fully specified as, for instance, in the classical case of independent observations from the fully parametrized density. In regression analysis, however, the best linear unbiased estimators are sought, in time series analysis the so-called Gaussian estimators, and in survival analysis the partial likelihood estimators. As applied to these special models, our unified approach leads naturally to the same optimal estimators. This is indeed not hard to see by specifying in each special case the general notions and results of the present paper, but we purposely avoid here these specifications in order to concentrate full attention to the unified setting.

1.3 As the usual scheme for deriving the Cramer–Rao inequality assumes a full parametrization of an experiment (see e.g. Ibragimov and Has'minskii (1981), I.7 and II.11) and therefore becomes unapplicable here, we need in the first place a proper formulation of the problem, which is then simply solved by applying the Schwartz inequality. To this end a number of adequate definitions is introduced restricting the class of considered estimators, which otherwise are viewed as arbitrary processes of the same dimension as the parameter itself, calculable from observations.

Firstly, using the observations of the semimartingale, we form all kinds of local martingales as the stochastic integrals with respect to this semimartingale (in statistical context the corresponding predictable integrands are usually called the *scoring functions*), and then use them for estimation; cf. Jacod (1990) and the references therein, in particular Godambe and Heyde (1990), Greenwood and Wefelmeyer (1989), Gushchin (1990), Sörensen (1990). Due to the *representation property* (see e.g. Jacod and Shiryaev (1987), III.4) all local martingales are representable as such integrals plus, perhaps, some orthogonal term which will be assumed negligible in the sense indicated below. Besides, a local martingale used is assumed square integrable, which means according to Liptser and Shiryayev (1989), Lemma III.5.1, assertion 3, that a possible extra term is also assumed to be negligible. Specifically, for each fixed value of the parameter all estimators considered admit a martingale representation in the sense that they can be represented, after an appropriate centering and scaling, as a certain square integrable margingale plus a remainder term (absorbing eventually negligible terms mentioned above), which can be ignored when determining the principal part of estimation precision (see 1.4 below). Accordingly, we say that two estimators are *asymptotically equivalent* if they have one and the same martingale representation (with different remainders, of course). Hence, a particular scoring function defines a class of asymptotically equivalent estimators.

Secondly, the fact that the model is not fully defined entails here that we can use only certain integrals with respect to the observed semimartingale (cf. regression and time series where only the linear and, respectively, quadratic forms from observations are admissible). The margingales so obtained, as well as corresponding scoring functions, are called *admissible*. Correspondingly, an estimator is called *admissible* if it has the martingale representation with an admissible martingale.

1.4 The principal part of estimation precision then is naturally determined by the scaling factor and the sharp brackets of the involved martingale in the form of a dispersion ellipsoid, called below the *spread* of an estimator. By applying the Schwartz inequality we get the lower bound for the spread of all admissible estimators. As a side remark we notice that even superefficient estimators obey this lower bound. In order to give the lower bound the usual Cramer–Rao form, we have to exclude this kind of abnormalities restricting the class of estimators by a certain regularity assumption.

It seems natural to call an admissible estimator with martingale representation for each fixed value of the parameter *regular* if the representation extends to a shrinking neighbourhood of the fixed parameter value, with the appropriate shrinkage rate related as usual to the growth of information.

1.5 In view of the fact that the spread of an estimator is defined as an asymptotic notion—the principal part of the estimation precision—it makes a sense to assume the *asymptotic differentiability* (weakly in the class of all admissible scoring functions; see Section 4) of the predicable characteristics of an observed semimartingale. (One can easily trace the simplifications caused by the differentiability assumption for a fixed sample size like in Jacod (1990); see Ibragimov and Has'minskii (1981), I.7 for the classical result and Barndorff–Nielsen and Sörensen (1990) for examples).

The main statement of the present paper can be described now as follows: *under the differentiability condition just mentioned, the spread of a regular estimator obeys the Cramer–Rao lower bound.*

1.6 As was mentioned above, for fully defined models this lower bound is attained by a special scoring function, namely that of involved in the likelihood equation. Surely, if the solution (approximate, may be) to this equation has the martingale representation, then it is an optimal estimator. The question on existence of this representation lies beyond the scope of the present paper (see e.g. Ibragimov and Has'minskii (1981), I.8 and III.1 or Chitashvili *et al.*, 1990). For not necessarily fully defined models, however, the optimal scoring function can be viewed as the projection of the above scoring function to the space of admissible scoring functions. Note that generally the projection operation requires the knowledge of some extra parameters which are supposed known or at least estimable by the given sample, as for instance in linear regression with independent residuals where the best linear unbiased estimator involves the variances of residuales (they cancel only in the i.i.d. case).

2 PRELIMINARIES

2.1 Let $(\Omega. \mathscr{F}, \mathbf{F}, P)$ be a stochastic basis with a filtration $\mathbf{F} = (\mathscr{F}_t)_{t \geq 0}$. Assume for simplicity that $\mathscr{F}_0$ is trivial P a.s. Let X be an adapted $\mathbb{R}^d$-valued locally square integrable semimartingale having on a set $\Omega^P \subset \Omega$ with $P(\Omega^P) = 1$ the Doob–Meyer decomposition $X = X_0 + M + A$ with the compensator $A \in \mathscr{A}_{\text{loc}}$ and the martingale part $M = X^c + x*(\mu - \nu) \in \mathscr{M}^2_{\text{loc}}$. As usual X^c and μ are the continuous part and the jump measure of X with the quadratic variation C and the compensator ν respectively,

chosen to satisfy the following relations: for each $\Gamma \in \mathcal{B}(\mathbb{R}_+)$

$$\nu(\omega, \Gamma \times \{0\}) = 0, \quad a_t(\omega) \equiv \nu(\omega, \{t\} \times \mathbb{R}^d) \leq 1 \quad \text{indentically and} \quad C = c \cdot v \quad (2.1.1)$$

with a continuous increasing process v and a nonnegative definite $\mathbb{R}^d \times \mathbb{R}^d$-matrix valued predictable process c (see Jacod and Shiryaev (1987), Section II.2 for more details). Then the quadratic variation of M is $\langle M \rangle = C + xx^T * \nu - [A]$.

2.2 With the continuous part $X^c \in \mathcal{M}^2_{\text{loc}}$, we may associate the linear space $L^2(X^c)$ of all $\mathbb{R}^1 \times \mathbb{R}^d$-valued predictable processes H such that $H c H^T \cdot v \in \mathcal{A}^+_{\text{loc}}$; see Jacod and Shiryaev (1987), Section III.4a. For $H \in L^2(X^c)$ we define the stochastic integral $H \cdot X^c$ as in Jacod and Shiryaev (1987), Theorem III.4.5. For $\mathbb{R}^k \times \mathbb{R}^d$-matrix valued predictable processes H and K with rows in $L^2(X^c)$ we have

$$\langle H \cdot X^c, K \cdot X^c \rangle = H c K^T \cdot v. \quad (2.2.1)$$

2.3 Denote $\tilde{\Omega} = \Omega \times \mathbb{R}_+ \times \mathbb{R}^d$ and $\tilde{\mathcal{P}} = \mathcal{P} \otimes \mathcal{B}(\mathbb{R}^d)$ where $\mathcal{P}$ is the predictable σ-field on $\Omega \times \mathbb{R}_+$. Let W be a $\tilde{\mathcal{P}}$-measurable function on $\tilde{\Omega}$ such that for each Markov time T

$$I(T < \infty) \int |W(\omega, T, x)| \nu(\omega; \{T\} \times dx) < \infty \quad \text{P-a.s.}$$

