Academy of Sciences of the Czech Republic
INSTITUTE OF INFORMATION THEORY AND AUTOMATION
Pod vodárenskou věží 4, 182 08 Prague 8, Czech Republic

P. Tichavský

Few General Remarks to Asymptotic Analyses

November 1998

Comments should be sent to: tichavsk@utia.cas.cz

Few General Remarks to Asymptotic Analyses

Petr Tichavský

Abstract

The difference between covariance matrix belonging to an asymptotic distribution of an estimation error and the limit covariance matrix of this error is pointed out. Sufficient conditions for equality of these matrices are suggested.

I. INTRODUCTION

In statistical signal processing, most of theoretical results have form of asymptotic covariances of estimating errors. These results might be of great value for evaluating performances and mutual comparison of parameter estimators, or for optimal tuning of desing variables of estimation methods [1]. However, sometimes the asymptotic results are not correctly interpretated or they are misused to unwarranted conclusions. In this correspondence we want to address a very frequent source of errors, namely the difference between the covariance matrix belonging to an asymptotic distribution of a parameter estimation error and a limit covariance matrix of this error. Some sufficient conditions for equality of these matrices are suggested.

Let $\hat{\theta}_N$ be a random vector indexed by an integer N, which represents an estimate of an unknown parameter θ. N usually stands for length of the data used for the estimation. There exist two statements, which are often mutually mixed.

I. $\sqrt{N}(\hat{\theta}_N - \theta)$ is asymptotically Gaussian distributed with zero mean and covariance Σ for large N. This statement means that the probability distribution function of $\sqrt{N}(\hat{\theta}_N - \theta)$ approaches the Gaussian distribution function with parameters $(0, \Sigma)$ as N goes to infinity. It is sometimes written as [2]

$$\sqrt{N}(\hat{\theta}_N - \theta) \in AsN(0, \Sigma) \tag{1}$$

II. Characterization of the estimator by its bias and by its limit (normalized) covariance matrix,

$$\lim_{N\to\infty} N\,\mathrm{E}[(\hat{\theta}_N - \theta)(\hat{\theta}_N - \theta)^T] = \Sigma \tag{2}$$

[0]This work has been supported by the Grant Agency of the Czech Republic, grant # 102/97/0466.
The author is with the Institute of Information Theory and Automation, Academy of Sciences of Czech Republic, P.O. Box 18, 182 08 Prague, Czech Republic, e-mail: p.tichavsky@ieee.org, fax: +420-2-688 4702.

where "E" denotes the expectation operator and the superscript "T" denotes the transpose.

Clearly, the statement II does not imply I, because (asymptotic) first two moments do not determine unambiguously (asymptotic) distribution function of a random variable or vector. On the other hand, the following two examples show that neither I does not imply II, in general.

Example 1 Put $\hat{\theta}_N = \theta + \frac{1}{\sqrt{N}}g + \frac{1}{N}c$, where g and c are random variables having $N(0,1)$ and a Cauchy distribution, respectively. Due to well known properties of the Cauchy distribution, $\hat{\theta}_N$ has neither a finite mean nor variance, and the statement II is not fulfilled. On the other hand, the probability distribution function of $\sqrt{N}(\hat{\theta}_N - \theta) = g + \frac{1}{\sqrt{N}}\, c$ converges to the probability distribution of g, $N(0,1)$. Thus $\hat{\theta}_N$ obeys the statement I, but not II.

Example 2 Consider the data sample $\{x_n\}_{n=1}^{N}$ randomly drawn from the distribution $N(\theta, 1)$, where $\theta \neq 0$. The sample mean $\bar{x}_N = \frac{1}{N}\sum_{n=1}^{N} x_n$ is an example of an estimator of θ, which obeys both statements, I and II. On the other hand, if the quantity of interest is $s(\theta) = 1/\theta$ and is estimated by $\hat{s}_N = 1/\bar{x}_N$, the situation is different. The distribution of $\sqrt{N}[\hat{s}_N(\theta) - s(\theta)]$ can be shown to be asymptotically normal with zero mean and variance $1/\theta^4$ (we will prove it in Section 2), but the estimator itself does not have a finite mean and variance et al.

