The Annals of Statistics
2018, Vol. 46, No. 1, 1–29
https://doi.org/10.1214/16-AOS1532

CHERNOFF INDEX FOR COX TEST OF SEPARATE PARAMETRIC FAMILIES

BY XIAOOU LI*, JINGCHEN LIU[1],† AND ZHILIANG YING[1],†

University of Minnesota and Columbia University*†

The asymptotic efficiency of a generalized likelihood ratio test proposed by Cox is studied under the large deviations framework for error probabilities developed by Chernoff. In particular, two separate parametric families of hypotheses are considered [In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob.* (1961) 105–123; *J. Roy. Statist. Soc. Ser. B* **24** (1962) 406–424]. The significance level is set such that the maximal type I and type II error probabilities for the generalized likelihood ratio test decay exponentially fast with the same rate. We derive the analytic form of such a rate that is also known as the Chernoff index [*Ann. Math. Stat.* **23** (1952) 493–507], a relative efficiency measure when there is no preference between the null and the alternative hypotheses. We further extend the analysis to approximate error probabilities when the two families are not completely separated. Discussions are provided concerning the implications of the present result on model selection.

1. Introduction. Cox (1961, 1962) introduced the problem of testing two separate parametric families. Let $X_1, \dots, X_n$ be independent and identically distributed real-valued observations from a population with density f with respect to some baseline measure μ. Let $\{g_\theta, \theta \in \Theta\}$ and $\{h_\gamma, \gamma \in \Gamma\}$ denote two separate parametric families of density functions with respect to the same measure μ. Consider testing H_0: $f \in \{g_\theta, \theta \in \Theta\}$ against H_1: $f \in \{h_\gamma, \gamma \in \Gamma\}$. To avoid singularity, we assume that all the distributions in the families g_θ and h_γ are mutually absolutely continuous so that the likelihood ratio stays away from zero and infinity. Furthermore, we assume that the model is correctly specified, that is, f belongs to either the g-family or the h-family.

Recently revisiting this problem, Cox (2013) mentioned several applications such as the one-hit and two-hit models of binary dose-response and testing of interactions in a balanced 2^k factorial experiment. Furthermore, this problem has been studied in econometrics [Vuong (1989), White (1982a, 1982b), Pesaran (1974), Pesaran and Deaton (1978), Davidson and MacKinnon (1981)]. For more applications of testing separate families of hypotheses, see Berrington de González and

Received January 2016; revised November 2016.

[1] Supported in part by NSF Grant SES-1323977, NSF Grant IIS-1633360, NIH Grant R01GM047845 and Army Research Office Grant W911NF-15-1-0159.

MSC2010 subject classifications. Primary 62F03; secondary 62J12, 62F12.

Key words and phrases. Asymptotic relative efficiency, generalized likelihood ratio, generalized linear models, large deviation, model selection, nonnested hypotheses, variable selection.

Cox (2007) and Braganca Pereira (2005), and the references therein. Furthermore, there is a discussion of model misspecification, that is, f belongs to neither the g-family nor the h-family, which is beyond the current discussion. For semiparametric models, Fine (2002) proposed a similar test for nonnested hypotheses under the Cox proportional hazards model assumption.

Cox (1962) considers the test statistic $l = l_g(\hat{\theta}) - l_h(\hat{\gamma}) - E_{g_{\hat{\theta}}}\{l_g(\hat{\theta}) - l_h(\hat{\gamma})\}$, where $l_g(\theta)$ and $l_h(\gamma)$ are the log-likelihood functions under the g-family and the h-family, respectively, and $\hat{\theta}$ and $\hat{\gamma}$ are the corresponding maximum likelihood estimators. Rigorous distributional derivations of statistic l can be found in Huber (1967) and White (1982a, 1982b). In this paper, we consider the generalized likelihood ratio statistic

$$(1) \qquad \mathrm{LR}_n = \frac{\max_{\gamma\in\Gamma}\prod_{i=1}^n h_\gamma(X_i)}{\max_{\theta\in\Theta}\prod_{i=1}^n g_\theta(X_i)} = e^{l_h(\hat{\gamma}) - l_g(\hat{\theta})}$$

that is slightly different from Cox's approach. We are interested in the Chernoff efficiency, whose definition is provided in Section 2.1, of the generalized likelihood ratio test.

In the hypothesis testing literature, there are several measures of asymptotic relative efficiency for simple null hypothesis against simple alternative hypothesis. Let n_1 and n_2 be the necessary sample sizes for each of two testing procedures to perform equivalently in the sense that they admit the same type I and type II error probabilities. Then the limit of ratio n_1/n_2 in the regime that both sample sizes tend to infinity represents the asymptotic relative efficiency between these two procedures.

Relative efficiency depends on the asymptotic manner of the two types of error probabilities with large samples. Under different asymptotic regimes, several asymptotic efficiency measures are proposed and they are summarized in Chapter 10 of Serfling (1980). Under the regime of Pitman efficiency, several asymptotically equivalent tests to the Cox test exist. Furthermore, Pesaran (1984) and Rukhin (1993) applied Bahadur's criterion of asymptotic comparison [Bahadur (1960, 1967)] to tests for separate families and compared different tests for lognormal against exponential distribution and for nonnested linear regressions. There are other efficiency measures that are frequently considered, such as Kallenberg efficiency [Kallenberg (1983)].

In the context of testing a simple null hypothesis against a fixed simple alternative hypothesis, Chernoff (1952) introduces a measure of asymptotic efficiency for tests based on sum of independent and identically distributed observations, a special case of which is the likelihood ratio test. This efficiency is introduced by showing no preference between the null hypothesis and the alternative hypothesis. The rejection region is setup such that the two types of error probabilities decay at the same exponential rate ρ. The rate ρ is later known as the Chernoff index. A brief summary of the Chernoff index is provided in Section 2.1.

The basic strategy of Chernoff (1952) is to apply large deviations techniques to the log-likelihood ratio statistic and compute/approximate the probabilities of the two types of errors. Under the situation when either the null hypothesis or the alternative hypothesis is composite, one naturally considers the generalized likelihood ratio test. To the authors' best knowledge, the asymptotic behavior of the generalized likelihood ratio test under the Chernoff's regime remains an open problem. This is mostly because large deviations results are not directly applicable as the test statistic is the ratio of the supremums of two random functions. This paper fills in this void and provides a definitive conclusion of the asymptotic efficiency of the generalized likelihood ratio test under Chernoff's asymptotic regime. We define the Chernoff index via the asymptotic decay rate of the maximal type I and type II error probabilities that is also the minimax risk corresponding to the zero-one loss function.

We compute the generalized Chernoff index of the generalized likelihood ratio test for two separate parametric families that keep a certain distance away from each other. That is, the Kullback–Leibler distance between g_θ and h_γ are bounded away from zero for all $\theta \in \Theta$ and $\gamma \in \Gamma$. We use $\rho_{\theta\gamma}$ to denote the Chernoff index of the likelihood ratio test for the simple null $H_0: \ f = g_\theta$ against simple alternative $H_1: \ f = h_\gamma$. Under mild moment conditions, we show that the exponential decay rate of the maximal error probabilities is simply the minimum of the one-to-one Chernoff index $\rho_{\theta\gamma}$ over the parameter space, that is, $\rho = \min_{\theta,\gamma} \rho_{\theta\gamma}$. This result suggests that the generalized likelihood ratio test is asymptotically the minimax strategy in the sense that with the same sample size it achieves the optimal exponential decay rate of the maximal type I and type II error probabilities when they decay equally fast. The present result can also be generalized to asymptotic analysis of Bayesian model selection among two or more families of distributions. A key technical component is to deal with the excursion probabilities of the likelihood functions, for which random field and nonexponential change of measure techniques are applied. This paper also in part corresponds to the conjecture in Cox (2013) "formal discussion of possible optimality properties of the test statistics would, I think, require large deviation theory" though we consider a slightly different statistic.

We further extend the analysis to the cases when the two families may not be completely separate, that is, one may find two sequences of distributions in each family and the two sequences converge to each other, or the two families may simply overlap, but not nested, as in the case of the Weibull family versus the gamma family. For this case, the generalized Chernoff index as described above is zero. An alternative and more meaningful formulation is to consider the asymptotic decay rate of the type I error probability under a fixed distribution g_{θ_0} which belongs to H_0, but is bounded away from H_1. Since the roles of H_0 and H_1 are switchable, it also gives the decay rate for the type II error. This formulation is clearly applicable to both separated and nonseparated families, and thus *it provides a means to approximate the error probabilities of the generalized likelihood ratio test for* *general parametric families*. The results established under this setting will have important theoretical as well as practical implications in hypothesis testing, model selection and other areas where maximum likelihood is employed. In particular, we show how the results are applied to selecting a set of covariates among competing sets in generalized linear models.

The rest of this paper is organized as follows. We present our main results for separate families of hypotheses in Section 2. Further extension to more than two families and Bayesian model selection is discussed in Section 3. Results for possibly nonseparate families are presented in Section 4. Numerical examples are provided in Section 5. Finally, a concluding remark is given in Section 6.

2. Main results.

2.1. *Simple null against simple alternative—a review of Chernoff index.*

In this section, we state the main results and their implications. To start with, we provide a brief review of Chernoff index for simple null versus simple alternative; then we proceed to the case of simple null versus composite alternative; furthermore, we present the generalized Chernoff index for the composite null versus composite alternative.