Associate with it the predictable process

$$\hat{W}_t(\omega) = \int W(\omega, t, x) \nu(\omega; \{t\} \times dx),$$

and note that $a = \hat{1}$ by (2.1.1), If $G^2(W) \in \mathcal{A}^+_{\text{loc}}$ with

$$G^2(W)_t = |W - \hat{W}|^2 * \nu_t + \sum_{s \leq t} (1 - a_s) |\hat{W}_s|^2, \quad (2.3.1)$$

then we say $W \in \mathcal{G}^2_{\text{loc}}(\mu)$. If W is $\mathbb{R}^k$-vector valued with components in $\mathcal{G}^2_{\text{loc}}(\mu)$, then

$$W * (\mu - \nu) \in \mathcal{M}^2_{\text{loc}}$$

and for a couple W and U

$$\langle W * (\mu - \nu), U * (\mu - \nu) \rangle_t = WU^T * \nu_t - \sum_{s \leq t} \hat{W}_s \hat{U}^T_s.$$

(cf. (2.3.1)) and

$$WU^T * \nu_t = \langle \tilde{W} * (\mu - \nu), U * (\mu - \nu) \rangle_t \quad \text{with} \quad \tilde{W} = W + 1_{\{a < 1\}} \hat{W}/(1 - a). \quad (2.3.2)$$

2.4 For brevity, use the following notations for $H \in L^2(X^c)$ and $W \in \mathscr{G}^2_{\text{loc}}(\mu)$:

$$M(H, W) = H \cdot X^c + W * (\mu - \nu) \quad \text{and} \quad \tilde{M}(H, W) = H \cdot X^c + \tilde{W} * (\mu - \nu). \tag{2.4.1}$$

By (2.2.1) and (2.3.2)

$$\langle \tilde{M}(H, W), M(K, U) \rangle_t = H c K^T \cdot v_t + W U^T * v_t, \tag{2.4.2}$$

while

$$\langle \tilde{M}(H, W) - M(H, W), M(K, U) \rangle_t = \sum_{s \le t} \hat{W}_s \hat{U}_s^T \tag{2.4.3}$$

and

$$\langle \tilde{M}(H, W), \tilde{M}(K, U) - M(K, U) \rangle_t = \sum_{s \le t} 1_{\{a_s < 1\}} \hat{W}_s \hat{U}_s^T / (1 - a_s). \tag{2.4.4}$$

2.5 Along with any $\mathbb{R}^d$-valued locally square integrable martingale $M \in \mathscr{M}^2_{\text{loc}}$, consider another locally square integrable martingale $m \in \mathscr{M}^2_{\text{loc}}$ of dimension d', say. Suppose that the quadratic variation $\langle M \rangle$ is positive definite at t- for t large enough, and define the $\mathbb{R}^{d'} \times \mathbb{R}^{d'}$-matrix valued predictable process

$$c(m, M) = \langle m \rangle - \langle m, M \rangle \langle M \rangle^{-1} \langle M, m \rangle. \tag{2.5.1}$$

In Section 5 we will need the following result concerning $c(m, M)$:

LEMMA 2.5.1 (Dzhaparidze and Spreij (1992)) *The process $c(m, M)$ defined by* (2.5.1) *is non decreasing, and $c(m, M) = 0$ iff there exists a $\mathscr{F}$-measurable random $(d' \times d)$-matrix C such that $m = CM$.*

Remark 2.5.2 C need not be $\mathscr{F}_0$-measurable. In Dzhaparidze and Spreij (1992) this result has been proved for the case where $\langle M \rangle^{-1}$ does not necessarily exist, and is replaced by $\langle M \rangle^+$, the Moore–Penrose inverse process. Notice too that even if C is not $\mathscr{F}_0$-measurable, it is such that the product CM is a martingale.

The process $c(m, M)$ is not symmetric. Instead we often use the so-called correlation process

$$\rho(m, M) = \langle m \rangle^{-1/2} \langle m, M \rangle \langle M \rangle^{-1/2} \tag{2.5.2}$$

which is simply related to $c(m, M)$ as follows:

$$\langle m \rangle^{1/2} c(m, M) \langle m \rangle^{1/2} = I - \rho(m, M) \rho(M, m) \ge 0. \tag{2.5.3}$$

The last inequality follows from the assertion of Lemma 2.5.1. In fact this is just the matrix version of the Schwartz inequality.

3 PARAMETRIZATION

3.1 Consider a set of probability measures $\mathbb{P}$, and suppose that under all $P \in \mathbb{P}$ a process X, adapted to a filtered measurable space $(\Omega, \mathcal{F}, \mathbf{F}, P)$, is an $\mathbb{R}^d$-valued locally square integrable semimartingale.

It will be supposed that a set of probability measures $\mathbb{P}$ allows the parametrization to be described in the present section.

3.2 For a fixed $P \in \mathbb{P}$ we single out in the linear spaces of integrands $L^2(X^c; P)$ and $\mathcal{G}^2_{\mathrm{loc}}(\mu; P)$, introduced in 2.2 and 2.3 respectively, the subspaces $\mathcal{H} \subset L^2(X^c; P)$ and $\mathcal{W} \subset \mathcal{G}^2_{\mathrm{loc}}(\mu; P)$ for all $P \in \mathbb{P}$, related by the condition that also $Hx \in \mathcal{W}$ for each $H \in \mathcal{H}$ to have that $\tilde{W}^P \in \mathcal{W}$; see (2.3.2).

Since for all $P \in \mathbb{P}$ the integrals $H \cdot A^P$ and $U * \nu^P$ with $H \in \mathcal{H}$ and $U = W - Hx \in \mathcal{W}$ are well defined, fixing P, $P' \in \mathbb{P}$ we may introduce the process

$$g^{P,P'}(H, W) = H \cdot (A^{P'} - A^P) + U * (\nu^{P'} - \nu^P) \quad \text{where} \quad U = W - Hx. \tag{3.2.1}$$

Note that on a set $\Omega^P \cap \Omega^{P'}$, which has by assumption in Section 2.1 full measure under P' if P' is locally dominated by P, we have

$$X^{cP'} - X^{cP} = x * (\nu^{P'} - \nu^P) - (A^{P'} - A^P) \tag{3.2.2}$$

and hence

$$M^{P'}(H, W) = M^P(H, W) - g^{P,P'}(H, W) \tag{3.2.3}$$

by definition in 2.4. In this case $g^{P,P'}(H, W)$ is the Girsanov correction term. Indeed, the density process of $P' \in \mathbb{P}$ relative to P, positive P'-a.s. for all $t \in \mathbb{R}_+$, is then the Dolean's exponential of the P-martingale $\tilde{M}^P = \tilde{M}^P(\beta, Y - 1)$ where $\beta \in \mathrm{L}^2(X^c; P)$ satisfies $\langle X^{cP'}, \tilde{M}^P \rangle = c^P \beta^T \cdot v$ and $Y - 1$ defined by $\nu^{P'} = Y \cdot \nu^P$, is such that

$$Y - 1 + 1_{\{a^P < 1\}}(a^{P'} - a^P)(1 - a^P)^{-1} \in \mathcal{G}_{\mathrm{loc}}(\mu; P) \tag{3.2.4}$$

where $a^{P'}$ and a^P are defined by (2.1.1) relative to P and P' respectively; see Jacod and Shiryaev (1987), III.5. Under these circumstances one can apply Girsanov's theorem as in Jacod and Shiryaev (1987), Lemma IV.3.19, to get

$$X^{cP'} - X^{cP} = x * (\nu^{P'} - \nu^P) - (A^{P'} - A^P) = -c^P \beta^T \cdot v \tag{3.2.5}$$

