

# A PARADOXICAL EFFECT OF NUISANCE PARAMETERS ON EFFICIENCY OF ESTIMATORS

Masayuki Henmi \*

## Abstract

This paper is concerned with parameter estimation in the presence of nuisance parameters. Usually, an estimator with known nuisance parameters is better than that with unknown nuisance parameters in reference to the asymptotic variance. However, it has been noted that the opposite can occur in some situations. In this paper we elucidate when and how this phenomenon occurs using the orthogonal decomposition of estimating functions. Most of the examples of this phenomenon are found in the case of semiparametric models, but this phenomenon can also occur in parametric models. As an example, we consider the estimation of the dispersion parameter in a generalized linear model. <sup>1 2</sup>

## 1 Introduction

In a statistical model with a number of parameters, only a portion of the parameters are often of interest. The rest are nuisance parameters. Let  $\mathcal{M} = \{p(x; \beta, \alpha)\}$  be a parametric model whose elements are specified by a vector of parameters of interest  $\beta$  and a vector of nuisance parameters  $\alpha$ . Then it is well known that under some regularity conditions the following inequality holds,

$$\text{Var}_A(\tilde{\beta}) \leq \text{Var}_A(\hat{\beta}), \quad (1)$$

where  $\tilde{\beta}$  and  $\hat{\beta}$  are the maximum likelihood estimators of  $\beta$  with known and unknown  $\alpha$  respectively.  $\text{Var}_A$  denotes the asymptotic covariance matrix of an estimator. For two symmetric matrices  $A$  and  $B$ ,  $A \leq B$  indicates that  $B - A$  is a positive semi-definite

---

\*Department of Statistical Science, The Graduate University for Advanced Studies, 4-6-7 Minami-azabu, Minato-ku, Tokyo 106-8569, Japan

<sup>1</sup>Running Head: A PARADOXICAL EFFECT OF NUISANCE PARAMETERS

<sup>2</sup>Keywords: Asymptotic variance, Estimating function, Nuisance parameter, Optimality, Orthogonal decomposition, Semiparametric model

matrix. However, inequality (1) is not always observed if we do not use the maximum likelihood method. Let  $\mathcal{M} = \{p(x; \beta, \alpha, k)\}$  be a semiparametric model with an infinite-dimensional nuisance parameter  $k$  as well as a vector of parameters of interest  $\beta$  and a vector of nuisance parameters  $\alpha$ . Then, in a certain special case, we can observe the inequality opposite to (1),

$$\text{Var}_A(\tilde{\beta}) \geq \text{Var}_A(\hat{\beta}),$$

when  $\beta$  is estimated by an estimating function depending on  $\alpha$ . Here,  $\tilde{\beta}$  and  $\hat{\beta}$  are estimators of  $\beta$  when  $\alpha$  is known and when  $\alpha$  is unknown and estimated respectively. In other words, the estimator with unknown nuisance parameters is better than that with known ones with respect to the asymptotic variance. We call this unusual phenomenon the inverse phenomenon of asymptotic variances. For example, Robins, Mark and Newey (1992) proposed a semiparametric model for causal inference and pointed out that this phenomenon can occur in their model. Moreover, these kinds of phenomena have also been noted in some other situations (Robins, Rotnitzky and Zhao, 1994, Lawless and Kalbfleisch, 1999). See also Fourdrinier and Strawderman (1996) for shrinkage estimation. The aim of this paper is to explore the structure of the inverse phenomenon of asymptotic variances systematically by examining estimating functions. Specifically, we focus on the orthogonal decomposition of estimating functions. This decomposition is obtained by decomposing an estimating function to the component in the space of optimal estimating functions and the component in its orthogonal complement. Here, optimal estimating functions mean that the estimators given by them have the minimum asymptotic variance of all estimating functions. The inverse phenomenon of asymptotic variances can occur when an estimating function for parameters of interest with known nuisance parameters is not optimal. Considering the orthogonal decomposition of estimating functions helps us elucidate how estimating nuisance parameters improves the asymptotic variance of estimators for parameters of interest.

This paper is organized as follows. In Section 2 we introduce the semiparametric model proposed by Robins *et al.* (1992) as an illustrative example. In Section 3 we describe the orthogonal decomposition of estimating functions for semiparametric models. Section 4 examines the structure of the inverse phenomenon using orthogonal decomposition. In Section 5 the parametric case is considered. The inverse phenomenon of asymptotic variances can also occur in parametric models if the maximum likelihood estimation method is not used. Its structure is essentially the same as in the semipara-

metric case. As an example we consider estimation of the dispersion parameter in a generalized linear model. Finally, in Section 6 we give some concluding remarks.

