ERROR EXPONENTS FOR COMPOSITE HYPOTHESIS TESTING WITH SMALL SAMPLES

Dayu Huang and Sean Meyn

CSL & ECE
University of Illinois at Urbana-Champaign
1308 West Main Street, Urbana, IL 61801, USA

ABSTRACT

We consider the small sample composite hypothesis testing problem, where the number of samples n is smaller than the size of the alphabet m. A suitable model for analysis is the high-dimensional model in which both n and m tend to infinity, and $n = o(m)$. We propose a new performance criterion based on large deviation analysis, which generalizes the classical error exponent applicable for large sample problems (in which $m = O(n)$). The results are:

(i) The best achievable probability of error P_e decays as $-\log(P_e) = (n^2/m)(1 + o(1))J$ for some $J > 0$, shown by upper and lower bounds.

(ii) A coincidence-based test has non-zero generalized error exponent J, and is optimal in the generalized error exponent of missed detection.

(iii) The widely-used Pearson's chi-square test has a *zero* generalized error exponent.

(iv) The contributions (i)-(iii) are established under the assumption that the null hypothesis is uniform. For the non-uniform case, we propose a new test with non-zero generalized error exponent.

Index Terms— chi-square test, high-dimensional model, goodness of fit, large deviations, composite hypothesis testing

1. INTRODUCTION

Composite hypothesis testing problems with small number of samples arise in many applications, such as security and biomedical research. To evaluate a test for these problems, since the exact formula for probability of error is usually complicated, we use asymptotic models and performance criteria that are both insightful and analytically tractable. One such approach is the so-called high-dimensional model, in which the number of samples n and the size of the alphabet m both increase to infinity.

Financial support from the National Science Foundation (NSF CCF 07-29031 and CCF 08-30776), ITMANET DARPA RK 2006-07284 and AFOSR grant FA9550-09-1-0190 is gratefully acknowledged. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF, DARPA or AFOSR.

A widely-used performance criterion is asymptotic consistency: Given some dependency of m on n, does the probability of error tend to zero as n, m tend to infinity? We then consider finer questions, such as the rate of convergence.

To this end, inspiration can be found in the criteria used for large sample problems, in which m is usually fixed or grows very slowly with n. A classical criterion is the error exponent: if the probability of error of a test decreases exponentially fast with respect to n, i.e., $P_e \approx \exp\{-nI\}$, then the rate I is called the error exponent. The popularity of the Generalized Likelihood Ratio Test (GLRT) for the composite testing problem, is partly due to the fact that it has optimal error exponent for fixed m [1]. On the other hand, for the small sample case where m grows very fast, the probability of error does not decay exponentially fast with respect to n; thus the classical error exponent concept is not applicable.

The goal of this paper is to demonstrate that the error exponent criterion can be extended to the small sample case and offers insights that are not available from asymptotic consistency, or criteria based on the central limit theorem.

1.1. Problem statement

Consider the following composite hypothesis testing problem: An i.i.d. sequence $\boldsymbol{Z}_1^n = \{Z_1, \dots, Z_n\}$ is observed, where $Z_i \in [m] := \{1, 2, \dots, m\}$. Denote the set of probability distribution over $[m]$ by $\mathcal{P}([m])$. The null hypothesis $H0$ is simple: Z_i has a *uniform* distribution π over $[m]$ (extensions to the non-uniform case are given in Section 4). The alternative hypothesis $H1$ is a composite one: Z_i has a unknown distribution $\mu \in \Pi_m$, which is given by

$$\Pi_m = \{\mu : d(\mu, \pi) \geq \varepsilon\} \tag{1}$$

where d is the total-variation metric:

$$d(\mu, \pi) = \sup\{|\mu(A) - \pi(A)| : A \subseteq [m]\} = \tfrac{1}{2}\|\mu - \pi\|_1.$$

A test $\phi = \{\phi_n\}_{n \geq 1}$ is a sequence of binary-valued function $\phi_n : [m]^n \to \{0, 1\}$. It decides in favor of $H1$ if $\phi_n = 1$ and $H0$ otherwise. Its performance is evaluated using the probability of false-alarm and worst-case probability of missed detection, defined respectively by

$$P_F(\phi_n) = \mathsf{P}_\pi\{\phi_n = 1\}, \quad P_M(\phi_n) = \sup_{\mu \in \Pi_m} \mathsf{P}_\mu\{\phi_n = 0\}.$$