Hence (3.2.3) holds with Girsanov's correction term

$$\begin{aligned} g^{P,P'}(H, W) &= Hc^P \beta^T \cdot v + W * (\nu^{P'} - \nu^P) \\ &= \langle M^P(H, W), \tilde{M}^P(\beta, Y - 1) \rangle. \end{aligned}$$

The last equality is verified by (2.4.2), (3.2.1), (3.2.3) and (3.2.4). It should be noted in addition that in the most general case where the local domination property does not

necessarily hold, Eq. (3.2.5) takes a more complicated form involving certain correction terms; see Jacod (1990) or Jacod and Shiryaev (1987), IV.3.

3.3 Turning back to the restrictions imposed on the sets $\mathscr{H}$ and $\mathscr{W}$, we suppose that for all $P \in \mathbb{P}$ and

$$H \in \mathscr{H} \subset L^2(X^c; P), \quad W \in \mathscr{W} \subset \mathscr{G}^2_{\text{loc}}(\mu; P)$$

and all t large enough $\langle M^P \rangle_t > 0$ P-a.s. where $M^P = M^P(H, W)$, and that

$$\mathscr{L}\{\langle M^P \rangle_t^{-1/2} M_t^P : t \text{ large enough} | P\}$$

is relatively compact with non degenerate limit points.

For all $P \in \mathbb{P}$ define the subset $[P]$ of $\mathbb{P}$ by

$$[P] = \{P' \in \mathbb{P} : H \cdot A^{P'} = H \cdot A^P \text{ for each } H \in \mathscr{H} \text{ and } W * \nu^{P'} = W * \nu^P \text{ for each } W \in \mathscr{W}\}. \tag{3.3.1}$$

Hence we have $g^{P,P'}(H, W) = 0$ for each $P' \in [P]$, $H \in \mathscr{H}$ and $W \in \mathscr{W}$ (see (3.2.1)). Therefore on a set $\Omega^P \cap \Omega^{P'}$ we have by (3.2.2) that $M^P(H, W) = M^{P'}(H, W)$ for each $P' \in [P]$, $H \in \mathscr{H}$ and $W \in \mathscr{W}$.

Suppose now that $[\mathbb{P}] = \{[P] : P \in \mathbb{P}\}$ allows a finite dimensional parametrization: there exists a one to one mapping

$$\vartheta : [\mathbb{P}] \to \Theta \subset \mathbb{R}^k. \tag{3.3.2}$$

Thus, by definition of $[P]$ (see (3.3.1)) this mapping induces only a partial parametrization upon the characteristics in 2.1 of the observed semimartingale X. In fact only integrals of type

$$H \cdot A^\theta \quad \text{and} \quad W * \nu^\theta \quad \text{for} \quad H \in \mathscr{H} \quad \text{and} \quad W \in \mathscr{W}, \tag{3.3.3}$$

in particular $\hat{W}^\theta$ and $a^\theta = \hat{1}^\theta$, are fully parametrized: apart from integrands $H \in \mathscr{H}$ and $W \in \mathscr{W}$ they depend on a parameter value $\theta \in \Theta$ only. Here and elsewhere below we substitute the index P by θ whenever $P \in [P] = \vartheta^{-1}(\theta)$ for some $\theta \in \Theta$.

3.4 We want to stress that our knowledge of $\mathbb{P}$ is expressed by the finite dimensional parametrization (3.3.2) in terms of the functional form of the integrals (3.3.3) only, with integrands $H \in \mathscr{H}$ and $W \in \mathscr{W}$. The problem of identifying the sets $[P]$, $P \in \mathbb{P}$ is then equivalent to estimating θ. Therefore, we say that the family of $\mathbb{R}^k$-valued martingale transforms

$$M^\theta(H, W) = H \cdot X^{c\theta} + W * (\mu - \nu^\theta), \quad \theta \in \Theta \tag{3.4.1}$$

we will deal with in the sequel, is *admissible* for the above estimation problem if $H \in \mathscr{H}$ and $W \in \mathscr{W}$, that is $\mathbb{R}^k \times \mathbb{R}^d$-matrix valued H's and $\mathbb{R}^k$-vector valued W's in (3.4.1) consist of $\mathbb{R}^d$-valued columns in $\mathscr{H}$ and components in $\mathscr{W}$ respectively.

3.5 We close this section with an important observation. Suppose we have parametrized the integrals (3.3.3), that is we have specified the functional dependence of these integrals on θ. In the practical situation one does this for all $\omega \in \Omega$. In a more sophisticated way one might then say that all measures $P \in [P] = \vartheta^{-1}(\theta)$ solve a martingale problem that is formulated by imposing that the integrals (3.3.3) are compensators of certain processes. Hence we are in a sense in this section in a converse situation as in Section 2. There the measures P define on the sets Ω^P the characteristics which can be changed arbitrarily outside this set, whereas here we have candidates for the characteristics, depending on θ, and we assume that there are measures P such that under these measures the candidates are indeed versions of the characteristics. As the consequence of this set up the processes $g^{P,P'}$, now denoted by $g^{\theta,\theta'}$, are defined on the whole set Ω and we may assume that Eqs. (3.2.3) and (3.4.1) are also valid on the whole set Ω.

This approach can be applied also to the situation where the measures are mutually singular (as, for instance, in case of $X_t = \theta(t + w_t)$, where w is a standard Brownian motion under all measures in $[P] = \vartheta^{-1}(\theta)$: here A_t^θ is defined to be θt, and of course $M_t^\theta = X_t - \theta t$ is then a martingale under any $P \in [P] = \vartheta^{-1}(\theta)$).

4 ASYMPTOTIC DIFFERENTIABILITY

4.1 Let ϕ_t be a certain predictable $\mathbb{R}^k \times \mathbb{R}^k$-matrix valued symmetric positive definite process, used below as a norming factor. It may depend on the parameter θ but this is irrelevant in the present context; see definition 4.1.1(iii) where ϕ is specified, as well as another norming factor ψ which is of the same type, but unlike ϕ it may depend on particular $H \in \mathscr{H}$ and $W \in \mathscr{W}$ involved in definition 4.1.1, so that $\psi = \psi(H, W)$.

To a fixed $\theta \in \Theta$ relate the set of directions $\mathscr{U}_t = \phi_t^{-1}(\Theta - \theta)$, and assume for simplicity that a perturbation $\theta + \phi_t u$ considered below of a parameter value θ in a direction u is again a parameter value: $\theta + \phi_t u \in \Theta$. Furthermore, considering below any parametrized predictable process $\{a_t(\theta)\}$ we will always assume that $\{a_t(\theta + \phi_t u)\}$ is a well defined predictable process.