## 2 Illustrative example

In this section, we give an illustrative example of the inverse phenomenon of asymptotic variances. We would like to estimate the causal effect of an exposure or treatment on an outcome of interest. In this case, as is widely known, if we ignore the effect of confounding factors that both covary with the exposure or treatment and are independent predictors of the outcome, the estimate of the causal effect is biased. Let  $Y, S$  and  $X = (X_2, \dots, X_K)$  be respectively a continuous outcome variable of interest, an indicator of exposure which takes the value of 1 when the subject is exposed and 0 otherwise, and a vector of variables of confounding factors. The following model proposed by Robins *et al.* (1992) is a semiparametric regression model to estimate the causal effect by adjusting for confounding factors,

$$Y = \beta S + h(X) + \epsilon, \quad \text{E}[\epsilon | S, X] = 0 \quad (2)$$

$$\text{P}(S = 1 | X) = \frac{\exp(\alpha_1 + \sum_{k=2}^K \alpha_k X_k)}{1 + \exp(\alpha_1 + \sum_{k=2}^K \alpha_k X_k)}, \quad (3)$$

where  $h(X)$  is an unknown real-valued function of  $X$ , and  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_K)$  is an unknown vector of nuisance parameters. The parameter  $\beta$  represents the average causal effect of an exposure or treatment on the outcome when a certain condition is satisfied. However, it has nothing to do with the estimation of  $\beta$ , so we omit it here (see, Robins *et al.*, 1992).

Next, we let  $\{(Y_i, S_i, X_i)\}_{i=1}^n$  be a random sample, that is, a set of independent and identically distributed random vectors under the above model. Robins *et al.* (1992) also proposed an estimating equation for  $\beta$  as follows,

$$\sum_{i=1}^n U(Y_i, S_i, X_i, \beta, \hat{\alpha}) = 0, \quad (4)$$

where  $\hat{\alpha}$  is the maximum likelihood estimator of  $\alpha$  from the logistic regression model (3) and

$$U(y, s, x, \beta, \alpha) = \{s - r(x; \alpha)\}(y - \beta s), \quad r(x; \alpha) = \frac{\exp(\alpha_1 + \sum_{k=2}^K \alpha_k x_k)}{1 + \exp(\alpha_1 + \sum_{k=2}^K \alpha_k x_k)}.$$

When the model is correct, the estimator  $\hat{\beta}$  of  $\beta$ , which is the solution of the estimating equation (4), is consistent and asymptotically normal under some regularity conditions. In addition its asymptotic variance is calculated as

$$\text{Var}_A(\hat{\beta}) = \text{Var}_A(\tilde{\beta}) - (Q^{-1}P)J^{-1}(Q^{-1}P)^T, \quad (5)$$

where  $\tilde{\beta}$  is the estimator of  $\beta$  with the true value  $\alpha_0$  of  $\alpha$  treated as known, which is the solution of (4) when one replaces  $\hat{\alpha}$  with  $\alpha_0$  and

$$P = E \left[ \frac{\partial U}{\partial \alpha}(Y, S, X, \beta, \alpha) \right], \quad Q = E \left[ \frac{\partial U}{\partial \beta}(Y, S, X, \beta, \alpha) \right]$$

$$J = E[M(S, X, \alpha)M(S, X, \alpha)^T], \quad M(s, x, \alpha) = \frac{\partial}{\partial \alpha} \log[r(x; \alpha)^s \{1 - r(x; \alpha)\}^{1-s}].$$

For a matrix  $A$ ,  $A^T$  denotes the transpose of  $A$ . Then, we find that the following inequality holds,

$$\text{Var}_A(\hat{\beta}) \leq \text{Var}_A(\tilde{\beta}), \quad (6)$$

since  $J$  is an positive definite matrix in equation (5). The equality holds if, and only if  $P = 0$ , that is,  $E[h(X) \frac{\partial r}{\partial \alpha}(X; \alpha)] = 0$ . One might feel that this is strange. Inequality (6) implies that a more precise estimate of  $\beta$  may be generated by estimating the nuisance parameter  $\alpha$  than by using the true value of  $\alpha$  even if the latter were known. This phenomenon was pointed out by Robins *et al.* (1992). They emphasized that this result depends on the fact that  $\hat{\alpha}$  is an efficient estimator of  $\alpha$ . In the following sections, we examine the structure of the inverse phenomenon of asymptotic variances using the orthogonal decomposition of estimating functions. It will be also made clearer what role the fact that  $\hat{\alpha}$  is an efficient estimator of  $\alpha$  plays in the inverse phenomenon.

### 3 The orthogonal decomposition of estimating functions

In this section we describe the orthogonal decomposition of estimating functions for semiparametric models, which is the key notion to understand the structure of the phenomenon mentioned above from our point of view.