In the large sample case where $m = O(n)$, the following error exponent criterion has been used to evaluate a test ϕ:

$$\begin{aligned} I_F(\phi) &:= -\limsup_{n\to\infty} n^{-1}\log(P_F(\phi_n)), \\ I_M(\phi) &:= -\limsup_{n\to\infty} n^{-1}\log(P_M(\phi_n)). \end{aligned} \tag{2}$$

In the small sample case where $n = o(m)$, this classical error exponent criterion given in (2) is not applicable since it is zero for all possible test. Our results imply that one should consider the following generalization, defined with respect to the normalization $r(n,m) = n^2/m$:

$$\begin{aligned} J_F(\phi,\mathbf{m}) &:= -\limsup_{n\to\infty} r^{-1}(n,m)\log(P_F(\phi_n)), \\ J_M(\phi,\mathbf{m}) &:= -\limsup_{n\to\infty} r^{-1}(n,m)\log(P_M(\phi_n)). \end{aligned} \tag{3}$$

The limits (but not $r(n,m)$) could depend on how m increases with n, represented by the second argument $\mathbf{m}$ in the notation of J_F and J_M. Note that to have a consistent test, it is necessary that $m = o(n^2)$ ([2]); thus $r(n,m)$ tends to infinity.

For the definition in (3) to be meaningful, we need

(i) There exists a test ϕ for which $\min\{J_F, J_M\}$ is bounded away from zero, uniformly for all sequences $\mathbf{m}$ satisfying $m = o(n^2)$ and $n = o(m)$.

(ii) For any test ϕ, $\min\{J_F, J_M\}$ is finite.

Precise statements are provided in Section 2.

The rest of the paper is organized as follows: Reviews of related work are given Section 1.2; The main results and preliminary analysis are contained in Section 2; The classical Pearson's chi-square test is shown to have a zero error exponent in Section 3. In Section 4, a new test for the case where π is not uniform is proposed and shown to have non-zero error exponents. Numerical results are presented in Section 5.

1.2. Related work

Three types of analysis of a test's performance are asymptotic consistency, Central Limit Theorem (CLT), and large deviation (error exponents). The existing results can be divided, according to the dependency of m on n, into three regimes: 1) fixed m. 2) $m = O(n)$. 3) $n = o(m)$. Representative works in each regime and criterion are listed:

	fixed m	$m = O(n)$	$n = o(m)$
consistency	well-known	well-known	[2][3]
CLT	[4]	[5]	[5]
error exponent	[1]	[6]	this paper

Their main results are summarized as follows:

[2] There is an asymptotically consistent test if and only if $m = o(n^2)$;

[3] There is a test using $n = O(m^{0.5}\text{polylog}(m))$ samples regardless of whether the null distribution is uniform.

[4] Pearson's test statistic is asymptotically χ^2-distributed;

[5] When $\varepsilon \asymp m^{-0.5}$ (CLT analysis), Pearson's chi-square test is asymptotically minimax;

[1] GLRT has optimal error exponents for fixed m;

[6] There exists a test with nonzero classical error exponents (see (2)) if and only if $m = O(n)$.

2. MAIN RESULTS

The main results require the following assumptions:

Assumption 1. *π is the uniform distribution over $[m]$.*

Assumption 2. *$n = o(m)$ and $m = o(n^2)$.*

Under Assumption 1 and 2, the following results hold:

Theorem 2.1 (Achievability)**.** *The following pair of generalized error exponents are achievable by the coincidence-based test ϕ^{K} given in Section 2.1: For $\tau \in [0, \kappa(\varepsilon)]$,*

$$J_F(\phi^{\mathsf{K}}) = \sup_{\theta\geq 0}\{\theta\tau - \tfrac{1}{2}(e^{2\theta} - 1 - 2\theta)\},$$

$$J_M(\phi^{\mathsf{K}}) = \sup_{\theta\geq 0}\{\theta(\kappa(\varepsilon)-\tau) - \tfrac{1}{2}(e^{-2\theta} - 1 + 2\theta)(1+\kappa(\varepsilon))\}$$

where

$$\kappa(\varepsilon) = \begin{cases} \frac{\varepsilon}{1-\varepsilon}, & \varepsilon \geq 0.5, \\ 4\varepsilon^2, & \varepsilon < 0.5. \end{cases} \tag{4}$$

