The Korean Communications
in Statistics Vol. 5, No. 3, 1998
pp. 927-934

An Approximate Parameter Orthogonality

Kwan Jeh Lee[1)]

Abstract

An approximate parameter orthogonality is defined, which is called an α-approximate orthogonality. The useful consequences of parameter orthogonality mentioned by Cox and Reid(1987) can be shared by an α-approximate orthogonality. If $\alpha \geq 1/2$, the consequences of orthogonality and α-approximate orthogonality are asymptotically equivalent.

1 Introduction

In multiparameter statistical models we often focus on one or a few parameters with the others being treated as nuisance parameters. The subject of inference on interest parameters in the presence of nuisance parameters is at the core of statistical problems with multi-dimensional parameters(Godambe 1976, Godambe and Thompson 1974). Even though there can be no generally optimal method for elimination of the effect of nuisance parameters, there are several approaches used to reduce the effect of nuisance parameters. A procedure used in the case is to replace the nuisance parameters in the likelihood function by their mles and examine the resulting "profile" likelihood as a function of the parameter of interest. This procedure is known to give inconsistent or inefficient estimates of the interest parameters. Another method, introduced by Anderson(1970), is to use the conditional likelihood. The conditional likelihood approach is achieved by conditioning the data on the minimal sufficient statistics for nuisance parameters. The merit of this approach is to focus the inference on a genuine likelihood which depends only on the parameter of interest so that the effects of nuisance parameters can be reduced. However, expect for some important special cases such as regular exponential families, the above approach may not be desirable as in general: the dimension of the minimal sufficient statistics is greater than the number of nuisance parameters. Other than two methods above, there are several approaches such as the modified profile likelihood of Barndorff-Nielsen(1983) and Barndorff-Nielsen and Cox(1994), the conditional profile likelihood of Cox and Reid(1987), the Bayesian approach of Box and Cox(1964) and Pericchi(1981), the orthogonalization of parameters of Cox and Reid(1987) and Amari(1985), etc.

---

1) Assistant Professor, Dept. of Statistics, Dongguk University, Seoul 100-715, Korea. This paper was supported in part by Research Foundation of Dongguk University.

There are a number of advantages, conceptual and mathematical, when parameters are orthogonal. Orthogonality of parameters implies that the Fisher information matrix is diagonal, that is, the corresponding components of the score statistic are uncorrelated. The formal definition and some asymptotic properties of orthogonal parameters will be discussed later.

In this paper we define the α-approximate orthogonality of parameters and see that the approximately orthogonal parameters share some asymptotic properties of orthogonal parameters.

2 Definition and Consequences of Orthogonal Parameters

We consider random variables (r.v.s) $X_1, X_2, \cdots, X_n$ which are independent and identically distributed (*i.i.d.*) according to a distribution with density $f(x|\theta)$, where the parameter θ is possibly a $p\times 1$ vector, and write $X=(X_1, \cdots, X_n)$.

The joint density of $X_1, X_2, \cdots, X_n$ is given by

$$f(x_1, \cdots, x_n|\theta) = \prod_{s=1}^{n} f(x_s|\theta) = L(\theta|x).$$

Each realization $x=(x_1, \cdots, x_n)$ produces a different function $L(\cdot|x)$, which called the likelihood function for θ. The logarithm of the likelihood is denoted by $l(\theta) = lnL(\theta|x)$. The vector of partial derivative functions of $l(\theta)$ with respect to the components of θ is

$$\dot{l}(\theta) = (\frac{\partial l(\theta)}{\partial \theta_1}, \frac{\partial l(\theta)}{\partial \theta_2}, \cdots, \frac{\partial l(\theta)}{\partial \theta_p})'.$$

The likelihood equation then is a vector equation representing a system of nonlinear equations in the p unknown θ_k. We shall use the notation $\ddot{l}(\theta)$ for the matrix of second partial derivatives with (s, t) position given by

$$\ddot{l}(\theta) = (\frac{\partial^2 l(\theta)}{\partial \theta_s \partial \theta_t})_{pxp}.$$

The Fisher information, or information matrix, for θ is defined as

$$I(\theta) = [-E(\ddot{l}(\theta))] = \frac{1}{n}(i_{st})_{pxp}. \tag{1}$$

Note here that i refers to information per observation, which will be assumed $O(1)$ as $n \to \infty$. We define the parameters are orthogonal if the Fisher information matrix is diagonal, that is, $i_{st} = 0$ for $s \neq t$ at (1) above. There is the construction of orthogonality in Amari(1985) and Cox and Reid(1987). If off-diagonal elements of the Fisher information matrix is zero for all parameters, it is sometimes called global orthogonality. Meanwhile, if $i_{st} = 0$ for $s \neq t$ holds for only s and t components of θ, then θ_s and θ_t are said to be locally orthogonal. It is also mentioned in Amari(1985) and Cox and Reid(1987) that the existence of an orthogonal paramerization is guaranteed when a parameter of interest is a scalar, whereas global orthogonality is possible only in special cases.