Let  $\mathcal{M} = \{p(x; \theta, k)\}$  be a semiparametric statistical model, that is, a family of probability density functions with respect to a common dominating measure  $\mu(dx)$ , whose

element is specified by a finite-dimensional parameter  $\theta = (\theta_1, \dots, \theta_m)^\top$  and an infinite-dimensional parameter  $k$ , typically lying in a space of functions. Here,  $\theta$  contains a parameter of interest and  $k$  is a nuisance parameter. Let  $u(x, \theta) = (u_1(x, \theta), \dots, u_m(x, \theta))^\top$  be a vector-valued smooth function of  $\theta$ , not depending on  $k$ , and of the same dimension as  $\theta$ . This function is called an estimating function for  $\theta$  when it satisfies the following conditions (Godambe, 1991, p.13),

$$\mathbb{E}_{\theta,k}[u(x, \theta)] = 0, \quad \mathbb{E}_{\theta,k}[\|u(x, \theta)\|^2] < \infty, \quad (7)$$

$$\det \mathbb{E}_{\theta,k} \left[ \frac{\partial u}{\partial \theta}(x, \theta) \right] \neq 0 \quad (8)$$

for all  $\theta$  and  $k$ , where  $\mathbb{E}_{\theta,k}$  denotes the expectation with respect to the distribution  $p(x; \theta, k)$ ,  $\det$  denotes the determinant of a matrix, and  $\|\cdot\|$  is the squared norm of vectors. Moreover, we assume that  $\int u(x, \theta)p(x; \theta, k)\mu(dx)$  is differentiable with respect to  $\theta$  and that differentiation and integration are interchangeable. When an estimating function  $u(x, \theta)$  exists, we have an estimator  $\hat{\theta}$  of  $\theta$  as the solution of the following estimating equation:

$$\sum_{i=1}^n u(x_i, \theta) = 0, \quad (9)$$

where  $x_1, \dots, x_n$  are  $n$  independent and identically distributed observations. The estimator  $\hat{\theta}$  is often called an M-estimator. Under some regularity conditions, it is consistent and asymptotically normally distributed with the asymptotic covariance matrix,

$$\text{Var}_A(\hat{\theta}) = W^{-1}VW^{-\top}, \quad (10)$$

where  $V = \mathbb{E}_{\theta,k}[u(x, \theta)u(x, \theta)^\top]$  and  $W = \mathbb{E}_{\theta,k}[\frac{\partial u}{\partial \theta}(x, \theta)]$ .

Now, under the above setting we describe the orthogonal decomposition of estimating functions. Let us consider the set of random variables defined by

$$\mathcal{H}_{\theta,k} = \{a(x) | \mathbb{E}_{\theta,k}[a(x)] = 0, \mathbb{E}_{\theta,k}[a(x)^2] < \infty\}. \quad (11)$$

This is a Hilbert space with the inner product  $\langle a(x), b(x) \rangle_{\theta,k} = \mathbb{E}_{\theta,k}[a(x)b(x)]$  for any two random variables  $a(x), b(x) \in \mathcal{H}_{\theta,k}$ . Then, condition (7) for estimating functions can be represented as

$$u_i(x, \theta) \in \mathcal{H}_\theta \text{ for all } i \text{ and } \theta, \quad (12)$$

where  $\mathcal{H}_\theta$  denotes the intersection of  $\mathcal{H}_{\theta,k'}$  over all  $k'$ . We assume that all components of the score function  $s(x, \theta, k)$  for  $\theta$  belong to  $\mathcal{H}_{\theta,k}$  and let  $s^I(x, \theta, k)$  be the vector comprised by the orthogonal projections of all components of  $s(x, \theta, k)$  onto  $\overline{\mathcal{H}}_\theta$ , which is the closure of  $\mathcal{H}_\theta$  with respect to the topology of  $\mathcal{H}_{\theta,k}$ . Then, the space  $\overline{\mathcal{H}}_\theta$  can be decomposed as

$$\overline{\mathcal{H}}_\theta = \mathcal{F}_{\theta,k}^I \oplus \mathcal{F}_{\theta,k}^A, \quad (13)$$

where  $\mathcal{F}_{\theta,k}^I$  denotes the linear space spanned by all components of  $s^I(x, \theta, k)$  and  $\mathcal{F}_{\theta,k}^A$  denotes the orthogonal complement of  $\mathcal{F}_{\theta,k}^I$  in  $\overline{\mathcal{H}}_\theta$ . We call the vector-valued function  $s^I(x, \theta, k)$  the information score function for  $\theta$  and assume that all components of  $s^I(x, \theta, k)$  are linearly independent. According to (12) and (13), any estimating function  $u(x, \theta)$  is represented by the following form for all  $k$ :

$$u(x, \theta) = T(\theta, k)s^I(x, \theta, k) + a(x, \theta, k), \quad (14)$$

where  $T(\theta, k)$  is an  $m \times m$  matrix and  $a(x, \theta, k)$  is a vector-valued function whose components belong to  $\mathcal{F}_{\theta,k}^A$ . Moreover, by condition (8) the orthogonal projections of all components of  $u(x, \theta)$  onto  $\mathcal{F}_{\theta,k}^I$  are linearly independent, and therefore  $T(\theta, k)$  is non-singular. Representation (14) is what we call the orthogonal decomposition of estimating functions for semiparametric models in this paper. This kind of decomposition has often been treated in the literature on estimating functions. In particular, Amari and Kawanabe (1997) consider the characterization of the orthogonal decomposition (13) from an information geometrical point of view. The terminology of an information score function is due to them. In the decomposition (14), the parameter  $k$  is fixed by an arbitrary possible value, and for an estimating function  $u(x, \theta)$ , its different expressions are obtained by values of  $k$ . When in particular we set  $k = k_0$ , which is the value of  $k$  corresponding to the unknown underlying distribution in  $\mathcal{M}$  that generates the data, the asymptotic covariance matrix of the estimator  $\hat{\theta}$  as the solution of the estimating equation (9) is calculated as follows (Amari and Kawanabe, 1997),