DEFINITION 4.1.1 For each fixed $\theta \in \Theta$ and each direction $u \in \mathscr{U}_t$ the compensators A^θ and v^θ are called *asymptotically differentiable* (weakly in $\mathscr{H}$ and $\mathscr{W}$, with norming factors ϕ and ψ) if there exist an $\mathbb{R}^k \times \mathbb{R}^d$-matrix valued predictable process $b^\theta \in \mathscr{H}$ and $\mathbb{R}^k$-vector valued predictable process $\lambda^\theta \in \mathscr{W}$ such that for each $\mathbb{R}^k \times \mathbb{R}^d$-matrix valued $H \in \mathscr{H}$ and $\mathbb{R}^k$-vector valued $W \in \mathscr{W}$ all integrals introduced below are well defined and in probability P for all $P \in [P] = \vartheta^{-1}(\theta)$ we have as $t \to \infty$ that

i) $\psi_t W * (v^{\theta+\phi_t u} - v^\theta)_t - \psi_t W \lambda^{\theta T} * v_t^\theta \phi_t u \to 0$,

ii) $\psi_t H \cdot (A^{\theta+\phi_t u} - A^\theta)_t - \psi_t (H c^\theta b^{\theta T} \cdot v_t + H x \lambda^{\theta T} * v_t^\theta) \phi_t u \to 0$ and

iii) the norming factors ϕ and ψ are such that $\Phi_t \to \Phi$ and $\Psi_t \to \Psi$ where Φ and $\Psi = \Psi(H, W)$ are certain non singular (random) matrices, while

$$\Phi = \langle \tilde{M} \rangle^{1/2} \phi \quad \text{and} \quad \Psi = \psi \langle M \rangle^{1/2}$$

with (cf. (2.4.2)–(2.4.4))

$$\langle M\rangle_t = \langle M^\theta(H, W)\rangle_t = Hc^\theta H^T \cdot v_t + WW^T * v_t^\theta - \sum_{s\le t} \hat{W}_s \hat{W}_s^T$$

and

$$\langle \tilde{M}\rangle_t = \langle \tilde{M}^\theta(b^\theta, \lambda^\theta)\rangle_t = b^\theta c^\theta b^{\theta T} \cdot v_t + \lambda^\theta \lambda^{\theta T} * v_t^\theta + \sum_{s\le t} 1_{\{a_s^\theta < 1\}} \hat{\lambda}_s^\theta \hat{\lambda}_s^{\theta T}/(1 - a_s^\theta).$$

Here and elswhere below we usually use the following arbridged notation

$$M = M^\theta = M^\theta(H, W) \quad \text{and} \quad \tilde{M} = \tilde{M}^\theta = \tilde{M}^\theta(b^\theta, \lambda^\theta).$$

4.2 The choice of the norming factors ϕ and ψ in (iii), with the same asymptotic behaviour as $\langle \tilde{M}\rangle^{-1/2}$ and $\langle M\rangle^{-1/2}$ respectively, is motivated as follows.

Define first $\mathring{A}^\theta = A^\theta - x * v^\theta$. Note that $-(\mathring{A}^{\theta'} - \mathring{A}^\theta) = X^{c\theta'} - X^{c\theta}$ on the set where (3.2.2) holds. Then (i) and (ii) in Definition 4.1.1 are equivalent to (i) and

(ii) $$\psi_t\{H \cdot (\mathring{A}^{\theta+\phi_t u} - \mathring{A}^\theta)_t - Hc^\theta b^{\theta T} \cdot v_t \phi_t u\} \to 0$$

in probability $P \in [P] = \vartheta^{-1}(\theta)$. Next, by (3.2.3)

$$g^{\theta,\theta+\phi_t u}(H, W)_t = H \cdot (A^{\theta+\phi_t u} - A^\theta)_t + U * (v^{\theta+\phi_t u} - v^\theta)_t, \quad U = W - Hx, \quad (4.2.1)$$

so that (i) and (ii′) are equivalent to

$$\psi_t g^{\theta,\theta+\phi_t u}(H, W)_t - \psi_t \langle M, \tilde{M}\rangle_t \phi_t u \to 0 \quad \text{in probability} \quad P \in [P] = \vartheta^{-1}(\theta) \quad (4.2.2)$$

with

$$\langle M, \tilde{M}\rangle = \langle M^\theta(H, W), \tilde{M}^\theta(b^\theta, \lambda^\theta)\rangle = Hc^\theta b^{\theta T} \cdot v + W\lambda^{\theta T} * v^\theta;$$

cf. (2.4.2) and (2.3.2) with $\tilde{\lambda}^\theta = \lambda^\theta + 1_{\{a^\theta<1\}}\hat{\lambda}^\theta/(1 - a^\theta)$. Due to (2.4.1) and (4.2.1)

$$M^{\theta+\phi_t u}(H, W) - M^\theta(H, W) = -g^{\theta,\theta+\phi_t u}(H, W),$$

hence (4.2.2) in turn is equivalent to

$$\psi_t \xi^{\theta,\theta+\phi_t u}(H, W)_t \to 0 \quad \text{as} \quad t \to \infty \quad \text{in probability} \quad P \in [P] = \vartheta^{-1}(\theta) \quad (4.2.3)$$

where

$$\xi^{\theta,\theta'}(H, W) = M^{\theta'}(H, W) - M^\theta(H, W) + \langle M, \tilde{M}\rangle(\theta' - \theta). \quad (4.2.4)$$

Thus, we have shown that the following statement is true.

STATEMENT 4.2.1 *The asymptotic differentiability at* $\theta \in \Theta$ *and each direction* $u \in \mathscr{U}_t$ *of the compensators* A^θ *and* ν^θ (in the sense of Definition 4.1.1) *is equivalent to* (4.2.3).

Turning back to the choice of the norming factors ϕ and ψ, observe that since we are interested in *weak* differentiability of the functionals A^θ and ν^θ (or equivalently, $\mathring{A}^\theta$ and ν^θ) acting on (H, W), the natural scaling of the differences $H \cdot (\mathring{A}^{\theta'} - \mathring{A}^\theta)$ and $W * (\nu^{\theta'} - \nu^\theta)$ (or equivalently of the difference $M^{\theta'} - M^\theta = -g^{\theta, \theta'}$) should be related to an L^2-norm of the pair (H, W). The reasonable choice is then a positive definite square root of the predictable process $\langle M \rangle$; cf. (4.2.3). This explains the choice of ψ. Furthermore, in order to give the weak asymptotic derivative a sensible meaning the norming process ϕ has to be such that the scaled difference

$$\psi_t \{ M^{\theta + \phi_t u}(H, W)_t - M^\theta(H, W)_t \}$$

is bounded by a finite random variable $P \in [P] = \vartheta^{-1}(\theta)$-a.s. But then, if differentiability (that is (4.2.3)) holds also $\psi_t \langle M, \tilde{M} \rangle_t \phi_t$ is bounded in the same way. Again, exploiting the fact that we require *weak* differentiability, we have to choose the norming ϕ such that these quantities are bounded no matter what H and W are. But then, using the Schwartz inequality for matrices the only way to guarantee this is by choosing ϕ such that $\phi_t \langle \tilde{M} \rangle_t \phi_t$ is bounded as in (iii). Certainly, to make the notion of differentiability the strongest possible we should require that ϕ_t tends to zero, but not too fast, otherwise this would render the notion vacuous.