The estimators in both of the examples, 1 and 2, do not have finite means and variances for a common reason, which consists in a slow decay of the probability densities of the estimators outside the center of their distributions. Such distributions are called heavy tailed. Note that most of the results published in the literature are of the type I, because they are easier to obtain. Authors of the engineering literature use basically three ways how to face the possible difference between the statements I and II. Some of them simply ignore the difference. Some other authors are aware of the difference but deal only with the asymptotic distribution of the estimate and with the corresponding covariance matrix as the quantity of interest. It is quite common to write [2, p. 242]

$$\mathrm{cov}[\hat{\theta}_N] \sim \frac{1}{N}\Sigma \tag{3}$$

whenever the asymptotic distribution of $\sqrt{N}(\hat{\theta}_N - \theta)$ is $N(0, \Sigma)$ despite the fact that $N\mathrm{cov}[\hat{\theta}_N]$ may not exist or may converge to a matrix different from Σ. This approach might be reasonable in some engineering applications. If the distribution function of $\hat{\theta}_N$ is approximately normal, it can be regarded as a weighted sum of a Gaussian distribution with the weight close to one and some unknown distribution. If the latter distribution function has heavy tails, it might happen that the estimate does not have a finite variance. However, if it is possible to repeat the experiment, then the outcomes of the experiments that deteriorate significantly from other ones can be omitted. The variance of the main cluster of the obtained data is the quantity which people usually need to know, and the true variance of the estimator may be a misleading information in fact. The appearance of the outliers is usually studied separately in connection with the so called threshold effect at low signal-to-noise (SNR) scenarios.

Finally, the third, very small group of the authors so far, makes an effort to ensure validity of both statements, I and II for their estimators.

Some handy tools and theorems for doing asymptotic analysis are summarized in Section 2. Some additional conditions that assure validity of the implication I $\Rightarrow$ II are suggested in Section 3.

II. Tools for Asymptotic Analysis

In this section we review the main tools, stochastic order symbols, which are benefitial to use in asymptotical calculations in probability theory. First, however, we recollect the standard order-of-convergence symbols $O(\cdot)$, $o(\cdot)$ symbols for deterministice sequences.

A sequence $\{x_n\}$ is said to be of order $o(1)$ if it converges to zero for n going to infinity. It is said to be of order $O(1)$, if it is bounded in absolute value by a fixed constant. Let $\{a_n\}$ be a sequence of positive real numbers. The notation $x_n = o(a_n)$ means that $x_n/a_n = o(1)$. The notation $x_n = O(a_n)$ means that $x_n/a_n = O(1)$. A vector sequence $\{x_n\}$ is said to be of order $o(a_n)$ or $O(a_n)$ if the sequence of the vector norms $\{\|x_n\|\}$ is $o(a_n)$ or $O(a_n)$, respectively.

A. Stochastic Order Symbols O_p, o_p

Mann and Wald [5] introduced the symbols $o_p(\cdot)$ and $O_p(\cdot)$ as a way to simplify asymptotic calculations in probability theory. Note in advance that these symbols are tightly connected with the convergence in the sense of probability and convergence in distribution and thus they can be used primarily only to derive results of the type I. In particular, these results do not say anything about convergence of moments.

A sequence of scalar random variables $\{X_n\}$ on a common probability space $(\Omega, \mathcal{A}, P)$ is said to be of order $o_p(1)$ if it converges to zero in probability, that is, if for all $\delta > 0$

$$\lim_{n\to\infty} P(\{|X_n| \geq \delta\}) = 0 \ .$$

A sequence of scalar random variables X_n is said to be of order $O_p(1)$ if it is *bounded in probability*, what means that for all $\varepsilon > 0$ there exists $\delta > 0$ such that, for all n,

$$P(\{|X_n| \geq \delta\}) < \varepsilon \ .$$

The notation $X_n = o_p(a_n)$ means that $X_n/a_n = o_p(1)$. The notation $X_n = O_p(a_n)$ means that $X_n/a_n = O_p(1)$. A vector sequence $\{X_n\}$ is said to be of order $o_p(a_n)$ or $O_p(a_n)$ if the sequence of the vector norms $\{\|X_n\|\}$ is $o_p(a_n)$ or $O_p(a_n)$, respectively.