Under the context of simple null hypothesis versus simple alternative hypothesis, we have the null hypothesis H_0: $f = g$ and the alternative hypothesis H_1: $f = h$. We write the log-likelihood ratio of each observation as $l^i = \log h(X_i) - \log g(X_i)$. Then the likelihood ratio is $\mathrm{LR}_n = \exp(\sum_{i=1}^n l^i)$. We use l to denote the generic random variable equal in distribution to l^i. We define the moment generating function of l under distribution g as $M_g(z) = E_g(e^{zl}) = \int\{h(x)/g(x)\}^z g(x)\mu(dx)$, which must be finite for $z \in [0, 1]$ by the Hölder inequality. Furthermore, we define the rate function $m_g(t) = \max_z[zt - \log\{M_g(z)\}]$. The following large deviations result is established in Chernoff (1952).

PROPOSITION 1. *If* $t < E_g(l)$, *then* $\log P_g(\mathrm{LR}_n < e^{nt}) \sim -n \times m_g(t)$; *if* $t > E_g(l)$, *then* $\log P_g(\mathrm{LR}_n > e^{nt}) \sim -n \times m_g(t)$.

We write $a_n \sim b_n$ if $a_n/b_n \to 1$ as $n \to \infty$. The above proposition provides an asymptotic decay rate of the type I error probability: for any $t > E_g(l)$, $P_g(\mathrm{LR}_n > e^{nt}) = e^{-\{1+o(1)\}n\times m_g(t)}$, as $n \to \infty$. Similarly, we switch the roles of g and h and define $M_h(z)$ and $m_h(t)$ by flipping the sign of the log-likelihood ratio $l = \log g(X) - \log h(X)$ and computing the expectations under h. One further defines $\rho(t) = \min\{m_g(t), m_h(-t)\}$ that is the slower rate among the type I and type II error probabilities. A measure of efficiency is given by

$$\rho = \max_{E_g(l)<t<E_h(l)} \rho(t) \tag{2}$$

that is known as the Chernoff index between g and h.

In the decision framework, we consider the zero-one loss function

$$L(C,f,X_1,\dots,X_n)=\begin{cases}1 & \text{if } f=g \text{ and } (X_1,\dots,X_n)\in C,\\ 1 & \text{if } f=h \text{ and } (X_1,\dots,X_n)\notin C,\\ 0 & \text{otherwise,}\end{cases} \tag{3}$$

where $C\subset R^n$ and f is a density function. Then the risk function is

$$R(C,f)=E_f\{L(C,f,X_1,\dots,X_n)\}=\begin{cases}P_g(C) & \text{if } f=g,\\ P_h(C^c) & \text{if } f=h.\end{cases} \tag{4}$$

The Chernoff index is the asymptotic exponential decay rate of the minimax risk $\min_C \max_f R(C,f)$ within the family of tests. In the following section, we will generalize the Chernoff efficiency following the minimaxity definition.

Using the fact that $M_g(z)=M_h(1-z)$, one can show that the optimization in (2) is solved at $t=0$ and

$$\rho=\rho(0). \tag{5}$$

Both $m_g(t)$ and $m_h(-t)$ are monotone functions of t and (5) suggests that $\rho=m_g(0)=m_h(0)$. To achieve the Chernoff index, we reject the null hypothesis if the likelihood ratio statistic is greater than 1 and the type I and type II error probabilities have identical exponential decay rate ρ.

To have a more concrete idea of the above calculations, Figure 1 in the supplementary material [Li, Liu and Ying (2017)] shows one particular $-\log\{M_g(z)\}$ as a function of z where $g(x)$ is lognormal and $h(x)$ is exponential. There are several useful facts. First, $-\log\{M_g(z)\}$ is a concave function of z and $-\log\{M_g(0)\}=-\log\{M_g(1)\}=0$. There are several useful facts. First, $-\log\{M_g(z)\}$ is a concave function of z and $-\log\{M_g(0)\}=-\log\{M_g(1)\}=0$. The maximization $\max_z[zt-\log\{M_g(z)\}]$ is solved at $d\log\{M_g(z)\}/dz=t$. Furthermore, the Chernoff index is achieved at $t=0$. We insert $t=0$ into the maximization and the Chernoff index is $\rho=\max_z[-\log\{M_g(z)\}]$.

2.2. *Generalized Chernoff index for testing composite hypothesis.*

In this subsection, we develop the corresponding results for testing composite hypotheses. Some technical conditions are required as follows:

A1 Complete separation: $\min_{\theta\in\Theta,\gamma\in\Gamma} E_{g_\theta}\{\log g_\theta(X)-\log h_\gamma(X)\}>0$.

A2 The parameter spaces Θ and Γ are compact subsets of R^{d_g} and R^{d_h} with continuously differentiable boundaries $\partial\Theta$ and $\partial\Gamma$, respectively.

A3 Define $l_{\theta\gamma}=\log h_\gamma(X)-\log g_\theta(X)$, $S_1=\sup_{\theta,\gamma}|\nabla_\theta l_{\theta\gamma}|$, and $S_2=\sup_{\theta,\gamma}|\nabla_\gamma l_{\theta\gamma}|$. There exists some $\eta, x_0>0$, that are independent with θ and γ, such that for $x>x_0$

$$\sup_{\theta\in\Theta,\gamma\in\Gamma}\max\{P_{g_\theta}(S_i>x), P_{h_\gamma}(S_i>x)\}\le e^{-(\log x)^{1+\eta}} \qquad (i=1,2). \tag{6}$$

REMARK 2. Condition A3 requires certain tail conditions of S_i. It excludes some singularity cases. This condition is satisfied by most parametric families. For instance, if $g_\theta(x) = g_0(x)e^{\theta x - \varphi_g(\theta)}$ and $h_\gamma = h_0(x)e^{\gamma x - \varphi_h(\gamma)}$ are exponential families, then $|\nabla_\theta l_{\theta\gamma}| = |x - \varphi_g'(\theta)| \le |x| + O(1)$. Thus (6) is satisfied if $|x|$ has a finite moment generating function.

If $g_\theta = g(x - \theta)$ is the location family, then $|\nabla_\theta l_{\theta\gamma}| = |\frac{g'(x-\theta)}{g(x-\theta)}|$ usually has a finite moment generating function for light-tailed distributions (Gaussian, exponential, etc.) and is usually bounded for heavy-tailed distributions (e.g., t-distribution). Similarly, one may verify (6) for scale families. Thus, A3 is a weak condition and is applicable to most parametric families practically in use.

We start the discussion for a simple null hypothesis against a composite alternative hypothesis

$$H_0: \ f = g \quad \text{and} \quad H_1: \ f \in \{h_\gamma : \gamma \in \Gamma\}. \tag{7}$$

In this case, the likelihood ratio takes the following form:

$$\mathrm{LR}_n = \frac{\max_{\gamma\in\Gamma} \prod_{i=1}^n h_\gamma(X_i)}{\prod_{i=1}^n g(X_i)}. \tag{8}$$

For each distribution h_γ in the alternative family, let ρ_γ be the Chernoff index of the likelihood ratio test for $H_0: \ f = g$ against $H_1: \ f = h_\gamma$, whose form is given as in (2). The first result is given as follows.

LEMMA 3. *Consider the hypothesis testing problem given as in* (7) *and the generalized likelihood ratio test with rejection region* $C_\lambda = \{(x_1, \dots, x_n) : \mathrm{LR}_n > \lambda\}$ *where* LR_n *is given by* (8). *If conditions* A1–3 *are satisfied and we choose* $\lambda = 1$, *then the asymptotic decay rate of the type I and maximal type II error probabilities are identical, more precisely,*

$$\log P_g(C_1) \sim \sup_{\gamma\in\Gamma} \log P_{h_\gamma}(C_1^c) \sim -n \times \min_\gamma \rho_\gamma.$$

For composite null versus composite alternative

$$H_0: \ f \in \{g_\theta : \theta \in \Theta\} \quad \text{against} \quad H_1: \ f \in \{h_\gamma : \gamma \in \Gamma\}, \tag{9}$$

similar results can be obtained. The generalized likelihood ratio statistic is given by (1). For each single pair (g_θ, h_γ), we let $\rho_{\theta\gamma}$ denote the corresponding Chernoff index of the likelihood ratio test for $H_0: f = g_\theta$ and $H_1: f = h_\gamma$. The following theorem states the main result.