For the sake of simplicity the norming factors ϕ and ψ in (iii) will be identified below with $\langle \tilde{M} \rangle^{-1/2}$ and $\langle M \rangle^{-1/2}$ respectively, as the necessary modifications to the general case are obvious. The relation (4.2.2) for instance can be rewritten then as follows:

$$\psi_t g^{\theta, \theta + \phi_t u}(H, W)_t - \rho(M, \tilde{M})_t u \to 0 \quad \text{in probability} \quad P \in [P] = \vartheta^{-1}(\theta)$$

where $\rho = \rho(M, \tilde{M})$ is the correlation process between M and $\tilde{M}$; see (2.5.2).

4.3 As was mentioned in Section 3.2, in the specific situation in which the model is fully parametrized and all measures involved are mutually locally absolutely continuous, Girsanov's theorem applies and the process $g^{\theta, \theta'}$ is Girsanov's correction term. Hence for instance b^θ in Definition 4.1.1 is the derivative in the above sense of β^θ that replaces β in the definition of the martingale $\tilde{M}^P(\beta, Y - 1)$. Similarly, in this case λ^θ can be interpreted as "logarithmic derivative" of ν^θ. Moreover, $\langle \tilde{M} \rangle$ is the genuine Fisher information process (see Jacod (1990)).

5 ADMISSIBLE ESTIMATORS

5.1 To estimate the unknown parameter value $\theta \in \Theta \subset \mathbb{R}^k$ at time instant t, a certain class of $\mathscr{F}_t$-adapted statistics, say $\{\hat{\theta}_t\}$, is considered as a class of potential estimators. We consider here an asymptotic setting of the estimation problem by assuming that when $t \to \infty$ an estimator $\hat{\theta}_t$ "estimates" the unknown parameter value θ in the sense that the appropriately scaled difference $\mathscr{B}_t(\hat{\theta}_t - \theta)$ has a non degenerate limit

distribution, where the scaling $\mathscr{B}_t$ is a $\mathbb{R}^k \times \mathbb{R}^k$-matrix valued predictable process, non singular $P \in [P] = \vartheta^{-1}(\theta)$-a.s. for t large enough (depending usually on the parameter θ but this is irrelevant in the present context). For the sake of generality, however, we do not exclude the possibility of a certain bias in estimation by taking into consideration also estimators $\hat{\theta}_t$ for which the limiting distribution of the scaled difference $\mathscr{B}_t(\hat{\theta}_t - a_t(\theta))$ is non degenerate with a certain deterministic function $a_t: \Theta \to \mathbb{R}^k$ for each fixed t, violating the condition

$$\mathscr{B}_t(\theta - a_t(\theta)) \to 0 \quad \text{as} \quad t \to \infty \quad \text{in probability} \quad P \in [P] = \vartheta^{-1}(\theta). \tag{5.1.1}$$

We will say that such estimators $\hat{\theta}_t$ are (asymptotically) *biased*. The difference

$$d_t(\theta) = \theta - a_t(\theta) \tag{5.1.2}$$

will be called the (asymptotic) bias of $\hat{\theta}_t$. Accordingly, we will say that an estimator $\hat{\theta}_t$ is (asymptotically) *unbiased* if it "estimates" θ in the sense mentioned above, i.e. if $\mathscr{B}_t(\hat{\theta}_t - \theta)$ has a non degenerate limit distribution.

5.2 In this paper we will restrict our attention to estimators called *admissible* as they will be represented below by means of admissible martingales; cf. (3.4.1). Note meanwhile that by this representation we will associate with a particular admissible martingale $M^\theta(H, W)$, for fixed $H \in \mathscr{H}$ and $W \in \mathscr{W}$, a set of asymptotically equivalent estimators $[\hat{\theta}_t(H, W)]$. The corresponding $H \in \mathscr{H}$ and $W \in \mathscr{W}$ are usually called the *scoring functions.*

DEFINITION 5.2.1 Let $\mathscr{B}_t$ be as above, and $\mathscr{A}_t(\theta)$, $\theta \in \Theta$ and $\mathbb{R}^k$-vector valued $\mathscr{F}_t$-adapted process for each fixed $\theta \in \Theta$. An (asymptotically) unbiased estimator $\hat{\theta}_t$ of θ is called *admissible* if it is representable for each fixed $\theta \in \Theta$ by means of an admissible martingale $M^\theta(H, W) = M^\theta$ as follows:

$$\mathscr{B}_t(\hat{\theta}_t - \mathscr{A}_t(\theta)) = \langle M^\theta \rangle_t^{-1/2} M_t^\theta \tag{5.2.1}$$

with some $\mathscr{A}_t$ and $\mathscr{B}_t$ such that

$$\delta(\theta)_t = \mathscr{B}_t(\mathscr{A}_t(\theta) - \theta) \to 0 \quad \text{as} \quad t \to \infty \quad \text{in probability} \quad P \in [P] = \vartheta^{-1}(\theta). \tag{5.2.2}$$

An admissible (asymptotically) biased estimator $\hat{\theta}_t$ with the bias (5.1.2) is defined similarly but with

$$\delta(\theta)_t = \mathscr{B}_t(\mathscr{A}_t(\theta) - a_t(\theta)) \to 0 \quad \text{as} \quad t \to \infty \quad \text{in probability} \quad P \in [P] = \vartheta^{-1}(\theta) \tag{5.2.3}$$

instead of (5.2.2).

Obviously, (5.2.1) and (5.2.2) (or (5.2.1) and (5.2.3)) are equivalent to

$$B_t(\hat{\theta}_t - \theta) = M_t^\theta + \eta(\theta)_t \quad (\text{or } B_t(\hat{\theta}_t - a_t(\theta)) = M_t^\theta + \eta(\theta)_t) \tag{5.2.4}$$

where

$$\psi_t B_t = \mathscr{B}_t \quad \text{and} \quad \psi_t \eta(\theta)_t = \delta(\theta)_t \to 0 \tag{5.2.5}$$

as in (5.2.2) (or (5.2.3)). Recall that $\psi = \langle M^\theta \rangle^{-1/2}$; cf. the last paragraph in 4.2. Of course, if the asymptotic bias (5.1.2) is small in the sense of (5.1.1), then the two expressions in (5.2.4) are equivalent.

5.3 By the assumptions imposed in 3.1 on the right hand side of the representation (5.2.1) the scaling factor $\mathscr{B}$ characterizes the convergence rate of the estimator $\hat{\theta}$, and for t large enough the ellipsoid generated by the inverse of the symmetric matrix

$$\mathscr{B}_t^T \mathscr{B}_t = B_t^T \langle M^\theta \rangle_t^{-1} B_t \tag{5.3.1}$$

characterizes the *spread* of $\hat{\theta}$ around θ.

DEFINITION 5.3.1 Let $\hat{\theta}$ be an admissible estimator of θ for each fixed $\theta \in \Theta$ (see Definition 5.2.1). For fixed t large enough the ellipsoid generated by the inverse of the matrix (5.3.1) with $\mathscr{B}$ involved in (5.2.1) is called the *spread* of $\hat{\theta}$ around θ or $a_t(\theta)$ depending whether $\hat{\theta}$ is asymptotically unbiased or biased. In the latter case the *spread* of $\hat{\theta}$ around θ is defined as the ellipsoid generated by the matrix

$$\sum_t(\theta) = (\mathscr{B}_t^T \mathscr{B}_t)^{-1} + d_t(\theta) d_t(\theta)^T, \tag{5.3.2}$$

for the bias (5.1.2) which violates the Condition (5.1.1) has to be taken into account.