Theorem 2.2 (Converse)**.** *For any test ϕ satisfying*

$$\lim_{n\to\infty} P_F(\phi_n) = 0, \tag{5}$$

the following upper-bound on the generalized error exponent of missed detection holds:

$$J_M(\phi,\mathbf{m}) \leq \bar{J}(\varepsilon).$$

where

$$\bar{J}(\varepsilon) = \tfrac{1}{2}\big(\kappa(\varepsilon) - \log(1+\kappa(\varepsilon))\big). \tag{6}$$

Corollary 2.3 (Optimality of coincidebased test)**.** *The coincidence based test achieves the upper-bound in Theorem 2.2:*

$$\lim_{n\to\infty} P_F(\phi_n^K) = 0, \quad J_M(\phi^K) = \bar{J}(\varepsilon),$$

where $\bar{J}(\varepsilon)$ is given in (6)*.*

We remark that

(i) The main results hold for any possible sequences $\mathbf{m}$ satisfying Assumption 2 (hence we dropped the argument $\mathbf{m}$). An example is $m = n^\alpha$ where $1 < \alpha < 2$.

(ii) Corollary 2.3 implies that the upper-bound is tight when P_F is only required to satisfy (5).

(iii) For other tests, the value of $J_F(\phi,\mathbf{m})$ and $J_M(\phi,\mathbf{m})$ might depend on the sequences $\mathbf{m}$. For example, given two tests with different generalized error exponents, the third test that switches between these tests according to the sequence $\mathbf{m}$ has generalized error exponents that depend on $\mathbf{m}$.

2.1. Achievability

Consider the coincidence-based statistic introduced in [2]:

$$K_n = \sum_{j=1}^{m} \mathbb{I}\{n\Gamma_j^n = 1\} - n, \tag{7}$$

where $\Gamma_j^n = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{Z_i = j\}$ is the empirical distribution. The test is given by $\phi^{\mathsf{K}} = \mathbb{I}\{K_n \leq \mathsf{E}_{\pi^n}[K_n] - \tau_n\}$. The sequence of thresholds $\{\tau_n\}$ we consider has the limit

$$\tau := \lim_{n\to\infty} m\tau_n/n^2. \tag{8}$$

The choice of τ determines the trade-off between J_F and J_M.

To simplify the exposition, we restrict ourselves to distributions in the set

$$\mathcal{P}([m])^{(\eta)} := \{\nu \in \mathcal{P}([m]) : \max_j \nu_j \le m^{-1}\eta\} \quad (9)$$

where η is a large constant. Note that $\pi \in \mathcal{P}([m])^{(\eta)}$. We show in the full version of the paper that distributions in this set (with appropriate choice of η) achieves the worst-case probability of missed detection.

We begin with the expectation and variance of K_n:

$$\mathsf{E}_\nu[K_n] = -\tfrac{n^2}{m}(m\textstyle\sum_{j=1}^m \nu_j^2) + O(\tfrac{n^3}{m^2}),$$

$$\mathsf{Var}_\nu[K_n] = 2\tfrac{n^2}{m}(m\textstyle\sum_{j=1}^m \nu_j^2)(1+o(1)).$$

Chebyshev's bound is used in [2] to establish asymptotic consistency. In fact, applying this bound with $\tau = \frac{1}{2}\kappa(\varepsilon)$ gives

$$P_F(\phi_n^K) \le \frac{\mathsf{Var}_\pi[K_n]}{(\mathsf{E}_{\pi^n}[K_n]-\tau_n)^2} = O(\frac{m}{n^2}),\ P_M(\phi_n^K) = O(\frac{m}{n^2}).$$

However this bound is too loose for our purpose. A tighter bound is obtained via Chernoff,

$$\mathsf{P}_\pi\{K_n \le \mathsf{E}_\pi[K_n]-\tau_n\} \le \exp\{-\theta(\mathsf{E}_\pi[K_n]-\tau_n)+\Lambda_{\pi,K_n}(\theta)\},$$

where $\Lambda_{\nu,K_n}(\theta) = \log(\mathsf{E}_\nu[\exp\{\theta K_n\}])$ is the log-moment-generating function. The probability of missed detection is bounded similarly. The main job is then to obtain an approximation to Λ_{ν,K_n}, given in the following proposition:

Proposition 2.4. *For $\nu \in \mathcal{P}([m])^{(\eta)}$, the n-sample logarithmic moment generating function for the statistic K_n has the following asymptotic expansion*

$$\Lambda_{\nu,K_n}(\theta) = \frac{n^2}{m}(m\sum_{j=1}^m \nu_j^2)\{-\theta+\tfrac{1}{2}[e^{-2\theta}-1+2\theta]\} + O(\frac{n^3}{m^2}) + O(1).$$

This leads directly to the generalized error exponent for the false-alarm. Finding the generalized error exponents for the missed detection is more involved, because the alternative hypothesis is a composite one. The key step is to identify the sequence of "dominating" distributions $\mu \in \Pi_m$, under which the associated probability of missed detection is approximately the largest. In view of Proposition 2.4, such distributions should minimize the quantity $m\sum_{j=1}^m \mu_j^2$. The following elementary result serves this purpose:

Lemma 2.5.

$$\inf_{\mu\in\Pi_m} m\sum_{j=1}^m \mu_j^2 = (1+\kappa(\varepsilon))(1+o(1)).$$

The infimum is achieved approximately by the bi-uniform distribution μ^* given below:

1. *When $\varepsilon \ge 0.5$,*
$$\mu_j^* = \begin{cases} \frac{1}{\lfloor m(1-\varepsilon)\rfloor}, & j \le \lfloor m(1-\varepsilon)\rfloor, \\ 0, & j > \lfloor m(1-\varepsilon)\rfloor. \end{cases}$$
2. *When $\varepsilon < 0.5$,*
$$\mu_j^* = \begin{cases} \frac{1}{m} + \frac{\varepsilon}{\lfloor m/2\rfloor}, & j \le \lfloor m/2\rfloor, \\ \frac{1}{m} - \frac{\varepsilon}{\lceil m/2\rceil}, & j > \lfloor m/2\rfloor. \end{cases}$$

Applying the Gärtner-Ellis Theorem to the sequence of "dominating" distributions, we show that the Chernoff bound is actually tight and obtain Theorem 2.1.

2.2. Converse

Let K_m denote the collection of subsets of $[m]$ that have cardinality $\lfloor m(1-\varepsilon)\rfloor$. For each set $\mathcal{U} \in K_m$, define the distribution $\mu_{\mathcal{U}}$ as

$$\mu_{\mathcal{U},j} = \begin{cases} \frac{1}{\lfloor m(1-\varepsilon)\rfloor}, & j \in \mathcal{U}; \\ 0, & j \in [m]\setminus\mathcal{U}. \end{cases} \quad (10)$$

The converse is proved by showing that the average likelihood ratio is lower-bounded uniformly for any $\boldsymbol{z}_1^n$,

$$\frac{1}{|K_m|}\sum_{\mathcal{U}\in K_m} \frac{\mu_{\mathcal{U}}^n}{\pi^n}(\boldsymbol{z}_1^n) \ge \exp\{-\tfrac{1}{2}\frac{n^2}{m}\frac{\varepsilon}{1-\varepsilon}(1+o(1))\}.$$

Consequently, $\sup_{\mu\in\Pi_m} \mathsf{P}_\mu\{\phi_n=0\}/\mathsf{P}_\pi\{\phi_n=0\}$ satisfies a similar bound. A refinement of the above argument leads to the tighter bound in Theorem 2.2.