THEOREM 4. *Consider a composite null hypothesis against a composite alternative hypothesis given as in* (9) *and the generalized likelihood ratio test with* *rejection region* $C_\lambda = \{(x_1, \dots, x_n) : \mathrm{LR}_n > \lambda\}$ *where* LR_n *is given by* (1). *If conditions* A1–3 *are satisfied and we choose* $\lambda = 1$, *then the asymptotic decay rate of the maximal type I and type II error probabilities are identical*, *more precisely*,

$$\sup_{\theta\in\Theta} \log P_{g_\theta}(C_1) \sim \sup_{\gamma\in\Gamma} \log P_{h_\gamma}\bigl(C_1^c\bigr) \sim -n \times \min_{\theta\in\Theta,\gamma\in\Gamma} \rho_{\theta\gamma}. \tag{10}$$

We call $\rho = \min_{\theta,\gamma} \rho_{\theta\gamma}$ the generalized Chernoff index between the two families $\{g_\theta\}$ and $\{h_\gamma\}$ that is the exponential decay rate of the maximal type I and type II error probabilities for the generalized likelihood ratio test. We would like to make a few remarks. Suppose that $\rho_{\theta\gamma}$ is minimized at θ_* and γ_*. The maximal type I and type II error probabilities of C_1 have identical exponential decay rate as that of the error probabilities of the likelihood ratio test for the simple null $H_0 : f = g_{\theta_*}$ versus simple alternative $H_1 : f = h_{\gamma_*}$ problem. Then, according to the Neyman–Pearson lemma, we have the following statement. Among all the tests for (9) that admit maximal type I error probabilities that decays exponentially at least at rate ρ, their maximal type II error probabilities decay at most at rate ρ. This asymptotic efficiency can only be obtained at the particular threshold $\lambda = 1$, at which the maximal type I and the type II error probabilities decay exponentially equally fast. Consider the loss function as in (3) and the risk function is

$$R(C, f) = \begin{cases} P_f(C) & \text{if } f \in \{g_\theta : \theta \in \Theta\}, \\ P_f(C^c) & \text{if } f \in \{h_\gamma : \gamma \in \Gamma\}. \end{cases} \tag{11}$$

According to the above discussion, the maximum risk of the rejection region $C_1 = \{\mathrm{LR}_n > 1\}$ achieves the same asymptotic decay rate as that of the minimax risk that is $\min_{C\subset R^n} \max_{f\in\{g_\theta\}\cup\{h_\gamma\}} \log\{R(C, f)\}/n \to -\rho$.

Upon considering the exponential decay rate of the two types of error probabilities, one can simply reduce the problem to testing $H_0 : f = g_{\theta_*}$ against $H_1 : f = h_{\gamma_*}$. Each of these two distributions can be viewed as the least favorable distribution if its own family is chosen to be the null family. The results in Lemma 3 and Theorem 4 along with their proofs suggest that the maximal type I and type II error probabilities are achieved at $f = g_{\theta_*}$ and $f = h_{\gamma_*}$. In addition, under the distribution g_{θ_*} and conditional on the event C_1, in which H_0 is rejected, the maximum likelihood estimator $\hat\gamma$ converges to γ_*; vice versa, under the distribution $f = h_{\gamma_*}$, if H_0 is not rejected, the maximum likelihood estimator $\hat\theta$ converges to θ_*.

2.3. *Relaxation of technical conditions.* The results of Lemma 3 and Theorem 4 require three technical conditions. Condition A1 ensures that the two families are separated and it is crucial for the exponential decay of the error probabilities. Condition A2, though important for the proof, can be relaxed for most parametric families. They can be replaced by certain localization conditions for the maximum likelihood estimator. We present one as follows:

A4 There exist parameter-dependent compact sets $A_\theta, \tilde{A}_\gamma \subset \Gamma$ and $B_\gamma, \tilde{B}_\theta \subset \Theta$ such that for all θ and γ

$$\begin{aligned} &\liminf_{n\to\infty} \frac{1}{n} \log P_{g_\theta}(\hat{\theta} \in \tilde{B}_\theta^c \text{ or } \hat{\gamma} \in A_\theta^c) < -\rho, \\ &\liminf_{n\to\infty} \frac{1}{n} \log P_{h_\gamma}(\hat{\theta} \in B_\gamma^c \text{ or } \hat{\gamma} \in \tilde{A}_\gamma^c) < -\rho, \end{aligned} \tag{12}$$

where $\hat{\theta}$ and $\hat{\gamma}$ are the maximum likelihood estimators under the two families. Condition A3 is satisfied if the maximization in the definition of S_i is taken on the set A_θ and $\tilde{B}_\theta$ when the tail is computed under g_θ and is taken on the set $\tilde{A}_\gamma$ and B_γ when the tail is computed under h_γ.

REMARK 5. Assumption A4 can be verified by means of large deviations of the maximum likelihood estimator; see Arcones (2006). Under regularity conditions, the probability that the maximum likelihood estimator deviates from the true parameter by a constant decreases exponentially. One can choose the constant large enough so that it decays at a faster rate than ρ and thus Assumption 4 is satisfied.

Consider the first probability in (12) under g_θ. We typically choose $\tilde{B}_\theta$ to be a reasonably large compact set containing θ, and thus $P_{g_\theta}(\hat{\theta} \in \tilde{B}_\theta^c)$ decays exponentially fast at a higher rate than ρ. For the choice of A_θ, we first define $\gamma_\theta = \arg\max_{\gamma\in\Gamma} E_{g_\theta}\{\log h_\gamma(X)\}$ that is the limit of $\hat{\gamma}$ under g_θ. Then we choose A_θ be a sufficiently large compact set containing γ_θ so that the decay rate of $P_{g_\theta}(\hat{\gamma} \in A_\theta^c)$ is higher than ρ. Similarly, we can choose B_γ and $\tilde{A}_\gamma$. Furthermore, the maximum score function for a single observation over a compact set usually has a sufficiently light tail to satisfy condition A4, for instance, $P_{g_\theta}(\sup_{\theta\in\tilde{B}_\theta,\gamma\in A_\theta} |\nabla_\theta l_{\theta\gamma}| > x) \le e^{-(\log x)^{1+\eta}}$.

COROLLARY 6. *Consider a composite null hypothesis against composite alternative hypothesis given as in* (9). *Suppose that conditions* A1 *and* A4 *are satisfied. Then the asymptotic decay rates of the maximal type I and type II error probabilities are identical*, *more precisely*, $\sup_{\theta\in\Theta} \log P_{g_\theta}(C_1) \sim \sup_{\gamma\in\Gamma} \log P_{h_\gamma}(C_1^c) \sim -n \times \min_{\theta,\gamma} \rho_{\theta\gamma}$.

The proof of this corollary is very similar to that of Theorem 4 and, therefore, is included in the supplementary material.

3. Extensions.

3.1. *On asymptotic behavior of Bayes factor.* The result in Theorem 4 can be further extended to the study of Bayesian model selection. Consider the two families in (9) each of which is endowed with a prior distribution on its own parameter space, denoted by $\phi(\theta)$ and $\varphi(\gamma)$. We use $\mathcal{M}$ to denote the family membership: $\mathcal{M}=0$ for the g-family and $\mathcal{M}=1$ for the h-family. Then the Bayes factor is

$$\text{(13)} \qquad \mathrm{BF} = \frac{p(X_1,\dots,X_n|\mathcal{M}=1)}{p(X_1,\dots,X_n|\mathcal{M}=0)} = \frac{\int_{\gamma\in\Gamma}\varphi(\gamma)\prod_{i=1}^n h_\gamma(X_i)\,d\gamma}{\int_{\theta\in\Theta}\phi(\theta)\prod_{i=1}^n g_\theta(X_i)\,d\theta}.$$

With a similar derivation as that of Bayesian information criterion [Schwarz (1978)], the marginalized likelihood $p(X_1,\dots,X_n|\mathcal{M}=i)$ is the maximized likelihood multiplied by a polynomial prefactor depending on the dimension of the parameter space. Therefore, we can approximate the Bayesian factor by the generalized likelihood ratio statistic as follows:

$$\text{(14)} \qquad \kappa^{-1}n^{-\beta} \le \frac{\mathrm{BF}}{\mathrm{LR}_n} \le \kappa n^{\beta}$$

for some κ and β sufficiently large. Therefore, $\log\mathrm{BF} = \log\mathrm{LR}_n + O(\log n)$. Since the expectation of $\log\mathrm{LR}_n$ is of order n, the $O(\log n)$ term does not affect the exponential rate. Therefore, we have the following result.

THEOREM 7. *Consider two families of distributions given as in* (9) *satisfying conditions* A1–3. *The prior densities φ and ϕ are positive and Lipschitz continuous. We select $\mathcal{M}=1$ if* BF >1 *and $\mathcal{M}=0$ otherwise where* BF *is given by* (13). *Then the asymptotic decay rate of selecting the wrong model are identical under each of the two families. More precisely,*

$$\begin{aligned}\log\int_{\theta\in\Theta} P_{g_\theta}(\mathrm{BF}>1)\phi(\theta)\,d\theta &\sim \sup_{\theta\in\Theta}\log P_{g_\theta}(\mathrm{BF}>1)\\ &\sim \log\int_{\gamma\in\Gamma} P_{h_\gamma}(\mathrm{BF}\le 1)\varphi(\gamma)\,d\gamma\\ &\sim \sup_{\gamma\in\Gamma}\log P_{h_\gamma}(\mathrm{BF}\le 1)\\ &\sim -n\times\min_{\theta,\gamma}\rho_{\theta\gamma}.\end{aligned}$$

The proof of the above theorem is an application of Theorem 4 and (14), and thus we omit it. The above result does not rely on the validity of the prior distributions. Therefore, model selection based on Bayes factor is asymptotically efficient even if the prior distribution is misspecified. That is, the Bayes factor is calculated based on the probability measures with density functions φ and ϕ that are different from the true prior probability measures under which θ and γ are generated.

3.2. *Extensions to more than two families.* Suppose that there are K nonoverlapping families $\{g_{k,\theta_k}:\theta_k\in\Theta_k\}$ for $k=1,\dots,K$, among which we would like to select the true family to which the distribution f belongs. Let $L_k(\theta_k) =$

$\prod_{i=1}^{n} g_{k,\theta_k}(X_i)$ be the likelihood of family k. A natural decision is to select the family that has the highest likelihood, that is, $\hat{k} = \arg\max_{k=1,\dots,K} \sup_{\theta_k} L_k(\theta_k)$. According to the results in Theorem 4, we obtain that $\sup_{k,\theta_k} \log P_{g_{k,\theta_k}}(\hat{k} \neq k) \sim -n\rho$, where ρ is the smallest generalized Chernoff indices, defined as in Theorem 4, among all the $(K-1)K/2$ pairs of families. To obtain the above limit, one simply considers each family k as the null hypothesis and the union of the rest $K-1$ altogether as the alternative hypothesis.