5.4 Denote

$$D = B^{-1} \langle M, \tilde{M} \rangle - I \tag{5.4.1}$$

where $\tilde{M} = \tilde{M}^\theta(b^\theta, \lambda^\theta)$ as in 4.1. By Lemma 2.5.1 we have

$$c(\tilde{M}, M) = \langle \tilde{M} \rangle - \langle \tilde{M}, M \rangle \langle M \rangle^{-1} \langle M, \tilde{M} \rangle \geq 0.$$

Therefore

$$(\mathscr{B}^T \mathscr{B})^{-1} = (I + D)(\langle \tilde{M} \rangle - c(\tilde{M}, M))^{-1} (I + D)^T \geq (I + D) \langle \tilde{M} \rangle^{-1} (I + D)^T. \tag{5.4.2}$$

This means that the spread of $\hat{\theta}_t$ around θ (or $a_t(\theta)$) exceeds the ellipsoid generated by the matrix on the right hand side of the last inequality. This lower bound for the spread around θ (or $a_t(\theta)$) of any admissible estimator lies at the basis of the Cramer–Rao inequality which will be obtained in Section 6. Meanwhile, even the spread of supperefficient estimators satisfy (5.4.2). In order to exclude such abnormalities and, consequently, render the inequality (5.4.2) in the usual Cramer–Rao form, we shall, according to the common practice, restrict the class of estimators by certain

regularity assumptions; see 6.1 below. Since $\langle \tilde{M} \rangle$ can be interpreted in the present setting as the *Fisher information matrix* (see Section 4.3), we say that the inequality (5.4.2) takes usual Cramer–Rao form if the matrix D on the right hand side is replaced by the "derivative" (in the sense of Remark 6.1.2 below) of the bias (5.1.2) with respect to θ; see (6.2.2) below. Consider for instance the following situation. We will return to this situation in Section 6.2.

5.5 The inequality (5.4.2) already gives the desired Cramer–Rao lower bound for estimators admitting the representation

$$\langle M, \tilde{M} \rangle_t (\hat{\theta}_t - \theta) = M_t^\theta + \eta(\theta)_t, \tag{5.5.1}$$

i.e. the representation (5.2.4) with special $B = \langle M, \tilde{M} \rangle$ and $a(\theta) = \theta$, for in this case $D = 0$ and hence

$$(\mathscr{B}^T \mathscr{B})^{-1} \geq \langle \tilde{M} \rangle^{-1}. \tag{5.5.2}$$

Note that the matrix valued process D is related to the bias of an admissible estimator $\hat{\theta}$ in the following sense. If an estimator satisfies (5.5.1), then $D = 0$ and $d(\theta) = 0$, so that by (4.2.4) and (5.2.4) it also satisfies the following relation: for each $\theta' \in \Theta$

$$\langle M, \tilde{M} \rangle_t (\hat{\theta}_t - \theta') = M^{\theta'}(H, W)_t - \xi^{\theta, \theta'}(H, W)_t + \eta(\theta)_t. \tag{5.5.3}$$

Next, evaluate (5.5.3) at $\theta' = \theta + \phi_t u$ under condition (4.2.3) to see that $\xi^{\theta, \theta + \phi_t u}$ has the same behaviour when $t \to \infty$ as $\eta(\theta)$, i.e. it can be absorbed in the remainder term. Thus the estimator $\hat{\theta}$ has the linear representation not only at θ but also in its neighbourhood $\theta' = \theta + \phi_t u$.

Now, assume D does not vanish, then the bias appears in the representation, as even if $d(\theta) = 0$ we get by (4.2.4) and (5.2.4) that

$$B_t(\hat{\theta}_t - [\theta' + D_t(\theta' - \theta)]) = M^{\theta'}(H, W)_t - \xi^{\theta, \theta'}(H, W)_t + \eta(\theta)_t$$

where $-\xi^{\theta, \theta'}(H, W) + \eta(\theta)$ at $\theta' = \theta + \phi_t u$ can be considered as a remainder term.

5.6 According to Lemma 2.5.1, we get equality in (5.4.2) iff $\tilde{M} = CM$ with some random matrix C, not depending on time. Hence equality in (5.4.2) is only attained for estimators that have the representation (5.2.4) of the following special form:

$$CB_t(\hat{\theta}_t - \theta) = \tilde{M}_t^\theta + C\eta(\theta)_t.$$

Notice that $C\eta(\theta)_t$ is indeed a remainder term in the sense of (5.2.5): since now

$$\langle \tilde{M} \rangle = C \langle M \rangle C^T$$

(see Dzhaparidze and Spreij (1992)), we immediately get

$$(C\eta(\theta)_t)^T \langle \tilde{M} \rangle_t^{-1} C\eta(\theta)_t \to 0 \quad \text{as} \quad t \to \infty \quad \text{in probability} \quad P \in [P] = \vartheta^{-1}(\theta).$$

6 REGULAR ESTIMATORS. THE CRAMER–RAO INEQUALITY

6.1 As is known in classical statistics, the minimization problem of the spread of an estimator by proving the Cramer–Rao inequality (see, e.g. Ibragimov and Has'minskii (1981), or in a more related context, Kutoyants (1984) and Jacod (1990), as well as the references therein) can be effectively solved only under certain regularity conditions imposed on estimators. In fully (or partially) specified models with LAN property, more sophisticated Hajek's type regularity is required. As our parametrization in Section 3 admits such models only as special cases, and besides our assumptions are too wide to admit establishing asymptotic distributions of estimators, Hajek's definition cannot be taken over here. However our definition of regularity can be, in principle, considered as a wide sense version of Hajek's regularity, but we don't dwell on this subject here. On the other hand our scheme includes also classical regression models in which the Gauss–Markov estimator has minimal spread as $t \to \infty$ among asymptotically linear unbiased estimators (there is a certain relationship between our definition of regularity and the asymptotic unbiasedness of estimators, but again we prefer to avoid this subject here).

The common idea hidden behind any definition of regularity of estimators representable as in (5.2.4) consists, roughly speaking, in admitting differentiability in a certain appropriate sense of the both sides of the representation. Our Definition 6.1.1 below is also based on this consideration. Namely, the class of all admissible estimators $\{[\hat{\theta}_t(H, W)],\ H \in \mathscr{H} \text{ and } W \in \mathscr{W}\}$ with the scoring functions $H \in \mathscr{H}$ and $W \in \mathscr{W}$ is restricted by the regularity assumption: an estimator $\hat{\theta}_t(H, W) = \hat{\theta}_t$ with the scoring functions $H \in \mathscr{H}$ and $W \in \mathscr{W}$ assumes the representation of type (5.2.4) not only at a fixed $\theta \in \Theta$ but also at $\theta + \phi_t u \in \Theta$ with the same ϕ as in 4.1(iii) and all directions $u \in \mathscr{U}_t$.