3. PEARSON'S TEST IS NOT OPTIMAL

Pearson's chi-square test is a classical test for this composite hypothesis testing problem. The test statistic χ_n^2 is given by

$$\chi_n^2 = \sum_{j=1}^m (n\pi_j)^{-1}(n\Gamma_j^n - n\pi_j)^2.$$

The test $\phi^{\mathsf{P}} = \mathbb{I}\{\chi_n^2 \ge \tau_n\}$ is asymptotically consistent:

Lemma 3.1. *There exists a sequence $\{\tau_n\}$ such that*

$$\lim_{n\to\infty} P_F(\phi_n^{\mathsf{P}}) = 0,\ \lim_{n\to\infty} P_M(\phi_n^{\mathsf{P}}) = 0.$$

Moving beyond consistency, when $\varepsilon = O(m^{-\frac{1}{2}})$, and CLT analysis is applied, Pearson's chi-square test has been shown to be asymptotically minimax [5]. In the asymptotic minimaxity, two tests are compared using the *absolute difference* between probabilities of error; in generalized error exponents, a finer comparison using the *ratio* between probabilities of error is considered, and Pearson's chi-square test is *not* optimal:

Theorem 3.2. *Assume in addition that $m = o(n^2/\log(n)^2)$. For any sequence of threshold $\{\tau_n\}$ satisfying*

$$\lim_{n\to\infty} P_M(\phi_n^{\mathsf{P}}) = 0, \quad (11)$$

the error exponent for probability of false alarm is zero, i.e.,

$$J_F(\phi^{\mathsf{P}}, \mathbf{m}) = 0.$$

Comparing to the coincidence-based test in which each individual summand in the definition of K_n is no larger than 1, the summand in χ_n^2 scales as $(n\Gamma_j^n)^2$. It becomes very large when $n\Gamma_j^n$ is large for some j, leading to a false alarm. Considering the following event,

$$A_n := \{(z_1,\ldots,z_n) : n\Gamma^n(1) = \lfloor 3n/\sqrt{m}\rfloor\}, \quad (12)$$

we claim that this event is likely to cause a false alarm:

$$\mathsf{P}_\pi\{\phi_n^{\mathsf{P}}(Z_1^n) = 1 | A_n\} = 1 - o(1).$$

On the other hand, the probability of A_n decays slowly:

$$\mathsf{P}_\pi(A_n) = \exp\{-1.5(n/\sqrt{m})\log(m)(1+o(1))\}. \quad (13)$$

Combining these two equality gives a lower-bound

$$P_F(\phi_n^{\mathsf{P}}) \ge \mathsf{P}_\pi(A_n)\mathsf{P}_\pi\{\phi_n^{\mathsf{P}}(Z_1^n) = 1 | A_n\}$$

which decays as $(n/\sqrt{m})\log(m)$, slower than n^2/m, and thus $J_F(\phi^{\mathsf{P}}) = 0$.

4. NON-UNIFORM NULL HYPOTHESIS

The coincidence-based test (7) only works for uniform π. For a non-uniform π that satisfies

Assumption 4.1. $\pi \in \mathcal{P}([m])^{(\eta)}$,

we propose the following test statistic:

$$T_n = \sum_{j=1}^{m} \tfrac{1}{2} n^2 \pi_j^2 \mathbb{I}\{n\Gamma_j^n = 0\} - n\pi_j \mathbb{I}\{n\Gamma_j^n = 1\} + \mathbb{I}\{n\Gamma_j^n = 2\}.$$

The new test is given by $\phi_n^{\mathsf{T}} = \mathbb{I}\{T_n \geq \tau_n\}$. The expectation of T_n is:

$$\mathsf{E}_\nu[T_n] = \tfrac{1}{2} \frac{n^2}{m} (m \sum_{j=1}^{m} (\nu_j - \pi_j)^2)(1 + o(1)).$$

The proposed test has nonzero error exponents:

Theorem 4.2. *Suppose Assumption 4.1 holds. For $\tau \in (0, 2\varepsilon^2)$ with τ defined in* (8)*, the following lower-bounds on the error exponents hold:*

$$J_F(\phi^{\mathsf{T}}, \mathbf{m}) \geq \underline{J}_F > 0, \quad J_M(\phi^{\mathsf{T}}, \mathbf{m}) \geq \underline{J}_M > 0.$$

where $\underline{J}_F$ and $\underline{J}_M$ do not depend on $\mathbf{m}$.

5. NUMERICAL EXPERIMENTS

The empirical performance of the coincidence-based test ϕ^{K} is shown in Fig. 1. The *slope* from the theoretical prediction almost matches the actual value. The difference between theoretical prediction and actual value is due to higher-order terms ($O(n^3/m^2)$) in Proposition 2.4, which are not negligible for the range of n and m plotted.