With the same argument as in Section 3.1, we consider Bayesian model selection among the K families each of which is endowed with a prior $\phi_k(\theta_k)$. Consider the marginalized maximum likelihood estimator $\hat{k}_B = \arg\max_k \int L_k(\theta_k)\phi_k(\theta_k)\,d\theta_k$ that admits the same misclassification rate $\sup_{k,\theta_k} \log P_{g_{k,\theta_k}}(\hat{k}_B \neq k) \sim \sup_k \log \int P_{g_{k,\theta_k}}(\hat{k}_B \neq k)\phi_k(\theta_k)\,d\theta_k \sim -n\rho$.

4. Results for possibly nonseparated families.

4.1. *The asymptotic approximation of error probabilities.* We extend the results to the cases when the g-family and the h-family are not necessarily separated, that is,

$$\min_{\theta\in\Theta,\gamma\in\Gamma} E_{g_\theta}\{\log g_\theta(X) - \log h_\gamma(X)\} = 0. \tag{15}$$

In the case of (15), the Chernoff index is trivially zero. We instead derive the asymptotic decay rate of the following error probabilities. For some $\theta_0 \in \Theta$ such that $\min_\gamma E_{g_{\theta_0}}\{\log g_{\theta_0}(X) - \log h_\gamma(X)\} > 0$, we consider the type I error probability $P_{g_{\theta_0}}(\mathrm{LR}_n > e^{nb})$ as $n \to \infty$ where LR_n is the generalized likelihood ratio statistic as in (1). For b, we require that

$$\sup_{\gamma\in\Gamma} E_{g_{\theta_0}}\{\log h_\gamma(X) - \log g_{\theta_0}(X)\} < b \tag{16}$$

ensuring that $P_{g_{\theta_0}}(\mathrm{LR}_n > e^{nb})$ eventually converges to zero.

The statement of the theorem requires the following construction. For each θ and γ, we first define the moment generating function of $\log h_\gamma(X) - \log g_\theta(X) - b$ as $M_{g_{\theta_0}}(\theta,\gamma,\lambda) = E_{g_{\theta_0}}[\exp\{\lambda(\log h_\gamma(X) - \log g_\theta(X) - b)\}]$ and consider the optimization problem

$$M^\dagger_{g_{\theta_0}} \triangleq \inf_{\theta\in\Theta} \sup_{\gamma\in\Gamma} \inf_{\lambda\geq 0} M_{g_{\theta_0}}(\theta,\gamma,\lambda). \tag{17}$$

Under Assumption A2, there exists at least one solution to the above optimization and let $(\theta^\dagger,\gamma^\dagger,\lambda^\dagger)$ denote one of them. Furthermore, we define a measure $Q^\dagger$ that is absolutely continuous with respect to $P_{g_{\theta_0}}$:

$$\frac{dQ^\dagger}{dP_{g_{\theta_0}}} = \exp\{\lambda^\dagger(\log h_{\gamma^\dagger}(X) - \log g_{\theta^\dagger}(X) - b)\}/M^\dagger_{g_{\theta_0}}. \tag{18}$$

REMARK 8. The inner most optimization of the rate function (17) has a close connection to the Kullback–Leibler divergence. In particular, for each θ and γ, $\inf_{\lambda \geq 0} M_{g_{\theta_0}}(\theta, \gamma, \lambda)$ has the following representation. Define a family of measure $dQ_\lambda / dP_{g_{\theta_0}} = \exp\{\lambda(\log h_\gamma(X) - \log g_\theta(X) - b)\}/M_{g_{\theta_0}}(\theta, \gamma, \lambda)$ and the Kullback–Leibler divergence $D(Q_\lambda | P_{g_{\theta_0}}) = E^{Q_\lambda}\{\log(dQ_\lambda / dP_{g_{\theta_0}})\}$. Then we have the following representation:

$$-\inf_{\lambda \geq 0} \log M_{g_{\theta_0}}(\theta, \gamma, \lambda) = \inf_{\lambda : E^{Q_\lambda}\{\log h_\gamma(X) - \log g_\theta(X)\} \geq b} D(Q_\lambda | P_{g_{\theta_0}}).$$

It is straightforward to verify that the optimization on the right-hand side is solved on the boundary. Let λ^* solve $E^{Q_{\lambda^*}}\{\log h_\gamma(X) - \log g_\theta(X)\} = b$. Thus, the rate function is $-\inf_{\lambda \geq 0} \log M_{g_{\theta_0}}(\theta, \gamma, \lambda) = D(Q_{\lambda^*} | P_{g_{\theta_0}})$.

DEFINITION 9 (Solid tangent cone). For a set $A \subset R^d$ and $x \in A$, the solid tangent cone $T_x A$ is defined as $T_x A = \{y \in R^d : \exists\ y_m \text{ and } \lambda_m \text{ such that } y_m \to y, \lambda_m \to 0 \text{ as } m \to \infty, \text{ and } x + \lambda_m y_m \in A\}$.

If A has continuously differentiable boundary and $x \in \partial A$, then $T_x A$ consists of all the vectors in R^d that have negative inner products with the normal vector to ∂A at x pointing outside of A; if x is in the interior of A, then $T_x A = R^d$. We consider the following technical conditions for the main theorem in this section:

A5 The moment generating function $M_{g_{\theta_0}}$ is twice differentiable at $(\theta^\dagger, \gamma^\dagger, \lambda^\dagger)$.

A6 Under $Q^\dagger$, the solution to the Euler condition is unique, that is, the equation with respect to θ and γ

$$E^{Q^\dagger}\{y^\top \nabla_\theta \log g_\theta(X)\} \leq 0 \qquad \text{for all } y \in T_\theta \Theta,$$
$$E^{Q^\dagger}\{y^\top \nabla_\gamma \log h_\gamma(X)\} \leq 0 \qquad \text{for all } y \in T_\gamma \Gamma$$

has a unique solution $(\bar{\theta}, \bar{\gamma})$. In addition, $E^{Q^\dagger}\{\sup_{\theta \in \Theta} |\nabla^2_\theta \log g_\theta(X)|\} < \infty$ and $E^{Q^\dagger}\{\sup_{\gamma \in \Gamma} |\nabla^2_\gamma \log h_\gamma(X)|\} < \infty$. We also assume that under $Q^\dagger$, $\sqrt{n}(\hat{\theta} - \bar{\theta}) = O_{Q^\dagger}(1)$ and $\sqrt{n}(\hat{\gamma} - \bar{\gamma}) = O_{Q^\dagger}(1)$ as $n \to \infty$, where $\hat{\theta}$ and $\hat{\gamma}$ are maximum likelihood estimators $\hat{\theta} = \arg\sup_\theta \sum_{i=1}^n \log g_\theta(X_i)$ and $\hat{\gamma} = \arg\sup_\gamma \sum_{i=1}^n \log h_\gamma(X_i)$. A random sequence $a_n = O_{Q^\dagger}(1)$ if it is tight under measure $Q^\dagger$.

A7 We assume that g_{θ_0} does not belong to the closure of the family of distributions $\{h_\gamma : \gamma \in \Gamma\}$, that is, $\inf_{\gamma \in \Gamma} D(g_{\theta_0} \| h_\gamma) > 0$.

Assumption A6 requires $n^{-1/2}$ convergence of $\hat{\theta}$ and $\hat{\gamma}$ under $Q^\dagger$. It also requires the local maximum of the function $E^{Q^\dagger}\{\log g_\theta(X)\}$ and $E^{Q^\dagger}\{\log h_\gamma(X)\}$ to be unique. We elaborate the Euler condition for $\theta \in \text{int}(\Theta)$ and $\theta \in \partial\Theta$ separately. If $\theta \in int(\Theta)$, then $T_\theta \Theta = R^{d_g}$. The Euler condition is equivalent to

$E^{Q^\dagger}\{\nabla_\theta \log g_\theta(X)\} = 0$, which is the usual first-order condition for a local maximum. If $\theta \in \partial\Theta$, then the Euler condition requires that the directional derivative of $E^{Q^\dagger}\{\log g_\theta(X)\}$ along a vector pointing toward inside Θ is nonpositive. Assumption A7 guarantees that the probability $\lim_{n\to\infty} P_{g_{\theta_0}}(\mathrm{LR}_n > e^{nb}) = 0$ for some b. Unlike other conditions, it is not symmetric between h_γ and g_θ. This is because we are interested in computing the type I error under g_{θ_0}. In the case of type II error computation, we need to switch their roles.

THEOREM 10. *Under Assumptions* A2–A3 *and* A5–A7, *for each b satisfying* (16), *we have* $\log P_{g_{\theta_0}}(\mathrm{LR}_n > e^{nb}) \sim -n \times \rho^\dagger_{g_{\theta_0}}$, *where* $\rho^\dagger_{g_{\theta_0}} = -\log M^\dagger_{g_{\theta_0}}$ *and* $M^\dagger_{g_{\theta_0}}$ *is defined in* (17).

The following corollary illustrates how the result of the above theorem is applied to the special case in which $b = 0$, the most common choice of threshold in model selection between two parametric families.