For list of many useful properties of the $o_p(\cdot)$ and $O_p(\cdot)$ symbols we refer to [8, Appendix C]. Below we list two theorems that make a connecting line between the $o_p(\cdot)$ and $O_p(\cdot)$ symbols and the convergence in distribution, and a theorem about the Taylor series expansion of functions of random variables. These theorems together make a solid ground for an asymptotic analysis.

Theorem 1. If a sequence of scalar or vector random variables X_n converges to X in distribution, then $X_n = O_p(1)$. ■

Theorem 2. Let X_n be a sequence of scalar or vector random variables converging in distribution to X, and let Y_n be another sequence such that $Y_n = X_n + o_p(1)$. Then Y_n converges in distribution to X. ■

Finally, introduce a notation required in Theorem 3. Let $g(x)$ be a function from $\mathcal{R}^K$ to $\mathcal{R}^M$ that is continuous and possesses continuous partial derivatives up to order m. Define $g^{(i)}(x;t)$ by the following recursion,

$$g^{(0)}(x;t) = g(x)$$

$$g^{(i)}(x;t) = \sum_{k=1}^{K} \frac{\partial g^{(i-1)}(x;t)}{\partial x_k}\, t_k, \qquad i = 1, 2, \ldots, m \ .$$

Theorem 3. Let X_n be a sequence of vector random variables satisfying $X_n = x_0 + O_p(a_n)$, where x_0 is a fixed point in $\mathcal{R}^K$ and $\{a_n\}$ is a sequence of positive numbers that tends to zero as $n \to \infty$. Let $g(x)$ be a function from $\mathcal{R}^K$ to $\mathcal{R}^M$ that is continuous and possesses continuous partial derivatives up to order m. Then

$$g(X_n) = \sum_{i=0}^{m} \frac{1}{i!}\, g^{(i)}(x_0; x_n - x_0) + o_p(a_n^m) \ . \tag{4}$$

Corollary 1. In the special case of $m = 1$, the Theorem 3 can be written as

$$X_n = x_0 + O_p(a_n) \quad \Rightarrow \quad g(X_n) = g(x_0) + G(X_n - x_0) + o_p(a_n) \tag{5}$$

where G is the matrix of partial derivatives (the Jacobi matrix) of $g(x)$ evaluated for $x = x_0$. ■

Corollary 2 In this corollary we show how Theorems 1-3 can be used for analysis of asymptotic distribution of an estimator. Let X_n be a random estimate of a vector x_0 such that

$$\frac{1}{a_n}(X_N - x_0) \in AsN(0, \Sigma) \tag{6}$$

where a_n is a real-valued sequence converging to zero. X_n can be interpreted as a sample statistic (usually it is a vector of sample moments). Assume that the parameter of interest θ is given as $\theta = g(x_0)$, where $g(x)$ is a continuous function possessesing continuous first-order partial derivatives in a neighborhood of the point $x = x_0$. We wish to find an asymptotic distribution of $\frac{1}{a_n}[g(X_N) - g(x_0)]$.

Applying Theorem 1 to (6) we get $\frac{1}{a_n}(X_N - x_0) = O_p(1)$ and hence $X_n = x_0 + O_p(a_n)$. Corollary 1 implies that $g(X_n) = g(x_0) + G(X_n - x_0) + o_p(a_n)$ and hence

$$\frac{1}{a_n}[g(X_N) - g(x_0)] = G\frac{1}{a_n}(X_n - x_0) + o_p(1) \ . \tag{7}$$

It follows from (6) that $G\frac{1}{a_n}(X_n - x_0)$ has the asymptotic distribution $N(0, G\Sigma G^T)$. Using Theorem 2 leads to the conclusion (cf. Theorem 3.16 in [8]) that

$$\frac{1}{a_n}[g(X_N) - g(x_0)] \in AsN(0, G\Sigma G^T) \ . \tag{8}$$

The proposition that (6) $\Rightarrow$ (8) is sometimes called the *delta method.* For instance, in Example 2 we have $a_n = 1/\sqrt{N}$, $\Sigma = 1$ and $g(x) = 1/x$. A simple computation gives the desired result $\sqrt{N}(\hat{s}_N - 1/\theta) \in AsN(0, 1/\theta^4)$.