$$\text{Var}_A(\hat{\theta}) = (G^I)^{-1} + (TG^I)^{-1}G^A(TG^I)^{-T}, \quad (15)$$

where  $G^I = E_{\theta_0, k_0}[s^I(x, \theta_0, k_0)s^I(x, \theta_0, k_0)^T]$ ,  $G^A = E_{\theta_0, k_0}[a(x, \theta_0, k_0)a(x, \theta_0, k_0)^T]$ ,  $T = T(\theta_0, k_0)$ , and  $\theta_0$  denotes the true value of  $\theta$ . In equation (15),  $G^A$  is a positive semi-definite matrix. Hence,  $\text{Var}_A(\hat{\theta}) \geq (G^I)^{-1}$  and the equality holds only when  $G^A = 0$ . This implies that if  $s^I(x, \theta, k_0)$  satisfies the conditions to be an estimating function, it is

an optimal estimating function in the sense that the asymptotic covariance matrix of the estimator is minimum among all estimating functions. However, it should be noted that generally,  $s^I(x, \theta, k_0)$  cannot be used since it usually depends on the unknown true value  $k_0$  of  $k$ . According to the above discussion, the orthogonal decomposition of estimating functions represents how an estimating function fails to reach the optimal state. Then, we call the first and second terms of the right side in the orthogonal decomposition (14) the optimal and non-optimal parts of  $u(x, \theta)$ , respectively.

## 4 The inverse phenomenon of asymptotic variances

In this section we examine the structure of the inverse phenomenon of asymptotic variances. The model in the example given in Section 2 is a semiparametric model with both finite and infinite-dimensional nuisance parameters. In fact, under (2) and (3), the joint probability density function of the observed variables  $Y, S$  and  $X$  can be written as follows:

$$p_{Y SX}(y, s, x; \beta, \alpha, h, g, f) = g(y - \beta s - h(x)|s, x)p_{S|X}(s|x; \alpha)f(x), \quad (16)$$

where  $g(\epsilon|s, x)$  denotes the conditional density function of the error  $\epsilon$  given  $S = s$  and  $X = x$ ,  $p_{S|X}(s|x; \alpha)$  denotes the conditional probability function of  $S$  given  $X = x$  and is written as  $\{r(x; \alpha)\}^s\{1 - r(x; \alpha)\}^{1-s}$  from (3), and  $f(x)$  denotes the marginal density function of  $X$ . While the parameter  $\beta$  is of interest,  $\alpha$  is a finite-dimensional nuisance parameter. The functions  $h, g$  and  $f$  play a role of infinite-dimensional nuisance parameters. The inverse phenomenon of asymptotic variances is the phenomenon in which the asymptotic variance of the estimator of  $\beta$  with unknown  $\alpha$  is less than that with known  $\alpha$ . As is shown in Section 2, this phenomenon can occur under the above model when  $\alpha$  is estimated by the maximum likelihood method and  $\beta$  is estimated by the estimating function  $U(y, s, x, \beta, \alpha)$ , which depends on  $\alpha$ . This implies that generally, the inverse phenomenon of asymptotic variances occurs under some special conditions.

Let  $\mathcal{M} = \{p(x; \theta, k)\}$  be a semiparametric model with a finite-dimensional parameter  $\theta = (\beta^T, \alpha^T)^T$  and an infinite-dimensional nuisance parameter  $k$ . Here,  $\beta$  and  $\alpha$  are parameters of interest and of nuisance respectively. Let  $u(x, \theta) = (u_\beta(x, \theta)^T, u_\alpha(x, \theta)^T)^T$  be an estimating function for  $\theta$ . The two components  $u_\beta(x, \theta)$  and  $u_\alpha(x, \theta)$  are marginal estimating functions for  $\beta$  and  $\alpha$ , that is, estimating functions for  $\beta$  and  $\alpha$  when  $\alpha$  and  $\beta$  are fixed, respectively. The following theorem gives one sufficient condition for the

inverse phenomenon to occur.