COROLLARY 11. *Let* $C_n = \{\mathrm{LR}_n > 1\}$ *be the critical region corresponding to* $b = 0$ *in Theorem* 10. *For all g-family, h-family and the underlying true distribution f satisfying the assumptions of Theorem* 10, *we have* $\log P_f(C_n) \sim n \log(M^\dagger_f)$, *where* $\log(M^\dagger_f) < 0$ *is defined similarly as in* (17) *for f in the g-family.*

REMARK 12. Corollary 11 suggests that the probability of selecting the wrong model decays to zero exponentially fast if the true data generating distribution f stays away from the family that it does not belong to. In the case that f lies precisely on the boundary between the two families, the model selection or testing of hypothesis is an ill-posed problem.

Theorem 10 provides a means to approximate the type I and type II error probabilities for general parametric families. The above results are applicable to both cases that the two families are separated or not separated.

According to standard large deviations calculation for random walk, we have that for each $\theta \in \Theta$ and $\gamma \in \Gamma$, $\log P_{g_{\theta_0}}(\sum_{i=1}^n \log h_\gamma(X_i) - \log g_\theta(X_i) - nb > 0) \sim n \inf_\lambda \log M_{g_{\theta_0}}(\theta, \gamma, \lambda)$. This fact together with Theorem 10 imply that $\log P_{g_{\theta_0}}(\mathrm{LR}_n > 1) \sim \inf_\theta \sup_\gamma \log P_{g_{\theta_0}}(\sum_{i=1}^n \log h_\gamma(X_i) - \log g_\theta(X_i) > nb) \sim \log P_{g_{\theta_0}}(\sum_{i=1}^n \log h_{\gamma^\dagger}(X_i) - \log g_{\theta^\dagger}(X_i) > nb)$. The exponential decay rate of the error probabilities under g_{θ_0} is the same as the exponential decay rate of the probability that $h_{\gamma^\dagger}$ is preferred to $g_{\theta^\dagger}$. Furthermore, compared with the results in Lemma 3 that is the simple versus the composite case, the exponential decay rate obtained in Theorem 10 involves one additional minimization with respect to θ, accounting for the maximization of the likelihood over the null hypothesis; see (17).

One application of Theorem 10 is to compute the power function asymptotically. Consider the fixed type I error α and the critical region of the generalized likelihood ratio test is determined by the quantile of a χ^2 distribution, that is, $\{\mathrm{LR}_n > e^{\lambda_\alpha}\}$ where $2\lambda_\alpha$ is the $(1-\alpha)$th quantile of the χ^2 distribution. This correspond to choosing $b = o(1)$. For a given alternative distribution h_γ, one can compute the type II error probability asymptotically by means of Theorem 10 switching the role of the null and the alternative families. Thus, the power function can be computed asymptotically.

4.2. *Application to variable selection in generalized linear models.* We discuss the application of Theorem 10 to model selection for generalized linear models [McCullagh and Nelder (1989)]. Let Y_i be the response of the ith observation and $X^{(i)} = (X_{i1}, \dots, X_{ip})^T$ and $Z^{(i)} = (Z_{i1}, \dots, Z_{iq})^T$ be two sets of predictors, $i = 1, \dots, n$. Consider a generalized linear model with canonical link function and the true conditional distribution of Y_i is

$$g_i(y_i, \beta^0) = \exp\{(\beta^0)^T X^{(i)} y_i - b((\beta^0)^T X^{(i)}) + c(y_i)\}, \qquad i = 1, 2, \dots, n, \tag{19}$$

where $f(y) = e^{c(y)}$ is the base-line density, $b(\cdot)$ is the logarithm of the moment generating function, $\beta^0 = (\beta_1^0, \dots, \beta_p^0)^T$ is the vector of true regression coefficients, and X is the set of true predictors. Let the null hypothesis be $H_0 : g_i(y_i, \beta) = \exp\{\beta^T X^{(i)} y_i - b(\beta^T X^{(i)}) + c(y_i)\}$, $i = 1, 2, \dots, n$, and the alternative hypothesis be $H_1 : h_i(y_i, \gamma) = \exp\{\gamma^T Z^{(i)} y_i - b(\gamma^T Z^{(i)}) + c(y_i)\}$, $i = 1, 2, \dots, n$. We further assume that H_1 does not contain (19). Conditional on the covariates X and Z, we consider the asymptotic decay rate of the type I error probability $P_{\beta^0}(\mathrm{LR}_n \ge 1)$, where $\mathrm{LR}_n = \frac{\sup_\gamma \prod_{i=1}^n h_i(Y_i,\gamma)}{\sup_\beta \prod_{i=1}^n g_i(Y_i,\beta)}$ is the generalized likelihood ratio.

We present the construction of the rate function as follows. For each $\beta \in R^p$, $\gamma \in R^q$ and $\lambda \in R$, define $\tilde{\rho}_n(\beta, \gamma, \lambda) = \frac{1}{n}\sum_{i=1}^n \{\lambda[b(\gamma^T Z^{(i)}) - b(\beta^T X^{(i)})] + b((\beta^0)^T X^{(i)}) - b((\beta^0)^T X^{(i)} + \lambda(\gamma^T Z^{(i)} - \beta^T X^{(i)}))\}$. Taking derivative with respect to λ, we have $\frac{\partial}{\partial\lambda}\tilde{\rho}_n(\beta, \gamma, \lambda) = \frac{1}{n}\sum_{i=1}^n \{b(\gamma^T Z^{(i)}) - b(\beta^T X^{(i)}) - b'((\beta^0)^T X^{(i)} + \lambda(\gamma^T Z^{(i)} - \beta^T X^{(i)}))(\gamma^T Z^{(i)} - \beta^T X^{(i)})\}$. From the fact that $b(\cdot)$ is a convex function, we have $\limsup_{\lambda\to+\infty} \frac{\partial}{\partial\lambda}\tilde{\rho}_n(\beta, \gamma, \lambda) < 0$, provided that $\beta^T X^{(i)} \neq \gamma^T Z^{(i)}$ for some i. Define the set

$$B_n = \left\{\beta : \inf_\gamma \frac{\partial}{\partial\lambda}\tilde{\rho}_n(\beta, \gamma, 0) \ge 0\right\}.$$

Then for $\beta \in B_n$ and $\gamma \in R^q$, there is a $\lambda \ge 0$ such that $\frac{\partial}{\partial\lambda}\tilde{\rho}_n(\beta, \gamma, 0) = 0$. Since b is convex, $\beta^0 \in B_n$, implying that B_n is never empty. Now,

$$\begin{aligned}&\frac{\partial^2}{(\partial\lambda)^2}\tilde{\rho}_n(\beta, \gamma, \lambda)\\ &\qquad = -\frac{1}{n}\sum_{i=1}^n b''((\beta^0)^T X^{(i)} + \lambda(\gamma^T Z^{(i)} - \beta^T X^{(i)}))(\gamma^T Z^{(i)} - \beta^T X^{(i)})^2 < 0,\end{aligned}$$

if $\beta^T X^{(i)} \neq \gamma^T Z^{(i)}$ for some i. Therefore, there is a unique solution to the maximization $\sup_\lambda \widetilde{\rho}_n(\beta,\gamma,\lambda)$. We further consider the optimization:

$$\widetilde{\rho}_n^\dagger = \sup_{\beta \in B_n} \inf_{\gamma} \sup_{\lambda \geq 0} \widetilde{\rho}_n(\beta,\gamma,\lambda). \tag{20}$$

We consider the following technical conditions:

A8 For each n, the solution to (20) exists, denoted by $(\beta_n^\dagger, \gamma_n^\dagger, \lambda_n^\dagger)$. There exists a constant κ_1 such that $\|\beta_n^\dagger\| \leq \kappa_1$, $\|\gamma_n^\dagger\| \leq \kappa_1$, and $\lambda_n^\dagger \leq \kappa_1$ for all n. Here, $\|\cdot\|$ is the Euclidean norm.

A9 There exists a constant $\delta_1 > 0$ such that $\inf_\gamma \sup_\lambda \widetilde{\rho}_n(\beta^0,\gamma,\lambda) > \delta_1$ for all n.

A10 There exists a constant κ_2 such that $\|X^{(i)}\| \leq \kappa_2$ and $\|Z^{(i)}\| \leq \kappa_2$ for all i. Additionally, there exits $\delta_2 > 0$ such that for all n the smallest eigenvalue of $\frac{1}{n}\sum_{i=1}^n X^{(i)} X^{(i)T}$ is bounded below by δ_2.

A11 For any compact set $K \subset R$, $\inf_{u \in K} b''(u) > 0$. In addition, $b(\cdot)$ is four-time continuously differentiable.

Assumption A8 requires that the solution of the optimization (20) does not tend to infinity as n increases, which is a mild condition. In particular, if the Kullback–Leibler divergence $D(g_i(\cdot,\beta^0)|g_i(\cdot,\beta))$ tend to infinity uniformly for all i as $\|\beta\|$ goes to infinity, then B_n is a bounded subset of R^p and $\|\beta_n^\dagger\|$ is also bounded. Similar checkable sufficient conditions can be obtained for $\gamma_n^\dagger$ and $\lambda_n^\dagger$.

THEOREM 13. *Under Assumptions* A8–A11, *conditional on the covariates* $X^{(i)}$ *and* $Z^{(i)}$, $i = 1,\dots,n$, *we have* $\log P_{\beta^0}(\mathrm{LR}_n \geq 1) \sim -n \times \widetilde{\rho}_n^\dagger$, *where* $\widetilde{\rho}_n^\dagger$ *is defined in* (20).