B. Order-of-Magnitude Symbols O_m, o_m

The order-of-magnitude symbol O_m was introduced by Porat and Friedlander in [8] and used for asymptotic statistical analysis of certain polynomial phase signals estimation method.

Let $\{X_n\}$ be a sequence of (real or complex) random variables. The notation $X_n = O_m(1)$ means that all the moments of X_n are bounded uniformly in n. The notation $X_n = O_m(a_n)$ means that $X_n/a_n = O_m(1)$. Analogously to the O_p notation, one could write $X_n = o_m(1)$ if all moments of X_n converge to zero, and $X_n = o_m(a_n)$ would mean that $X_n/a_n = o_m(1)$.

The order-of-magnitude notation directly characterizes moments of random variables. However, its utilization in statistical analysis of estimators is strongly limited. First, the assumption that all moments of a random variable are bounded, is very restrictive. It is valid for random variables with Gaussian and some other short-tailed distributions, but excludes many important distributions with longer tails. The second limitation of the O_m notation is that a desired Taylor series expansion similar to that in Corollary 1 does not hold true in general. For convergence of moments of $g(x)$ requirements about a nice local behavior of function $g(x)$ in a neighborhood of the point $x = x_0$ are not sufficient. Since moment computations involve integration through the whole space, one would have to assume boundeness of all derivatives of the function g over the whole space.

III. Towards proving the convergence of moments

In this section we return to Corollary 2 in the previous section and address the problem of convergence of the covariance matrix of the term $\frac{1}{a_n}[g(X_N) - g(x_0)]$ to the quantity $G\Sigma G^T$ appearing in (8). Indeed, for proving the convergence some additional assumptions are necessary.

The first assumption is boundeness of the estimator $g(x)$. Very often it happens that the estimated quantity must lie for some reason in a boundened space. For example, if the estimated quantity is a frequency, it must lie in a Nyquist interval. In Example 2, for instance, we might know that θ is always greater or equal in absolute value than δ. Then, the estimator can be modified to become bounded in the following way,

$$\hat{s}_N^\delta = \begin{cases} \frac{1}{\bar{x}_N} & \text{for } |\bar{x}_N| \geq \frac{1}{\delta} \\ \frac{1}{\delta}\,\text{sign}(\bar{x}_N) & \text{for } |\bar{x}_N| < \frac{1}{\delta} \ . \end{cases} \tag{9}$$

Second, we will need stronger conditions for convergence of X_n to x_0 than only $X_n = x_0 + O_p(a_n)$. In particular, we will assume the following moment conditions,

$$\frac{1}{a_n^2} E\{[X_n - x_0][X_n - x_0]^T\} = \Sigma + o(1) \tag{10}$$

$$\frac{1}{a_n^2} E(\|X_n - x_0\|^4) = o(1) \ . \tag{11}$$

The following Theorem generalizes the method of the asymptotic analysis used in [6, 7].

Theorem 4. Let X_n be a sequence of vector random variables satisfying the moment conditions (10)-(11), where $\{a_n\}$ is a sequence of positive numbers tending to zero as $n \to \infty$. Let $g(x)$ be a *bounded* function from $\mathcal{R}^K$ to $\mathcal{R}^M$ that is continuous and possesses continuous *second*-order partial derivatives in a neighborhood of a fixed point $x = x_0$. Let G denote the Jacobian of $g(x)$ at point $x = x_0$. Then

$$\frac{1}{a_n^2} E\{[g(X_n) - g(x_0)][g(X_n) - g(x_0)]^T\} = G\Sigma G^T + o(1) \ . \tag{12}$$

Proof: See Appendix.

Example 2, continued. Theorem 4 can be applied to Example 2 with the boundened estimator introduced in (9). The only nontrivial assumption which has to be verified is (11). Since $\bar{x}_N$ has the distribution $N(\theta, \frac{1}{N})$, its centered fourth-order moment is

$$E[(\bar{x}_N - \theta)^4] = \frac{3}{N^2} = O(a_N^4) = o(a_N^2) \tag{13}$$

as required. The conclusion is that $N\ E[(\hat{s}_N^\delta - 1/\theta)^2] = 1/\theta^2 + o(1)$, i.e. that the asymptotic variance and limit variance of the estimator coincide.