DEFINITION 6.1.1 An estimator $\hat{\theta}_t(H, W)$ of the value θ with scoring functions $H \in \mathscr{H}$ and $W \in \mathscr{W}$, is called *regular* (with a centering a and scaling B) if

i) it is representable at each fixed $\theta \in \Theta$ in all directions $u \in \mathscr{U}_t$ as follows

$$B_t(\hat{\theta}_t - a_t(\theta + \phi_t u)) = M^{\theta+\phi_t u}(H, W)_t + \eta(u, \theta)_t \tag{6.1.1}$$

where $\phi = \langle \tilde{M} \rangle^{-1/2}$;

ii) the remainder term $\eta(u, \theta)_t$ (depending on $H \in \mathscr{H}$ and $W \in \mathscr{W}$ of course) is such that as $t \to \infty$

$$\psi_t \eta(u, \theta)_t \to 0 \quad \text{in probability} \quad P \in [P] = \vartheta^{-1}(\theta) \tag{6.1.2}$$

where $\psi = \langle M^\theta \rangle^{-1/2}$;

iii) there exists a function $\mathring{a}$: $[0, \infty) \times \Theta \to \mathbb{R}^d \times \mathbb{R}^d$ such that the following holds for each fixed $\theta \in \Theta$ and all directions $u \in \mathscr{U}_t$:

$$\mathscr{B}_t \alpha_t^{\theta, \theta+\phi_t u} \to 0 \quad \text{as} \quad t \to \infty \quad \text{in probability} \quad P \in [P] = \vartheta^{-1}(\theta) \tag{6.1.3}$$

where

$$\alpha^{\theta,\theta'} = a(\theta') - a(\theta) - \mathring{a}(\theta)(\theta' - \theta). \tag{6.1.4}$$

Remark 6.1.2 The matrix $\mathring{a}$ is the "asymptotic derivative" of a with respect to θ. Notice that if the bias (5.1.2) is small in the sense of (5.1.1) not only at a fixed θ but also in a shrinking neighbourhood, that is we may replace here θ with $\theta + \phi_t u$, then in (6.1.3) we may take $\mathring{a}_t(\theta) = I$.

Remark 6.1.3 As $u = 0$ we get (5.2.4) with an admissible $M^\theta(H, W)$. Note that by (5.2.5) and (6.1.2) the remainder term $\eta(u, \theta)_t$ in (6.1.1) is asymptotically differentiable (in accordance with Definition 4.1) with derivative equal to zero in the sense that for each fixed $\theta \in \Theta$ and each direction $u \in \mathscr{U}_t$ we have that in probability $P \in [P] = \mathfrak{I}^{-1}(\theta)$

$$\psi_t\{\eta(u, \theta)_t - \eta(\theta)_t\} \to 0.$$

Conversely, if an estimator $\hat{\theta}_t$ satisfies (5.2.4) with an addmissible $M^\theta(H, W)$ and certain $\eta(\theta)$, then it satisfies also (6.1.1) with $\eta(u, \theta)$ such that

$$\begin{aligned} \psi\{\eta(u, \theta) - \eta(\theta)\} &= -\psi\{M^{\theta'} - M^\theta\} - \mathscr{B}\{a(\theta') - a(\theta)\} \\ &= \psi\xi^{\theta, \theta'} - \mathscr{B}\alpha^{\theta, \theta'} + \varepsilon u \end{aligned} \tag{6.1.5}$$

where $\theta' = \theta + \phi_t u$ and

$$\varepsilon = \rho - \mathscr{B}\mathring{a}(\theta)\phi \tag{6.1.6}$$

with $\rho = \rho(M, \tilde{M})$; cf. (2.5.2), (4.2.4) and (6.1.4). Therefore the following statement is true.

STATEMENT 6.1.4 *If the compensators A^θ and v^θ are asymptotically differentiable at $\theta \in \Theta$ and each direction $u \in \mathscr{U}_t$ in the sense of Definition* 4.1.1, *then by Statement* 4.2.1 *the regularity of $\hat{\theta}$ is equivalent to the condition that the centering a and scaling B involved in its representation are such that as $t \to \infty$ the last term on the right hand side of* (6.1.5) *vanishes, that is $\varepsilon_t \to 0$ in probability $P \in [P] = \mathfrak{I}^{-1}(\theta)$ where ε is given by* (6.1.6).

6.2 As in Section 5.5, suppose an estimator $\hat{\theta}$ satisfies the first of expressions (5.2.4) with special $B = \langle M, \tilde{M}\rangle$ which by (5.4.1) means $D = 0$, or more generally it satisfies the second of expressions (5.2.4) where $a(\theta)$ is differentiable in the sense of (6.1.3) with a special B such that

$$\mathring{a}(\theta) = I + D = B^{-1}\langle M, \tilde{M}\rangle. \tag{6.2.1}$$

It follows then from (6.1.5) and (6.1.6) that under the conditions of Statement 6.1.4 the estimator $\hat{\theta}$ is regular. Moreover, by (5.4.2) its spread around $a(\theta)$ satisfies the Cramer–Rao inequality

$$(\mathscr{B}^T\mathscr{B})^{-1} \geq \mathring{a}\langle\tilde{M}\rangle^{-1}\mathring{a}^T; \tag{6.2.2}$$

cf. (5.5.2). Therefore the following question is important. Suppose $\hat{\theta}$ is regular. Does it then admit the representation (5.2.4) with B satisfying (6.2.1)? The answer given in the next section turns out to be affirmative if we sharpen Definition 4.1.1 a little bit.

6.3 Consider a regular estimator $\hat{\theta}(H, W)$ and suppose that it satisfies (6.1.1) not only with $\phi = \langle \tilde{M} \rangle^{-1/2}$ but also with $\phi' = \phi\rho^{-1}$ where $\rho = \rho(M, \tilde{M})$; cf. (2.5.2). Note that $\phi\phi^T = \phi'\rho\rho^T\phi'^T \leq \phi'\phi'^T$ by (2.5.3). Hence the perturbation rate ϕ' is not faster then ϕ. Of course, if the correlation process ρ stays bounded away from zero there is no need in introducing ϕ'. Next, suppose that for the scoring functions H and W the relation (i) and (ii) in Definition 4.1 are satisfied with ϕ' instead of ϕ, that is the differentiability still takes place, despite the above remark that the perturbation rate ϕ' is not faster then ϕ.

PROPOSITION 6.3.1 *Under the conditions of the present section*

i) *$\hat{\theta}$ admits also the following representation*

$$\langle M, \tilde{M} \rangle_t \mathring{a}_t(\theta)^{-1}(\hat{\theta}_t - a_t(\theta)) = M^\theta(H, W)_t + \eta'(\theta)_t \tag{6.3.1}$$

where $\eta'(\theta)$ is again a remainder term:

$$\psi_t \eta'(\theta)_t \to 0 \quad \text{in probability} \quad P \in [P] = \vartheta^{-1}(\theta) \quad \text{as} \quad t \to \infty. \tag{6.3.2}$$

ii) *The spread of $\hat{\theta}$ around $a(\theta)$ satisfies the Cramer–Rao inequality* (6.2.2).

Proof Assertion (ii) follows from (i) since by definition the spread of $\hat{\theta}$ around $a(\theta)$ is generated by the matrix $\Gamma\Gamma^T$ where

$$\Gamma = \mathring{a}\langle M, \tilde{M} \rangle^{-1} \langle M \rangle^{1/2}, \tag{6.3.3}$$

so that

$$\Gamma\Gamma^T - \mathring{a}\langle \tilde{M} \rangle^{-1} \mathring{a}^T \geq \mathring{a}\langle M, \tilde{M} \rangle^{-1} c(M, \tilde{M}) \langle \tilde{M}, M \rangle^{-1} \mathring{a}^T \geq 0$$

by Lemma 2.5.1.