For generalized linear models, the moment generating function of likelihood ratio is $E_{\beta^0}(\exp\{\lambda \sum_{i=1}^n [\log h_i(Y_i,\gamma) - \log g_i(Y_i,\beta)]\}) = e^{-n\widetilde{\rho}_n(\beta,\gamma,\lambda)}$. Therefore, $\widetilde{\rho}_n^\dagger$ is a natural generalization of $\rho_{g_{\theta_0}}^\dagger$ for the nonidentically distributed case.

Theorem 13 provides the asymptotic rate of selecting the wrong model by maximizing the likelihood. The asymptotic rate as a function of the true regression coefficients β^0 quantifies the strength of the signals. The larger the rate is, the easier it is to select the correct variables. The rate also depends on covariates. If Z is highly correlated with X, then the rate is small. Overall, the rate serves as an efficiency measure of selecting the true model from families that misspecifies the model.

5. Numerical examples.

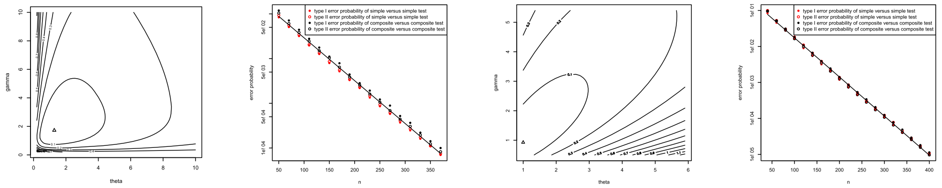
In this section, we present numerical examples to illustrate the asymptotic behavior of the maximal type I and type II error probabilities and the sample size tends to infinity. The first one is an example of continuous distributions and the second one is an example of discrete distributions. The third one is an example of linear regression models where the null hypotheses and alternative are not separated. In these examples, we compute the error probabilities using importance sampling corresponding to the change of measure in the proof with sufficiently large number of Monte Carlo replications to ensure that our estimates are sufficiently accurate.

EXAMPLE 14. Consider the case of lognormal versus exponential distributions. For $x > 0$, let $g_\theta(x) = \frac{1}{x(2\pi\theta)^{1/2}} e^{-\frac{(\log x)^2}{2\theta}}$, $\Theta = (0, +\infty)$ and $h_\gamma(x) = \frac{1}{\gamma} e^{-\frac{x}{\gamma}}$, $\Gamma = (0, +\infty)$ be the density functions of the lognormal and the exponential distributions.

For each θ and γ, we compute $\rho_{\theta\gamma}$ numerically. Figure 1(a) shows the contour plot of $\rho_{\theta,\gamma}$. The minimum of $\rho_{\theta\gamma}$ is 0.020 and is obtained at $(\theta^*, \gamma^*) = (1.28, 1.72)$. From the theoretical analysis, the maximal type I and type II error probabilities for the test decay at rate $e^{-n\rho_{\theta^*\gamma^*}}$.

Figure 1(b) is the plot of the maximal type I and type II error probabilities as a function of the sample size for the composite versus composite test $H_0 : f \in \{g_\theta; \theta \in \Theta\}$ against $H_1 : f \in \{h_\gamma; \gamma \in \Gamma\}$ and simple versus simple test $H_0 : f = g_{\theta_*}$ against $H_1 : f = h_{\gamma_*}$. We also fit a straight line to the logarithm of error probabilities against the sample sizes using least squares and the slope is -0.022. This confirms the theoretical findings. The error probabilities shown in Figure 1(b) range from 7×10^{-5} to 0.12 and the range for sample size is from 50 to 370.

EXAMPLE 15. We next consider the case of Poisson versus geometric distributions. Let $g_\theta(x) = \frac{e^{-\theta}\theta^x}{x!}$, $\Theta = [1, +\infty)$, and $h_\gamma(x) = \frac{\gamma^x}{(1+\gamma)^{x+1}}$, $\Gamma = [0.5, +\infty)$, for $x \in \mathbb{Z}^+$. The parameter γ is the failure to success odds. The minimum Chernoff



(a) Contour plot for $\rho_{\gamma,\theta}$ in Example 14. The triangle point indicates the minimum.

(b) Decay rate of type I and type II error probabilities (y-coordinate) as a function of sample size (x-coordinate) in Example 14.

(c) Contour plot for $\rho_{\gamma,\theta}$ in Example 15. The triangle point indicates the minimum.

(d) Maximal type I and type II error probabilities (y-coordinate) as a function of sample size (x-coordinate) in Example 15.

FIG. 1. *Contour plots of $\rho_{\gamma,\theta}$ and decay of error probabilities.*

index without constraint is attained at $\theta = \gamma = 0$ and $\rho_{00} = 0$. Thus we truncate the parameter spaces away from zero to separate the two families.

The Chernoff index $\rho_{\theta,\gamma}$ can be computed numerically and is minimized at $(\theta^*, \gamma^*) = (1, 0.93)$, with $\rho_{\theta^*,\gamma^*} = 0.023$. Figure 1(c) shows the contour plot of $\rho_{\theta,\gamma}$. Same as in the previous example, we compute the maximal type I and type II error probabilities of the composite versus composite test and simple versus simple test. Figure 1(d) shows the maximal type I and type II error probabilities as a function of the sample size. The error probabilities appeared in Figure 1(d) range from 1.0×10^{-4} to 0.10 with the sample sizes range from 40 to 400. We also fit a straight line to the logarithm of error probabilities against the sample sizes and the slope is -0.025. This numerical analysis confirms our theorems.

EXAMPLE 16. We now consider two linear regression models with different covariates, $H_0 : Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon_1$ against $H_1 : Y = \beta_1 X_1 + \zeta_1 Z_1 + \varepsilon_2$, where (X_1, X_2, Z_1) jointly follows the multivariate Gaussian distribution with the mean $(0, 0, 0)^T$ and the covariance matrix Σ. The random noises ε_1 and ε_2 follow the normal distributions $N(0, \sigma_1^2)$ and $N(0, \sigma_2^2)$, respectively, and are independent of (X_1, X_2, Z_1). We assume the true model to be $Y = \beta_1^0 X_1 + \beta_2^0 X_2 + \varepsilon$, with the following parameters:

$$\beta_1^0 = 1, \qquad \beta_2^0 = 2, \qquad \varepsilon \sim N(0, 1), \quad \text{and} \quad \Sigma = \begin{bmatrix} 1 & 0.1 & 0.1 \\ 0.1 & 1 & 0.1 \\ 0.1 & 0.1 & 1 \end{bmatrix}.$$

Let $(X_{i1}, X_{i2}, Z_{i1}, Y_i)^T$ be i.i.d. copies of (X_1, X_2, Z_1, Y) generated under the true model, for $i = 1, \dots, n$. Let $\theta = (\beta_1, \beta_2)$ and $\gamma = (\beta_1, \zeta_1)$ be the regression coefficients for the null and the alternative hypotheses, respectively. The maximum likelihood estimators for θ and γ are the least square estimators $\hat{\theta} = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \tilde{Y}$ and $\hat{\gamma} = (\tilde{Z}^\top \tilde{Z})^{-1} \tilde{Z}^\top \tilde{Y}$, where $\tilde{Y} = (Y_1, \dots, Y_n)^T$ and $\tilde{X}$ and $\tilde{Z}$ are $n \times 2$ matrices with respective row vectors (X_{i1}, X_{i2}) and (X_{i1}, Z_{i1}), $i = 1, \dots, n$. We consider the error probability that the maximized log-likelihood of H_0 is smaller than that of H_1, equivalently, the residual sum of squares under H_0 is larger than that under H_1, that is, $P_{\beta^0,\Sigma}(\|\tilde{Y} - \tilde{X}\hat{\theta}\|^2 > \|\tilde{Y} - \tilde{Z}\hat{\gamma}\|^2)$. From the theoretical analysis, the above probability decays at rate $e^{-n\rho^\dagger_{g_{\theta_0}}}$ as $n \to \infty$, where the definition of $\rho^\dagger_{g_{\theta_0}}$ is given in Theorem 10. We solve the optimization problem (17) numerically and obtain $\rho^\dagger_{g_{\theta_0}} = 0.45$. We simulate the probability $P_{\beta^0,\Sigma}(\|\tilde{Y} - \tilde{X}\hat{\theta}\|^2 > \|\tilde{Y} - \tilde{Z}\hat{\gamma}\|^2)$ of different sample sizes and fit straight lines for $\log P_{\beta^0,\Sigma}(\|\tilde{Y} - \tilde{X}\hat{\theta}\|^2 > \|\tilde{Y} - \tilde{Z}\hat{\gamma}\|^2)$ against n using least square. The range of error probabilities is from 10^{-4} to 0.25 with sample size from 3 to 18. The range of error probabilities is from 1.2×10^{-8} to 4.0×10^{-6} with the sample size from 24 to 36. The fitted slope in the former case is -0.52 and the fitted slope in the latter case is -0.47. This confirms our theoretical results.

6. Concluding remarks.

The generalized likelihood ratio test of separate parametric families that was put forth by Cox in his two seminal papers has received a great deal of attention in the statistics and econometrics literature. The present investigation takes the viewpoint of an early work by Chernoff (1952) where testing a simple null versus a simple alternative is considered. By imposing that the two types of error probabilities decay at the same rate, we extend the Chernoff index to the case of the Cox test.