There are roughly four reasons for wishing to consider an orthogonal parameterization mentioned by Cox and Reid(1987): computation, approximation, interpretation, and elimination of nuisance parameters. For simplicity, suppose $\theta = (\psi, \lambda)$ has just two components. Then the consequences of orthogonality of ψ and λ are

(i) the mles of $\hat{\psi}$ and $\hat{\lambda}$ are asymptotically independent;
(ii) the asymptotic standard error for estimating ψ is the same whether λ is known or not;
(iii) $\hat{\psi}_\lambda - \hat{\psi} = O_p(n^{-1})$ provided $\hat{\lambda} - \lambda = O_p(n^{-\frac{1}{2}})$, where $\hat{\psi}_\lambda$ is the mle of ψ when λ is given;
(iv) the computation in numerical determination of the mles $(\hat{\psi}, \hat{\lambda})$ may be easier.

3 Main Results

The useful consequences of parameter orthogonality mentioned in the previous section can be achieved in some cases where the off-diagonal components of the Fisher information matrix are not zero exactly. We define some approximate orthogonality of parameters.

Definition 1 *Let i_{st} be (s, t)-component of Information matrix $I(\theta)$ for $\theta = (\theta_1, \cdots, \theta_p)$. If for $\alpha \geq 0$ and $s \neq t$*

$$i_{st} = O(n^{-\alpha}),$$

then θ_s and θ_t are said to be α-approximately orthogonal. If all the off-diagonal components of the Fisher information matrix are $O(n^{-\alpha})$, then θ is said to be globally α-approximately orthogonal.

For simplicity, suppose $\theta=(\psi, \lambda)$ has just two components. We also assume the regularity conditions required for maximum likelihood theory in Cramer(1946, p501) or Serfling(1980, p144). We have the following results.

Lemma 1 *Let (X, Y) be a bivariate random vector with joint pdf or pmf $f(x, y)$. Then $X$and Y are independent random variables if and only if there exist functions $g(x)$ and $h(y)$ such that, for every $x\in R$ and $y\in R$,*

$$f(x, y)=g(x)\, h(y).$$

The proof of the Lemma is in details in Casella and Berger(1990, p142).

Theorem 1 *If for $\theta=(\psi, \lambda)$ and $\alpha\geq 0$, parameters ψ and λ are α-approximately orthogonal, we have the following results.*

(i) the mles of $\hat{\psi}$ and $\hat{\lambda}$ are asymptotically independent;

(ii) the difference between the asymptotic standard error for estimating ψ with unknown λ and that with known λ is $O^{+}(n^{-\alpha})$, where $O^{+}(n^{-\alpha})$ denotes nonnegative $O(n^{-\alpha})$;

(iii) $\hat{\psi}_{\lambda}-\hat{\psi}=O_{p}(n^{-\min\{\alpha+1/2,\,1\}})$ provided $\hat{\lambda}-\lambda=O_{p}(n^{-\frac{1}{2}})$, where $\hat{\psi}_{\lambda}$ is the mle of ψ when λ is given.

Proof. Suppose two components of θ, ψ and λ, are α-approximately orthogonal, i.e.,

$$i_{\psi\lambda}=O(n^{-\alpha}).$$

Since the asymptotic distribution of $\hat{\psi}$ and $\hat{\lambda}$ is a bivariate normal and the mles are asymptotically efficient, $\sqrt{(n)}(\hat{\theta}-\theta)$ is asymptotically bivariate normal with (vector) mean zero and covariance matrix $\Sigma=I^{-1}(\theta)$, where we can write

$$\Sigma=\begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\psi\lambda} & i_{\lambda\lambda}\end{pmatrix}^{-1}.$$

Thus

$$\Sigma=\frac{1}{i_{\psi\psi}i_{\lambda\lambda}-i_{\psi\lambda}i_{\psi\lambda}}\begin{pmatrix} i_{\lambda\lambda} & -i_{\psi\lambda} \\ -i_{\psi\lambda} & i_{\psi\psi}\end{pmatrix}.$$