Let us now prove assertion (i). By (5.2.4) and (6.3.1) we get

$$\psi\eta' = \psi\eta + \zeta\psi(M + \eta) \quad \text{where} \quad \zeta = \varepsilon\rho^{-1}(I - \varepsilon\rho^{-1})^{-1}$$

with the same ε and ρ as in (6.1.6). Similar to (6.1.6) we immediately obtain that $\varepsilon\rho^{-1}$ and hence ζ tends to zero in probability $P \in [P] = \vartheta^{-1}(\theta)$ as $t \to \infty$. Therefore (6.3.3) yields (6.3.2), for by assumption η is a remainder term and $\{\psi_t M_t\}$ is a tight family (cf. Section 3.1).

6.4 Let us turn back to the general case where $\hat{\theta}$ is a regular estimator in the sense of Definition 6.1.1. Regarding a lower bound to the asymptotic spread the following proposition is true.

PROPOSITION 6.4.1 *Let $\hat{\theta}$ be a regular estimator with the spread generated by $(\mathscr{B}^T\mathscr{B})^{-1}$. Let the compensators A^θ and v^θ be asymptotically differentiable at $\theta \in \Theta$ and each*

direction $u \in \mathscr{U}_t$ in the sense of Definition 4.1.1. *Then for any symmetric positive definite matrix $\delta > 0$ the event*

$$\mathscr{B}_t\{(\mathscr{B}_t^T\mathscr{B}_t)^{-1} - \mathring{a}_t\langle\tilde{M}_t\rangle^{-1}\mathring{a}_t^T\}\mathscr{B}_t^t \geq -\delta \tag{6.4.1}$$

takes place with $P \in [P] = \vartheta^{-1}(\theta)$ probability tending to one as $t \to \infty$.

Proof With the notations used in Statement 6.1.4 we have

$$\mathscr{B}\mathring{a}\langle\tilde{M}\rangle^{-1/2} = \rho + \varepsilon.$$

Therefore the event (6.4.1) is equivalent to

$$(\rho_t + \varepsilon_t)(\rho_t + \varepsilon_t)^T \leq I + \delta,$$

so that the desired assertion follows from (2.5.3) and Statement 6.1.4 according to which $\rho_t\rho_t^T \leq I$ and $\varepsilon_t \to 0$ in probability $P \in [P] = \vartheta^{-1}(\theta)$.

7 OPTIMALITY

7.1 Throughout this section the compensators A^θ and ν^θ are asymptotically differentiable at $\theta \in \Theta$ and each direction $u \in \mathscr{U}_t$ in the sense of Definition 4.1.1, and all estimators mentioned are admissible in the sense of Definition 5.2.1.

The assertions of Propositions 6.3.1 and 6.4.1 can be interpreted as follows: the minimal possible spread around $a_t(\theta)$ of a regular estimator is generated by the matrix

$$\mathring{a}_t\langle\tilde{M}_t\rangle^{-1}\mathring{a}_t^T$$

where $\tilde{M} = \tilde{M}^\theta(b^\theta, \lambda^\theta)$ as usual. Hence the following definition

DEFINITION 7.1.1 A regular estimator $\hat{\theta}(H, W)$ is called *optimal* if it can be represented as in (5.2.4) with $a_t(\theta) = \theta$, and if the scoring functions $H \in \mathscr{H}$ and $W \in \mathscr{W}$ are such that the spread attains the lower bound which in this case is generated by $\langle\tilde{M}_t\rangle^{-1}$.

PROPOSITION 7.1.2 *A regular estimator $\hat{\theta}$ is optimal in the sense of Definition* 7.1.1 *iff it admits the following special form of the general representation* (5.2.4):

$$\langle\tilde{M}^\theta\rangle_t(\hat{\theta}_t - \theta) = \tilde{M}_t^\theta + \eta(\theta)_t. \tag{7.1.1}$$

Proof By definition the spread of an estimator $\hat{\theta}$ admitting the Representation (7.1.1) is generated by $\langle\tilde{M}\rangle^{-1}$, i.e. $\hat{\theta}$ is optimal.

Conversely, if $\hat{\theta}$ is optimal, then its spread is generated by the inverse of the matrix

$$B^T\langle M\rangle^{-1}B = \langle\tilde{M}\rangle. \tag{7.1.2}$$

In notations of (6.1.6) the equality (7.1.2) means that

$$(\rho - \varepsilon)^T(\rho - \varepsilon) = I. \tag{7.1.3}$$

Since $\hat{\theta}$ is regular, Statement 6.1.4 is true: $\varepsilon_t \to 0$ in probability $P \in [P] = \vartheta^{-1}(\theta)$. Hence $\rho_t^T \rho_t \to I$ by (7.1.3). Therefore we can apply Proposition 6.3.1, assertion (i) to write down the following representation for $\hat{\theta}$:

$$\langle M, \tilde{M}\rangle(\hat{\theta} - \theta) = M + \eta. \tag{7.1.4}$$

Thus $B = \langle M, \tilde{M}\rangle$ in (7.1.2), that is $c(\tilde{M}, M)_t = 0$ for all t. By Lemma 2.5.1 this implies $\tilde{M} = CM$ with a possibly random matrix C independent of t. Hence (7.1.4) can be rewritten as follows

$$C\langle M, \tilde{M}\rangle(\hat{\theta} - \theta) = \tilde{M} + C\eta; \tag{7.1.5}$$

cf. (5.5.1). Since $C\langle M, \tilde{M}\rangle = \langle \tilde{M}\rangle$ (see Dzhaparidze and Spreij (1992)) and $C\eta$ is a remainder term (see Section 5.5), (7.1.5) yields (7.1.1).

Acknowledgements

This paper is largely expository in nature and reflects the viewpoint of the authors on the presented subject, discussed with R. J. Chitashvili and J. Jacod at various stages of its preparation.

References

O. Barndorff-Nielsen and M. Sørensen. Asymptotic likelihood theory for stochastic processes. A review. *Research Report* no. 162. Department of Theoretical Statistics, Institute of Mathematics, University of Aarhus, (1989).

R. J. Chitashvili, N. L. Lazrieva and T. A. Toronjadze. Asymptotic theory of M-estimators in general statistical models, *Reports* BS R9019-20, CWI, Amsterdam, (1990).

K. Dzhaparidze and P. Spreij. On correlation calculus for multivariate martingales, to appear in *Stochastic Processes and their Applications*, (1992).

V. P. Godambe, C. Heyde. Quasi-likelihood and optimal estimation. *Intern. Statist. Review* **55** (1990), 231–244.

P. E. Greenwood and W. Wefelmeyer. Partially and fully specified semimartingale models and efficiency. *Preprints in Statistics.* University of Cologne, (1989).

A. A. Gushchin. Cramer–Rao type inequality for a filtered space. *Preprint.* (1990).

I. A. Ibragimov and R. Z. Has'minskii. *Statistical Estimation-Asymptotic Theory.* Berlin, Springer, (1981).

J. Jacod. Regularity, partial regularity, partial information process for a filtered statistical model. *Probab. Theory and Related Fields* **85** (1990), 305–336.

J. Jacod and A. N. Shiryaev, *Limit Theorems for Stochastic Processes.* Springer, New York, (1987).

Yu. A. Kutoyants. *Parameter Estimation for Stochastic Processes.* Heldermann Verlag, Berlin, (1984).

R. Sh. Liptser and A. N. Shiryayev, *Theory of Martingales.* Dordrecht, Kluwer Academic Publ. (1989).

M. Sørensen. On quasi likelihood for semimartingales, *Stochastic Processes and their Applications.* **35** (1990), 331–346.