When the two families are not completely separated, we derive the rate of exponential decay (type I) error probability of the Cox test under the true model-parameter specification. This formulation covers more broadly the model selection problem, including selection of covariates among competing sets of covariates for generalized linear models.

Our results are under the basic assumption that the data come from one of the parametric families under consideration. It is often the case that none is the true model. The finite sample deviation and large deviations bound of quasi-MLE under model misspecification is discussed in Spokoiny (2012). Similar results for semiparametric models are presented in Andresen and Spokoiny (2014). An interesting future development is the exact large deviations rate under this setting.

An initial motivation that led to the Cox formulation of the problem comes from the survival analysis where different models are used to fit failure time data. The econometrics literature also contains much subsequent development. Semiparametric models that contain infinite dimensional nuisance parameters are widely used in both econometrics and survival analysis. It would be of interest to develop parallel results for testing separate semiparametric models.

7. Proof of Lemma 3.

Throughout the proof, we adopt the following notation $a_n \cong b_n$ if $\log a_n \sim \log b_n$. Also, we use $E(X; A) = E\{XI(\omega \in A)\}$ to denote expectation of X on the set A. We define the log-likelihood ratio as $l_\gamma(x) = \log h_\gamma(x) - \log g(x)$. The generalized log-likelihood ratio statistic is defined as $l = \sup_\gamma \sum_{i=1}^n l_\gamma^i$ where $l_\gamma^i = l_\gamma(X_i)$. The generalized likelihood ratio test admits the rejection region $C_\lambda = \{e^l > \lambda\}$. We consider the case that $\lambda = 1$ and show that for this particular choice of λ the maximal type I and type II error probabilities decay exponentially fast with the same rate. We let $\gamma_* = \arg\inf \rho_\gamma$, and thus $\rho = \rho_{\gamma_*}$.

Based on Chernoff's calculation of large deviations for the log-likelihood ratio statistic, we proceed to the calculation of the type I error probability $P_g(l > 0) = P_g(\sup_\gamma \sum_{i=1}^n l_\gamma^i > 0)$. We now provide an approximation to it by establishing a lower bound and an upper bound. We start with the lower bound by noticing that $P_g(\sup_\gamma \sum_{i=1}^n l_\gamma^i > 0) \geq \sup_\gamma P_g(\sum_{i=1}^n l_\gamma^i > 0)$. This is further bounded bounded below by $e^{-\{1+o(1)\}n\rho}$, according to Proposition 1. where $\rho = \min \rho_\gamma$. For the upper bound, we split the probability, with some $\beta > 0$,

$$P_g\left(\sup_\gamma \sum_{i=1}^n l_\gamma^i > 0\right) \le P_g\left(\sup_\gamma \sum_{i=1}^n l_\gamma^i > 0, \sup_\gamma \left|\sum_{i=1}^n \nabla l_\gamma^i\right| < e^{n^{1-\beta}}\right) \tag{21}$$
$$+ P_g\left(\sup_\gamma \left|\sum_{i=1}^n \nabla l_\gamma^i\right| \ge e^{n^{1-\beta}}\right).$$

The first term on the right-hand side is bounded by Lemma 17.

LEMMA 17. *Consider a random function $\eta_n(\theta)$ living on a d-dimensional compact domain $\theta \in D$, where n is an asymptotic parameter that will be send to infinity. Suppose that $\eta_n(\theta)$ is almost surely differentiable with respect to θ and for each θ, there exists a rate $\rho(\theta)$ such that $P\{\eta_n(\theta) > \zeta_n\} \cong e^{-n\rho(\theta)}$ for all $\zeta_n/n \to 0$ as $n \to \infty$. This convergence is uniform in θ. Then we have the following approximation for all $\beta > 0$*:

$$\liminf_{n\to\infty} -\frac{1}{n} \log P\left\{\sup_{\theta\in D} \eta_n(\theta) > 0, \sup_{\theta\in D} |\nabla \eta_n(\theta)| < e^{n^{1-\beta}}\right\} \ge \min_\theta \rho(\theta).$$

The proof of this lemma is in the supplementary materials. It employs a change of measure defined on the continuous sample space. Similar techniques are used for the extreme analysis of stochastic systems driven by Gaussian processes [Liu and Xu (2012), Adler, Blanchet and Liu (2012), Li and Liu (2015), Liu, Lu and Zhou (2015), Liu and Xu (2014a, 2014b)].

With the aid of Proposition 1, we have that the random function $\sum_{i=1}^n l_\gamma^i$ satisfies the assumption in Lemma 17 with $\rho(\gamma) = \rho_\gamma$. Then the first term in (21) is bounded from the above by $e^{-\{1+o(1)\}n\rho}$. For the second term in (21), according to condition A3, we choose β sufficiently small such that

$$P_g\left(\sup_\gamma \left|\sum_{i=1}^n \nabla l_\gamma^i\right| \ge e^{n^{1-\beta}}\right) \le n P_g\left(\sup_\gamma |\nabla l_\gamma^i| > n^{-1} e^{n^{1-\beta}}\right) = o\left(e^{-n\rho}\right).$$

Thus, we obtain an upper bound $P_g(\sup_\gamma \sum_{i=1}^n l_\gamma^i > 0) \le e^{-n\{\rho+o(1)\}}$. Then the type I error probability is approximated by

$$e^{-n\rho} \cong \sup_\gamma P_g\left(\sum_{i=1}^n l_\gamma^i > 0\right) \le P_g\left(\sup_\gamma \sum_{i=1}^n l_\gamma^i > 0\right) \le e^{-n\{\rho+o(1)\}}. \tag{22}$$

We now consider the type II error probability $\alpha_2 = \sup_\gamma P_{h_\gamma}(l < 0)$. For each γ, note that $P_{h_\gamma}(l < 0) = P_{h_\gamma}(\sup_{\gamma_1} \sum_{i=1}^n l_{\gamma_1}^i < 0) \le P_{h_\gamma}(\sum_{i=1}^n l_\gamma^i < 0)$. Note that the right-hand side is the type II error probability of the likelihood ratio test. According to Chernoff's calculation, we have that $P_{h_\gamma}(l < 0) \le P_{h_\gamma}(\sum_{i=1}^n l_\gamma^i < 0) \cong e^{-n\rho_\gamma}$

for all γ. We take maximum with respect to γ on both sides and obtain that

$$\sup_{\gamma} P_{h_\gamma}(l < 0) \le \sup_{\gamma} P_g\left(\sum_{i=1}^{n} l_\gamma^i > 0\right) \cong e^{-n \min_\gamma \rho_\gamma}. \tag{23}$$

Thus, the maximal type II error probability has an asymptotic upper bound that decays at the rate of the Chernoff index.

In what follows, we show that this asymptotic upper bound is attained. Choose λ_n so that $P_g(\sup_\gamma \sum_{i=1}^n l_\gamma^i > 0) = P_g(\sum_{i=1}^n l_{\gamma_*}^i > n\lambda_n)$. Note that g is fixed, λ_n depends on g, and the probabilities on both sides of the above identity decay at the rate $e^{-n\rho}$. Together with the continuity of the large deviations rate function, it must be true that $\lambda_n \to 0-$. We apply Neyman–Pearson lemma to the simple null $H_0 : f = g$ versus simple alternative $H_1 : f = h_{\gamma*}$. Note that $\{\sum_{i=1}^n l_{\gamma_*}^i > n\lambda_n\}$ is a uniformly most powerful test and $\{\sup_\gamma \sum_{i=1}^n l_\gamma^i > 0\}$ is a test with the same type I error probability. Then we have that

$$P_{h_{\gamma*}}\left(\sup_{\gamma} \sum_{i=1}^{n} l_\gamma^i < 0\right) \ge P_{h_{\gamma*}}\left(\sum_{i=1}^{n} l_{\gamma_*}^i < n\lambda_n\right). \tag{24}$$

That is, the type II error probability of the generalized likelihood ratio test must be greater than that of the likelihood ratio test under the simple alternative $h_{\gamma*}$. Note that $\lambda_n \to 0-$. Thanks to the the continuity of the large deviations rate function, we have that

$$P_{h_{\gamma*}}\left(\sum_{i=1}^{n} l_{\gamma*}^i < n\lambda_n\right) \cong P_{h_{\gamma*}}\left(\sum_{i=1}^{n} l_{\gamma*}^i < 0\right) \cong e^{-n\rho}. \tag{25}$$

Combining (23), (24) and (25), we have $\sup_\gamma P_{h_\gamma}(l < 0) \cong e^{-n\rho}$ concluding the proof.