**THEOREM 1.** *Assume that the semiparametric model  $\mathcal{M} = \{p(x; \theta, k)\}$  and the estimating function  $u(x, \theta) = (u_\beta(x, \theta)^\top, u_\alpha(x, \theta)^\top)^\top$  satisfy the conditions,*

$$\mathbb{E}_{\theta, k}[s_\beta(x, \theta, k)s_\alpha(x, \theta, k)^\top] = 0 \quad (\forall \theta, \forall k), \quad (17)$$

$$s_\alpha(x, \theta, k) \text{ does not depend on } k, \quad (18)$$

$$u_\alpha(x, \theta) = s_\alpha(x, \theta), \quad (19)$$

where  $s_\beta(x, \theta, k)$  and  $s_\alpha(x, \theta, k) = s_\alpha(x, \theta)$  are the score functions for  $\beta$  and  $\alpha$ , respectively. Then, the following inequality holds:

$$\text{Var}_A(\hat{\beta}) \leq \text{Var}_A(\tilde{\beta}), \quad (20)$$

where  $\hat{\beta}$  is the estimator of  $\beta$  in the joint estimation of  $\beta$  and  $\alpha$  by  $u(x, \theta)$  while  $\tilde{\beta}$  is that in the single estimation of  $\beta$  by  $u_\beta(x, \theta)$  with known  $\alpha$ . The equality holds if, and only if  $\mathbb{E}_{\theta, k}[u_\beta(x, \theta)s_\alpha(x, \theta)^\top] = 0$ .

The example in Section 2 fits this theorem as a special case, in which the score function for  $\alpha$  depends neither on the infinite-dimensional parameters  $h, g$  and  $f$  nor on the parameter of interest  $\beta$ . The above theorem can be proved by direct calculation of the asymptotic covariance matrices of the estimators  $\hat{\beta}$  and  $\tilde{\beta}$ . However, the reason of the inverse phenomenon of asymptotic variances is not sufficiently explained by direct calculation. Then, we consider the orthogonal decomposition of estimating functions described in Section 3. It leads us to clear understanding of the structure of the inverse phenomenon. The following is a proof of the above theorem using the orthogonal decomposition.

Firstly, we note that under conditions (17) and (18) the following equations hold:

$$\mathbb{E}_{\theta, k}[s_\beta^\text{I}(x, \theta, k)s_\alpha^\text{I}(x, \theta, k)^\top] = 0 \quad (\forall \theta, \forall k), \quad (21)$$

$$s_\alpha^\text{I}(x, \theta, k) = s_\alpha(x, \theta), \quad (22)$$

where  $s_\beta^\text{I}(x, \theta, k)$  and  $s_\alpha^\text{I}(x, \theta, k)$  are the information score functions for  $\beta$  and  $\alpha$ , respectively. This is because the information score functions for  $\beta$  and  $\alpha$  are respectively the orthogonal projections of the score functions for  $\beta$  and  $\alpha$  onto the space  $\overline{\mathcal{H}}_\theta$  defined in Section 3 and because the score function for  $\alpha$  belongs to  $\overline{\mathcal{H}}_\theta$  due to (18). By equations

(21) and (22), the orthogonal decomposition of the marginal estimating function  $u_\beta(x, \theta)$  for  $\beta$  can be represented as follows:

$$u_\beta(x, \theta) = T_\beta(\theta, k_0)s_\beta^I(x, \theta, k_0) + T_\alpha(\theta, k_0)s_\alpha(x, \theta) + a(x, \theta, k_0), \quad (23)$$

where  $k_0$  denotes the true value of  $k$ . The first term in the right side of (23) is the optimal part of  $u_\beta(x, \theta)$  while the second and third terms compose the non-optimal part. Here,  $s_\alpha(x, \theta)$  and  $a(x, \theta, k_0)$  are orthogonal. Now, we consider the estimation of  $\theta = (\beta^\top, \alpha^\top)^\top$  by the estimating function  $u(x, \theta) = (u_\beta(x, \theta)^\top, s_\alpha(x, \theta)^\top)^\top$ . It should be noted that the term of the score function for  $\alpha$  is redundant in the decomposition (23) due to the existence of  $s_\alpha(x, \theta)$  as a marginal estimating function for  $\alpha$ . In other words,  $u(x, \theta)$  is equivalent to the estimating function  $u^*(x, \theta) = (u_\beta^*(x, \theta)^\top, s_\alpha(x, \theta)^\top)^\top$ , where

$$u_\beta^*(x, \theta) = T_\beta(\theta, k_0)s_\beta^I(x, \theta, k_0) + a(x, \theta, k_0), \quad (24)$$

in the sense that  $u(x, \theta)$  and  $u^*(x, \theta)$  give the same estimator. Here,  $u_\beta^*(x, \theta)$  is a marginal estimating function for  $\beta$  that usually depends on unknown  $k_0$ , and cannot be used in practice. It is hypothetical, but can be theoretically considered just like an information score function evaluated by unknown  $k_0$ . According to the orthogonality of  $s_\alpha(x, \theta)$  and  $a(x, \theta, k_0)$ , and equations (21) and (22),  $u_\beta^*(x, \theta)$  is orthogonal to  $s_\alpha(x, \theta)$ . Then, by the following theorem, we find that the asymptotic covariance matrix of the estimator of  $\beta$  in the joint estimation of  $\beta$  and  $\alpha$  by  $u^*(x, \theta)$  coincides with that in the single estimation of  $\beta$  by  $u_\beta^*(x, \theta)$  with known  $\alpha$ .