Note here that

$$\frac{1}{i_{\phi\phi}i_{\lambda\lambda}-i_{\phi\lambda}i_{\phi\lambda}}=\frac{1}{i_{\phi\phi}i_{\lambda\lambda}(1-i_{\phi\lambda}i_{\phi\lambda}/i_{\phi\phi}i_{\lambda\lambda})}$$

$$=\frac{1}{i_{\phi\phi}i_{\lambda\lambda}}\left(1+\frac{i_{\phi\lambda}i_{\phi\lambda}}{i_{\phi\phi}i_{\lambda\lambda}}+\left(\frac{i_{\phi\lambda}i_{\phi\lambda}}{i_{\phi\phi}i_{\lambda\lambda}}\right)^2+\cdots\right) \tag{2}$$

Since both $i_{\phi\phi}$ and $i_{\lambda\lambda}$ are assumed $O(1)$, and off-diagonal component $i_{\phi\lambda}$ is $O(n^{-\alpha})$, we can rewrite (2) as

$$\frac{1}{i_{\phi\phi}i_{\lambda\lambda}-i_{\phi\lambda}i_{\phi\lambda}}=\frac{1}{i_{\phi\phi}i_{\lambda\lambda}}(1+O^+(n^{-2\alpha})) \tag{3}$$

Using (3), we can write

$$\Sigma=\begin{pmatrix} i_{\phi\phi}^{-1}(1+O^+(n^{-2\alpha})) & O(n^{-\alpha}) \\ O(n^{-\alpha}) & i_{\lambda\lambda}^{-1}(1+O^+(n^{-2\alpha})) \end{pmatrix}, \tag{4}$$

which shows *(ii)*.

The exponent of bivariate normal density of $(\hat{\psi},\hat{\lambda})$ can be written

$$-\frac{n}{2}(\hat{\psi}-\psi,\hat{\lambda}-\lambda)\Sigma^{-1}(\hat{\psi}-\psi,\hat{\lambda}-\lambda)'$$

$$=-\frac{n}{2}(\hat{\psi}-\psi,\hat{\lambda}-\lambda)I(\theta)(\hat{\psi}-\psi,\hat{\lambda}-\lambda)'$$

$$=-\frac{n}{2}((\hat{\psi}-\psi)^2i_{\phi\phi}+(\hat{\lambda}-\lambda)^2i_{\lambda\lambda}+2(\hat{\psi}-\psi)(\hat{\lambda}-\lambda)i_{\phi\lambda}). \tag{5}$$

The last term of (5) is $O_p(n^{-\alpha})$ from the fact that $\hat{\psi}-\psi=O_p(n^{-\frac{1}{2}})$ and $\hat{\lambda}-\lambda=O_p(n^{-\frac{1}{2}})$. Note here that for $\alpha\geq 0$

$$\exp(O_p(n^{-\alpha}))=1+O_p(n^{-\alpha})+(O_p(n^{-\alpha}))^2/2+\cdots.$$

Thus the exponent of bivariate normal density can be factorized up to the order $O_p(n^{-\alpha})$. That is, we can write the exponent of bivariate normal density up to the order $O_p(n^{-\alpha})$,

$$-\frac{n}{2}((\hat{\psi}-\psi)^2i_{\psi\psi}+(\hat{\lambda}-\lambda)^2i_{\lambda\lambda})$$

By the Lemma above $(\hat{\psi}, \hat{\lambda})$ are asymptotically independent up to the order $O_p(n^{-\alpha})$. To show *(iii)*, we write the log-likelihood function near the mles $(\hat{\psi}, \hat{\lambda})$ as

$$l(\hat{\psi}, \hat{\lambda}) + \frac{n}{2}\{-\hat{j}_{\psi\psi}(\psi - \hat{\psi})^2 - 2\hat{j}_{\psi\lambda}(\psi - \hat{\psi})(\lambda - \hat{\lambda}) - \hat{j}_{\lambda\lambda}(\lambda - \hat{\lambda})^2\} + O_p(\| \theta - \hat{\theta} \|^3), \quad (6)$$

where, for example, $n\hat{j}_{\psi\psi} = [-\partial^2 l(\psi, \lambda)/\partial\psi^2]_{\theta = \hat{\theta}}$ and $\hat{j}_{\psi\psi} = i_{\psi\psi} + O_p(n^{-\frac{1}{2}})$. Differentiating (6) with respect to ψ, we have

$$\frac{\partial l(\psi, \lambda)}{\partial \psi} = \frac{n}{2}\{-2\hat{j}_{\psi\psi}(\psi - \hat{\psi}) - 2\hat{j}_{\psi\lambda}(\lambda - \hat{\lambda})\} + O_p(n^{-\frac{1}{2}}).$$

Thus $\hat{\psi}_\lambda$ satisfies

$$0 = \frac{n}{2}\{-2\hat{j}_{\psi\psi}(\hat{\psi}_\lambda - \hat{\psi}) - 2\hat{j}_{\psi\lambda}(\lambda - \hat{\lambda})\} + O_p(n^{-\frac{1}{2}})$$