IV. Conclusions

In Theorem 4 we presented sufficient conditions for equality of an asymptotic covariance matrix of an estimation error and a normalized limit mean square estimation error. These conditions include boundeness of the estimator and a condition on the fourth-order moment of the sample statistic used for estimation.

Appendix: Proof of Theorem 4

Put

$$\tilde{g}(x) = g(x_0) + G(x - x_0) \qquad \text{for } x \in \mathcal{R}^K\ . \tag{14}$$

Note that $g(x)$ and $\tilde{g}(x)$ have equal first-order Taylor series expansions at $x = x_0$, and hence $\|\tilde{g}(x) - g(x)\| = O(\|x - x_0\|^2)$ for $x \to x_0$. Since in addition $g(x)$ is bounded and $\tilde{g}(x)$ is linear in x, but $\|x - x_0\|^2$ is quadratic and thus asymptotically greater than $\|\tilde{g}(x) - g(x)\|$ for large $\|x - x_0\|$, there exists a constant β such that

$$\|\tilde{g}(x) - g(x)\| \leq \beta \|x - x_0\|^2 \qquad \text{for all } x \in \mathcal{R}^K\ . \tag{15}$$

It immediately follows that

$$E\{\|\tilde{g}(X_n) - g(X_n)\|^2\} \leq \beta^2 E\{\|X_n - x_0\|^4\} = o(a_n^2)\ . \tag{16}$$

In the sequel we shall assume without any lack of generality that $g(x_0) = 0$. The notations $g(X_n)$ and $\tilde{g}(X_n)$ will be simplified to g and $\tilde{g}$, respectively. From (10) we derive that

$$E(\tilde{g}\tilde{g}^T) = E[G(X_n - x_0)(X_n - x_0)^T G^T] = a_n^2\, G\Sigma G + o(a_n^2) = O(a_n^2)\ . \tag{17}$$

Using the Schwartz inequality we get

$$\begin{aligned}
& \left\| E(gg^T) - E(\tilde{g}\tilde{g}^T) \right\| = \left\| E[(g - \tilde{g} + \tilde{g})(g - \tilde{g} + \tilde{g})^T] - E(\tilde{g}\tilde{g}^T) \right\| \\
= \ & \left\| E[(g - \tilde{g})(g - \tilde{g})^T] + E[\tilde{g}(g - \tilde{g})^T] + E[(g - \tilde{g})\tilde{g}^T] \right\| \\
\leq \ & \| E[(g - \tilde{g})(g - \tilde{g})^T] \| + 2\sqrt{E\|\tilde{g}\|^2 \cdot E\|g - \tilde{g}\|^2} \\
= \ & o(a_n^2) + 2\sqrt{O(a_n^2) \cdot o(a_n^2)} = o(a_n^2) \ .
\end{aligned} \tag{18}$$

The desired result (12) follows from (17) and (18).

References

[1] P. Stoica, "Asymptomania ?", *IEEE Signal Processing magazine*, Vol. 15, No.1, p.16, January 1998.

[2] L. Ljung, *System Identification: Theory for the User.* Prentice–Hall, 1987.

[3] B. Porat, *Digital Processing of Random Signals: Theory and Methods.* Prentice–Hall, 1994.

[4] D. Pollard, *Convergence of Stochastic Processes.* Springer–Verlag, 1984.

[5] H.B. Mann and A. Wald, "On stochastic limit and order relationship", *Annals of Math. Statistics*, Vol. 14, pp. 217–226, 1943.

[6] P. Tichavský and A. Swami, "Statistical characterization of sample fourth-order cumulants of a noisy complex sinusoidal process", *IEEE Trans. on Signal Processing*, vol. 43, no.5, pp 1620-1630, July 1995.

[7] P. Händel and P. Tichavský, "Frequency rate estimation at high SNR", *IEEE Tr. on Signal Processing*, Vol. 45, No. 8, pp. 2101-2105.

[8] B. Porat and B. Friedlander, "Asymptotic statistical analysis of the high–order ambiguity function for parameter estimation of polynomial–phase signals", *IEEE Trans. on Information Theory*, Vol. 42, No. 3, pp. 995–1001, May 1996.