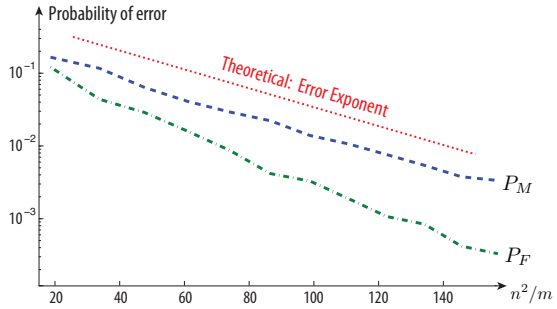

Fig. 1. Performance of ϕ^{K} with $\varepsilon = 0.45$.

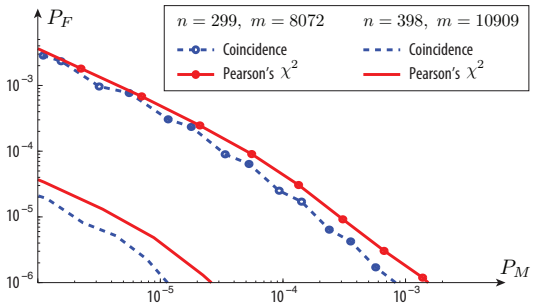

Fig. 2. Error probabilities for Pearson's chi-square test and coincidence-based test: Averaged over 1.5×10^8 runs; $\varepsilon = 0.85$.

The coincidence-based test and Pearson's test are compared in Fig. 2. The difference in performance is visible but not very significant. While Pearson's test has zero error exponent, its error decay given by $\sqrt{n^2/m \log(m)}$ in (13), is not much smaller than n^2/m for the range of n and m plotted. The difference of performance would be more significant when simulating with much larger n and m.

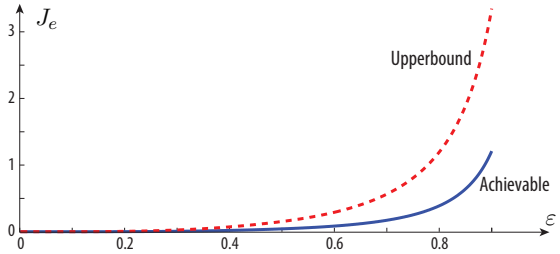

Fig. 3. Bounds on optimal $J_e = \min\{J_F, J_M\}$ for different ε: Upper-bound from Theorem 2.2; Lower-bound from Theorem 2.1 via ϕ^{K}.

6. CONCLUSION

We have shown that the classical error exponent for the composite hypothesis testing problem can be generalized to the small sample case, and this criterion offers insights that are not available in asymptotic consistency analysis, or central-limit-theorem analysis. Future research directions include:

(i) We conjecture that the converse also holds for a non-uniform π. The grouping idea in [6] could be helpful.

(ii) Fig. 3 shows a plot of the upper-bound and lower-bound on the optimal generalized error exponent for the probability of error $J_e = \min\{J_F, J_M\}$. There clearly is room for improvement of the bounds.

(iii) The generalized error exponent concept could be applied to small-sample classification problems.

(iv) In practice, the data might be real-valued. An important and classical problem is how to quantize the real line. The error exponent concept could be useful since it offers a clear view on how quantization affects the test performance via m and ε.

7. REFERENCES

[1] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Ann. Math. Statist.*, vol. 36, pp. 369 – 401, 1965.

[2] L. Paninski, "A coincidence-based test for uniformity given very sparsely sampled discrete data," *IEEE Trans. Inf. Theory*, vol. 54, no. 10, pp. 4750 – 4755, Oct. 2008.

[3] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White, "Testing random variables for independence and identity," in *Proc. 42nd IEEE Symposium on Foundations of Computer Science, 2001.*, October 2001, pp. 442 – 451.

[4] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *Ann. Math. Statist.*, vol. 9, pp. 60 – 62, 1938.

[5] M. S. Ermakov, "Asymptotic minimaxity of chi-square tests," *Theory Probab. Appl.*, vol. 42, p. 589, 1998.

[6] A. R. Barron, "Uniformly powerful goodness of fit tests," *Ann. Statist.*, vol. 17, no. 1, pp. 107 – 124, 1989.