It follows that

$$\begin{aligned}\hat{\psi}_\lambda - \hat{\psi} &= -\frac{\hat{j}_{\psi\lambda}}{\hat{j}_{\psi\psi}}(\lambda - \hat{\lambda}) + O_p(n^{-1}) \\ &= -\frac{i_{\psi\lambda}}{i_{\psi\psi}}(\lambda - \hat{\lambda}) + O_p(n^{-1}) \\ &= O_p(n^{-\min\left\{\frac{1}{2}+\alpha, 1\right\}})\end{aligned}$$

since $i_{\psi\lambda} = O(n^{-\alpha})$ and $\lambda - \hat{\lambda} = O_p(n^{-\frac{1}{2}})$. This completes the proof. The theorem is extended to the case of p-dimensional parameters, which will not be described here in details.

4 Example

For the location-scale families with density $(1/\lambda)f((x - \psi)/\lambda)$, $\lambda > 0$, $f(x) > 0$ for all x, the elements of the information matrix are

$$I(\psi, \lambda) = \begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\psi\lambda} & i_{\lambda\lambda} \end{pmatrix},$$

where

$$i_{\psi\psi} = \frac{1}{\lambda^2} \int [\frac{f'(y)}{f(y)}]^2 f(y)\, dy,$$

$$i_{\lambda\lambda} = \frac{1}{\lambda^2} \int [\frac{yf'(y)}{f(y)} + 1]^2 f(y)\, dy,$$

and

$$i_{\psi\lambda} = \frac{1}{\lambda^2} \int y[\frac{f'(y)}{f(y)}]^2 f(y)\, dy.$$

See Lehman(1983, p128) for details. Several people in the discussion of the paper by Cox and Reid(1987) pointed out the interesting orthogonality of $\psi + k\lambda$ and λ for suitable k in the location-scale model. However, k is a fairly complicated function of the distribution of the ancillary statistic or of the ancillary statistic itself so that it is difficult to calculate k in numerical works.

Instead, the covariance term $i_{\psi\lambda}$ is zero whenever f is symmetric about the origin. That is, if the standard density of the location-scale model is symmetric about the origin, then the parameters ψ and λ are orthogonal. Therefore, without using the complicated function of the ancillary statistic, we can obtain the advantages of the orthogonal parameters if $i_{\psi\lambda}$ is small enough to $O(n^{-\alpha})$, $\alpha \geq 0$. The intensive numerical works on the inference of the parameters in the location-scale model is prepared in other place, in consideration of the skewness and the location transformation rather than using the system of the partial differential equations in Cox and Reid(1987).

5 Concluding Remarks

We have started how to make inferences about the parameter of interest in the presence of a nuisance parameter(s); profile likelihood, conditional likelihood, Bayesian approach, orthogonality, etc. Among them, we considered the parameter ortogonality by Cox and Reid(1987), which is used as an aid to computation, approximation, interpretation, and elimination of the effect of a nuisance parameter(s). The useful results of parameter orthogonality can be achieved by an α-approximate orthogonality with $\alpha \geq 1/2$.

References

[1] Amari, S. I. (1985). *Differential Geometry in Statistics*. New York; Springer-Verlag.

[2] Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *J. R. Statist.* Soc. B 32, 283-301.

[3] Barndorff-Nielsen, O. E. (1983). On a formula for the distribution of a maximum likelihood estimator. *Biometrika*, Vol. 70, 343-365.

[4] Barndorff-Nielsen, O. E. and Cox, D. R. (1994). *Inference and Asymptotics*. Chapman and Hall.

[5] Box, G. E. P. and Cox, D. R. (1964). An analysis of transformation (with discussion). *J. R. Statist. Soc.*, B 26, 211-252.

[6] Casella, G. and Berger R. L. (1990). *Statistical Inference*. Wadsworth and Brooks/Cole

[7] Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional Inference (with discussion). *J. R. Statist. Soc,*. B 49, 1-39.

[8] Cramer, H. (1946). *Mathematical Methods of Statistics*. Princeton University

[9] Godambe, V. P. (1976). Conditional likelihood and unconditional optimum estimating equations. *Biometrika*, Vol. 63, 277-284.

[10] Godambe, V. P. and Thompson, M. E. (1974). Estimating equations in the presence of a nuisance parameter. *Ann. Statist.*, Vol. 2, 568-571.

[11] Lehmann, E. L. (1983). *Theory of Point Estimation*. John Wiley and Sons.

[12] Pericchi, L. R. (1981). A Bayesian approach to transformations to normality. *Biometrika*, Vol. 68, 35-43.

[13] Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley and Sons.