**THEOREM 2 (Insensitivity Theorem).** *Let  $\mathcal{M} = \{p(x; \theta, k)\}$  be an arbitrary semiparametric model with a finite-dimensional parameter  $\theta = (\beta^\top, \alpha^\top)^\top$  and an infinite-dimensional nuisance parameter  $k$ . Let  $w(x, \theta)$  be an arbitrary estimating function for  $\theta$  composed by marginal estimating functions  $w_\beta(x, \theta)$  for  $\beta$  and  $w_\alpha(x, \theta)$  for  $\alpha$ . If  $w_\beta(x, \theta)$  is orthogonal to the score function  $s_\alpha(x, \theta, k)$  for  $\alpha$ , that is,*

$$E_{\theta, k}[w_\beta(x, \theta)s_\alpha(x, \theta, k)^\top] = 0 \quad (\forall \theta, \forall k),$$

*then the asymptotic covariance matrix of the estimator of  $\beta$  in the joint estimation of  $\beta$  and  $\alpha$  by  $w(x, \theta)$  coincides with that in the single estimation of  $\beta$  by  $w_\beta(x, \theta)$  with known  $\alpha$ .*

This theorem was shown by Knudsen (1999) in the case of parametric models, but it also holds in the case of semiparametric models. From the above discussion the asymptotic covariance matrix of  $\hat{\beta}$ , which is the estimator of  $\beta$  in the joint estimation of  $\beta$  and  $\alpha$  by  $u(x, \theta)$ , coincides with that in the single estimation of  $\beta$  by  $u_\beta^*(x, \theta)$  with known  $\alpha$ . Hence, by equation (15) and the orthogonal decomposition (24), the asymptotic covariance matrix of  $\hat{\beta}$  is represented as

$$\text{Var}_A(\hat{\beta}) = (G_\beta^I)^{-1} + (T_\beta G_\beta^I)^{-1} G^A (T_\beta G_\beta^I)^{-T}, \quad (25)$$

where  $G_\beta^I = E_{\theta_0, k_0}[s_\beta^I(x, \theta_0, k_0)s_\beta^I(x, \theta_0, k_0)^T]$ ,  $G^A = E_{\theta_0, k_0}[a(x, \theta_0, k_0)a(x, \theta_0, k_0)^T]$ ,  $T_\beta = T_\beta(\theta_0, k_0)$ , and  $\theta_0$  is the true value of  $\theta$ . On the other hand, according to the decomposition (23), the asymptotic covariance matrix of the estimator  $\tilde{\beta}$  with known  $\alpha$  is represented as

$$\text{Var}_A(\tilde{\beta}) = (G_\beta^I)^{-1} + (T_\beta G_\beta^I)^{-1} (T_\alpha G_\alpha T_\alpha^T + G^A) (T_\beta G_\beta^I)^{-T}, \quad (26)$$

where  $G_\alpha = E_{\theta_0, k_0}[s_\alpha(x, \theta_0)s_\alpha(x, \theta_0)^T]$  and  $T_\alpha = T_\alpha(\theta_0, k_0)$ . By comparing (25) and (26), we find that the following inequality holds:

$$\text{Var}_A(\hat{\beta}) \leq \text{Var}_A(\tilde{\beta}). \quad (27)$$

The equality holds only when  $T_\alpha = 0$  because of the positive-definiteness of the matrix  $G_\alpha$ . This is equivalent to the condition  $E_{\theta, k}[u_\beta(x, \theta)s_\alpha(x, \theta)^T] = 0$ . Thus, Theorem 1 has been proved.

In the above discussion, the key point is to consider the orthogonal decomposition (23) for  $u_\beta(x, \theta)$ . Because of the orthogonality of the information score function  $s_\beta^I(x, \theta, k)$  for  $\beta$  and the score function  $s_\alpha(x, \theta)$  for  $\alpha$ , the non-optimal part of  $u_\beta(x, \theta)$  has a component of  $s_\alpha(x, \theta)$  unless  $u_\beta(x, \theta)$  and  $s_\alpha(x, \theta)$  are orthogonal. Therefore, in the single estimation of  $\beta$  by  $u_\beta(x, \theta)$  with known  $\alpha$ , there exists loss of asymptotic efficiency which comes from the component of  $s_\alpha(x, \theta)$ . However, by estimating  $\beta$  and  $\alpha$  simultaneously with  $u(x, \theta) = (u_\beta(x, \theta)^T, s_\alpha(x, \theta)^T)^T$ , the component of  $s_\alpha(x, \theta)$  vanishes and the asymptotic efficiency is improved. It should be noted that Insensitivity Theorem plays an important role here, that is, it converts the asymptotic efficiency in the joint estimation into that in the single estimation.

It should be also noted that if the marginal estimating function for  $\beta$  is optimal, the inverse phenomenon does not occur. This holds true whether or not the model  $\mathcal{M}$

and the estimating function  $u(x, \theta)$  satisfy conditions (17), (18) and (19), because the lower bound of the asymptotic covariance matrix of  $\hat{\beta}$  is not less than that of  $\tilde{\beta}$ . The inverse phenomenon of asymptotic variances indicates that the asymptotic efficiency of the estimator can be improved by estimating nuisance parameters under some special conditions in the situation where the optimal estimating function cannot be used.

## 5 Parametric case

In the preceding sections we considered the semiparametric case, but the inverse phenomenon of asymptotic variances can also occur in the parametric case. The structure is essentially the same as in the semiparametric case. The discussion in Section 4 is also applicable to the parametric case if a small modification is made; that is, to remove the infinite-dimensional parameter  $k$  and to replace information scores with ordinary scores.

For parametric models, the inverse phenomenon of asymptotic variances occurs in the following case: Let  $\mathcal{M} = \{p(x; \theta)\}$  be a parametric model with a vector of parameters  $\theta$ , which is composed by two vectors of parameters,  $\beta$  of interest and  $\alpha$  of nuisance. We assume that  $\beta$  and  $\alpha$  are orthogonal, that is,

$$E_{\theta}[s_{\beta}(x, \theta)s_{\alpha}(x, \theta)^T] = 0 \quad (\forall \theta), \quad (28)$$

where  $s_{\beta}(x, \theta)$  and  $s_{\alpha}(x, \theta)$  are the score functions for  $\beta$  and  $\alpha$ , respectively. In this situation we consider the estimation of  $\theta = (\beta^T, \alpha^T)^T$  by an estimating function  $u(x, \theta) = (u_{\beta}(x, \theta)^T, s_{\alpha}(x, \theta)^T)^T$ , where  $u_{\beta}(x, \theta)$  is an arbitrary marginal estimating function for  $\beta$ . Then, the following inequality for asymptotic covariance matrices holds:

$$\text{Var}_A(\hat{\beta}) \leq \text{Var}_A(\tilde{\beta}), \quad (29)$$

where  $\hat{\beta}$  is the estimator of  $\beta$  in the joint estimation of  $\beta$  and  $\alpha$  by  $u(x, \theta)$  and  $\tilde{\beta}$  is that given by  $u_{\beta}(x, \theta)$  when the true value of  $\alpha$  is known. The equality holds if, and only if the marginal estimating function  $u_{\beta}(x, \theta)$  and the score function  $s_{\alpha}(x, \theta)$  are orthogonal. When condition (28) holds, it is well known that the equality holds in (29) if  $u_{\beta}(x, \theta)$  coincides with  $s_{\beta}(x, \theta)$ . However, if not so, the asymptotic covariance matrix of the estimator of  $\beta$  in the case of estimating  $\alpha$  can be less than in the case of using the true value of  $\alpha$ .

Now, we give one example of the inverse phenomenon of asymptotic variances in the parametric case. Let  $Y$  and  $X$  be a response variable of interest and a vector of some

covariates, respectively. We assume that they are both random variables. A generalized linear model for the conditional distribution of  $Y$  given  $X = x$  is written as follows:

$$p(y|x; \beta, \phi) = \exp \left\{ \frac{y\theta - b(\theta)}{\phi} + c(y, \phi) \right\}, \quad (30)$$

$$g(\mu) = x^T \beta, \quad (31)$$

where  $\theta, \mu$  and  $\phi$  denote a natural, a mean and a dispersion parameter, respectively, and  $g$  is a link function. The vector of regression parameters  $\beta$  is usually the object of inference. Here, however we assume that the dispersion parameter  $\phi$  is of our interest and treat  $\beta$  as a vector of nuisance parameters. When we estimate  $\beta$  based on an observed random sample, the maximum likelihood method is usually used. However, estimation of the dispersion parameter  $\phi$  is not always the same. For example, the moment method is often used based on some reasons (see, for instance, McCullagh and Nelder, 1989, p.295). When the maximum likelihood method is applied for  $\beta$  and the moment method for  $\phi$ , the corresponding estimating function is as follows:

$$\begin{aligned} u(y, x, \phi, \beta) &= (u_\phi(y, x, \phi, \beta), s_\beta(y, x, \phi, \beta)^T)^T \\ &= \left( \phi - \frac{(y - \mu)^2}{V(\mu)}, \frac{y - \mu}{\phi V(\mu) g'(\mu)} x^T \right)^T, \end{aligned} \quad (32)$$

where  $V(\mu)$  denotes a variance function. The function  $s_\beta(y, x, \phi, \beta)$  is the score function for  $\beta$  and in addition, the two parameters  $\phi$  and  $\beta$  are orthogonal. Hence, this is a situation in which the inverse phenomenon of asymptotic variances can occur; that is, we observe

$$\text{Var}_A(\hat{\phi}) \leq \text{Var}_A(\tilde{\phi}), \quad (33)$$

where  $\hat{\phi}$  is the estimator of  $\phi$  in the case of estimating  $\beta$  and  $\tilde{\phi}$  is that in the case of using the true value of  $\beta$ . The condition for the equality to hold is as follows:

$$\text{E} \left[ \frac{V'(\mu)}{V(\mu)g'(\mu)} X^T \right] = 0. \quad (34)$$

If the model (30) is a normal distribution, this condition is satisfied because  $V(\mu) = 1$ . However, for instance, in the case of a gamma distribution, this condition is not always satisfied and the inverse phenomenon of asymptotic variances can occur. Since the dispersion  $\phi$  is usually a nuisance parameter, the efficiency of the estimator of  $\phi$  might be of little concern in practice. However, the fact that inequality (33) holds with respect to the estimation of  $\phi$  is of interest.

## 6 Concluding remarks

In this paper we have examined the structure of the inverse phenomenon of asymptotic variances using the orthogonal decomposition of estimating functions. If an optimal estimating function can be used as a marginal estimating function for a parameter of interest  $\beta$ , the asymptotic variance of the estimator with unknown nuisance parameter  $\alpha$  cannot be less than that with known  $\alpha$ . This is reasonable and compatible with our intuition. However, it is not always true unless the marginal estimating function for  $\beta$  is optimal. In fact, as is discussed in Section 4, when the marginal estimating function for  $\beta$  has a component of the score function for  $\alpha$  in the non-optimal part of its orthogonal decomposition, the asymptotic variance of the estimator of  $\beta$  decreases by estimating  $\alpha$  with the score function for  $\alpha$  rather than by using the true value of  $\alpha$ . The inverse phenomenon of asymptotic variances seems to be strange at least intuitively. However, the discussion on the orthogonal decomposition of estimating functions makes it clear how this inverse phenomenon occurs.

It should be noted that the inverse phenomenon of asymptotic variances can occur in both parametric and semiparametric models in principle. The inverse phenomenon comes from the structure of estimating functions. However, it seems that this phenomenon has fewer opportunities to occur in the parametric case than in the semiparametric case. This is because, in the parametric case, the optimal estimating function can be used under moderate regularity conditions, and other estimation methods are not used unless some special reason exists. On the other hand, in the semiparametric case, since the optimal estimating function generally depends on the unknown true value of infinite-dimensional nuisance parameters, it usually cannot be used even if it is possible to obtain its functional form explicitly. In the example given in Section 2, if we assume that the error  $\epsilon$  and  $(S, X)$  are independent, the information score function for the parameter of interest  $\beta$  can be calculated explicitly and depends on the unknown regression function  $h$  and the unknown marginal density function  $g$  of  $\epsilon$ . The marginal estimating function for  $\beta$  which is used there is obtained from the information score function by substituting the zero function and the normal density function with the mean zero and the variance constant for  $h$  and  $g$ , respectively. Therefore, if  $h$  and  $g$  actually coincide with the above functions, the inverse phenomenon of asymptotic variances never occurs. However, if not, and especially if  $h$  is not a zero function, the inverse phenomenon does occur. Of course, even in the semiparametric case, the results can change if we estimate the infinite-

dimensional nuisance parameters by some nonparametric approach. For the model in Section 2, no estimation of the regression function  $h$  is intended because it is difficult for epidemiological reasons. Instead of estimating  $h$ , model (3) is considered (see, Robins *et al.*, 1992).

The inverse phenomenon of asymptotic variances does not seem to occur in so many situations since the conditions in Theorem 1 are rather restrictive. However, this phenomenon naturally occurs in some situations as well as the example by Robins *et al.* (1992), for instance, in the problems of missing-data (Robins, *et al.*, 1994, Lawless *et al.*, 1999), measurement error (Carroll, Ruppert and Stefanski, 1995) and survey sampling (Rosenbaum, 1987). The inverse phenomenon of asymptotic variances gives great impact to the statistical community since it defies the common sense of statistical inference. We believe that our viewpoint helps us comprehend this phenomenon.

## REFERENCES

- Amari, S. and Kawanabe, M. (1997). Information geometry of estimating functions in semiparametric statistical models, *Bernoulli*, **3**, 29-54.
- Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*, Chapman and Hall, London.
- Fourdrinier, D. and Strawderman, W. E. (1996). A paradox concerning shrinkage estimators: should a known scale parameter be replaced by an estimated value in the shrinkage factor? *J. Multivar. Anal.*, **59**, 109-140.
- Godambe, V. P. (ed.) (1991). *Estimating Functions*, Oxford University Press, New York.
- Knudsen, S. J. (1999). *Estimating Functions and Separate Inference*, Monographs Vol.1., Dept. of Statistics and Demography, University of Southern Denmark.
- Lawless, J. F., Kalbfleisch, J. D. and Wild, C. J. (1999). Semiparametric methods for response-selective and missing data problems in regression, *J. R. Statist. Soc. B*, **61**, 413-438.
- McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, Chapman and Hall, London.

- Robins, J. M., Mark, S. D. and Newey W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders, *Biometrics*, **48**, 479-495.
- Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed, *J. Am. Statist. Ass.*, **89**, 846-866.
- Rosenbaum, P. R. (1987). Model-based direct adjustment, *J. Am. Statist. Ass.*, **82**, 387-394.