8. Proof of Theorem 4. The one-to-one log-likelihood ratio is $l_{\theta\gamma}(x) = \log h_\gamma(x) - \log g_\theta(x)$. The generalized log-likelihood ratio statistic is

$$l = \sup_{\gamma} \sum_{i=1}^{n} \log h_\gamma(X_i) - \sup_{\theta} \sum_{i=1}^{n} \log g_\theta(X_i) = \inf_{\theta} \sup_{\gamma} \sum_{i=1}^{n} l_{\theta\gamma}^i,$$

where $l_{\theta\gamma}^i = l_{\theta\gamma}(X_i)$ and the rejection region is $C_\lambda = \{e^l > \lambda\}$. We define that $\gamma(\theta) = \arg\inf_\gamma \rho_{\theta\gamma}$, and $\theta(\gamma) = \arg\inf_\theta \rho_{\theta\gamma}$, and $(\theta_*, \gamma_*) = \arg\inf_{\theta,\gamma} \rho_{\theta\gamma}$. Note that the null and the alternative are now symmetric, thus we only need to consider one of the two types of error probabilities. We consider the type II error probability. We now define $k_\theta = \sup_\gamma \sum_{i=1}^n l_{\theta\gamma}^i$. For each given θ and γ, we have a simple upper bound

$$P_{h_\gamma}(k_\theta < 0) \le P_{h_\gamma}\left(\sum_{i=1}^{n} l_{\theta\gamma}^i < 0\right) \cong e^{-n\rho_{\theta\gamma}}. \tag{26}$$

We now proceed to the type II error probability when h_γ is the true distribution, that is,

$$P_{h_\gamma}\Big(\inf_\theta k_\theta < 0\Big) \le P_{h_\gamma}\Big(\inf_\theta k_\theta < 0, \sup_\theta |\nabla k_\theta| < e^{n^{1-\beta}}\Big) + P_{h_\gamma}\Big(\sup_\theta |\nabla k_\theta| \ge e^{n^{1-\beta}}\Big).$$

For the first term on the right-hand side, we have, in view of Lemma 17 and (26), $P_{h_\gamma}(\inf_\theta k_\theta < 0, \sup_\theta |\nabla k_\theta| < e^{n^{1-\beta}}) \le e^{-n\{\inf_\theta \rho_{\theta\gamma}+o(1)\}}$. For the second term, according to condition A3, $P_{h_\gamma}\{\sup_\theta |\nabla(\sup_\gamma \sum_{i=1}^n l^i_{\theta\gamma})| \ge e^{n^{1-\beta}}\} = o(e^{-n\rho})$. Thus, $P_{h_\gamma}(\inf_\theta k_\theta < 0) = P_{h_\gamma}(l < 0) \le e^{-n\{\inf_\theta \rho_{\theta\gamma}+o(1)\}}$, which provides an upper bound for the type II error probability $\sup_\gamma P_{h_\gamma}(l < 0) \le e^{-n\{\inf_{\theta,\gamma} \rho_{\theta\gamma}+o(1)\}}$. We now provide a lower bound. For a given θ and $\gamma(\theta) = \arg\inf_\gamma \rho_{\theta\gamma}$, applying proof of Lemma 3 for the type II error probability by considering $H_0 : f = g_\theta$ and $H_1 : f \in \{h_\gamma : \gamma \in \Gamma\}$, we have $P_{h_{\gamma(\theta)}}(k_\theta < 0) \cong e^{-n\rho_{\theta\gamma(\theta)}}$. Thus, $P_{h_{\gamma(\theta)}}(\inf_\theta k_\theta < 0) \ge P_{h_{\gamma(\theta)}}(k_\theta < 0) \cong e^{-n\rho_{\theta\gamma(\theta)}}$. By setting $\theta = \theta_*$ in the above asymptotic identity, we conclude the proof.

9. Proof of Theorem 10. The proof of the theorem consists of establishing upper and lower bounds for the probability $P_{g_{\theta_0}}(\mathrm{LR}_n > e^{nb}) = P_{g_{\theta_0}}(\sup_{\gamma\in\Gamma} \inf_{\theta\in\Theta} \sum_{i=1}^n [\log h_\gamma(X_i) - \log g_\theta(X_i)] > nb)$.

Upper bound. The event $\{\sup_{\gamma\in\Gamma} \inf_{\theta\in\Theta} \sum_{i=1}^n \log h_\gamma(X_i) - \log g_\theta(X_i) > nb\}$ implies $\{\sup_{\gamma\in\Gamma} \sum_{i=1}^n \log h_\gamma(X_i) - \log g_{\theta^\dagger}(X_i) > nb\}$. Therefore, we have an upper bound:

$$P_{g_{\theta_0}}(\mathrm{LR}_n > e^{nb}) \le P_{g_{\theta_0}}\Bigg(\sup_\gamma \sum_{i=1}^n \log h_\gamma(X_i) - \log g_{\theta^\dagger}(X_i) > nb\Bigg). \tag{27}$$

We split the probability

$$\begin{aligned} &P_{g_{\theta_0}}\Bigg(\sup_\gamma \sum_{i=1}^n [\log h_\gamma(X_i) - \log g_{\theta^\dagger}(X_i)] > nb\Bigg) \\ &\quad\le P_{g_{\theta_0}}\Bigg(\sup_\gamma \sum_{i=1}^n [\log h_\gamma(X_i) - \log g_{\theta^\dagger}(X_i)] > nb, \\ &\qquad\quad \sup_\gamma \Bigg|\sum_{i=1}^n \nabla \log h_\gamma(X_i)\Bigg| < e^{n^{1-\beta}}\Bigg) \\ &\qquad + P_{g_{\theta_0}}\Bigg(\sup_\gamma \sum_{i=1}^n |\nabla_\gamma \log h_\gamma(X_i)| \ge e^{n^{1-\beta}}\Bigg). \end{aligned} \tag{28}$$

We establish upper bounds of the first and second terms in (28) separately. For the first term, let $\eta_n(\gamma) = \sum_{i=1}^n [\log h_\gamma(X_i) - \log g_{\theta^\dagger}(X_i)] - nb$. For each γ, the exponential decay rate of the probability

$$\log P_{g_{\theta_0}}\bigl(\eta_n(\gamma)\geq 0\bigr)\leq n\log\inf_{\lambda\geq 0}M_{g_{\theta_0}}(\lambda,\gamma,\theta^\dagger) \tag{29}$$

is established with standard large deviations calculation. By Lemma 17 and (29), the first term in (28) is bounded by $\sup_\gamma \inf_{\lambda\geq 0}\{M_{g_{\theta_0}}(\theta^\dagger,\lambda,\gamma)\}^{(1+o(1))n} = e^{-(1+o(1))n\rho^\dagger_{g_{\theta_0}}}$. For the second term, according to Assumption A3,

$$\begin{aligned}&P_{g_{\theta_0}}\Biggl(\sup_\gamma\sum_{i=1}^n\bigl|\nabla_\gamma\log h_\gamma(X_i)\bigr|\geq e^{n^{1-\beta}}\Biggr)\\&\quad\leq nP_{g_{\theta_0}}\Bigl(\sup_\gamma\bigl|\nabla_\gamma\log h_\gamma(X_i)\bigr|>n^{-1}e^{n^{1-\beta}}\Bigr)=o\bigl(e^{-n\rho^\dagger_{g_{\theta_0}}}\bigr).\end{aligned}$$

Combining the analyses for both the first and the second terms, we arrive at an upper bound $P_{g_{\theta_0}}(\mathrm{LR}_n>e^{nb})\leq e^{-(1+o(1))n\rho^\dagger_{g_{\theta_0}}}$.

Lower bound. Recall $\frac{dQ^\dagger}{dP_{g_{\theta_0}}}=\exp\{\lambda^\dagger(\log h_{\gamma^\dagger}(X)-\log g_{\theta^\dagger}(X))-nb\}/M^\dagger_{g_{\theta_0}}$. Then the probability can be written as

$$P_{g_{\theta_0}}(\mathrm{LR}_n>e^{nb})=E^{Q^\dagger}\Biggl\{\frac{dP_{g_{\theta_0}}}{dQ^\dagger};\sum_{i=1}^n\bigl[\log h_{\hat\gamma}(X_i)-\log g_{\hat\theta}(X_i)\bigr]>nb\Biggr\},$$

where $\hat\gamma$ and $\hat\theta$ are the maximum likelihood estimators for the h-family and the g-family, respectively. According to the definition of $Q^\dagger$, the above display is equal to

$$\begin{aligned}&e^{-n\rho^\dagger_{g_{\theta_0}}}E^{Q^\dagger}\Biggl\{e^{-\lambda^\dagger[\sum_{i=1}^n\log h_{\gamma^\dagger}(X_i)-\log g_{\theta^\dagger}(X_i)-nb]};\\&\qquad\times\sum_{i=1}^n\log h_{\hat\gamma}(X_i)-\log g_{\hat\theta}(X_i)>nb\Biggr\},\end{aligned}\tag{30}$$

where $\rho^\dagger_{g_{\theta_0}}=-\log M^\dagger_{g_{\theta_0}}$. We now establish a lower bound for

$$I\triangleq E^{Q^\dagger}\Biggl\{e^{-\lambda^\dagger[\sum_{i=1}^n\log h_{\gamma^\dagger}(X_i)-\log g_{\theta^\dagger}(X_i)-nb]};\sum_{i=1}^n\log h_{\hat\gamma}(X_i)-\log g_{\hat\theta}(X_i)>nb\Biggr\}.$$

Because $e^{-\lambda^\dagger[\sum_{i=1}^n\log h_{\gamma^\dagger}(X_i)-\log g_{\theta^\dagger}(X_i)-nb]}$ is positive, I is lower bounded by

$$\begin{aligned}&E^{Q^\dagger}\Biggl\{e^{-\lambda^\dagger[\sum_{i=1}^n\log h_{\gamma^\dagger}(X_i)-\log g_{\theta^\dagger}(X_i)-nb]};\\&\qquad\sum_{i=1}^n\log h_{\hat\gamma}(X_i)-\log g_{\hat\theta}(X_i)>nb,E_1\Biggr\},\end{aligned}\tag{31}$$

where $E_1 = \{|\sum_{i=1}^n \log h_{\gamma^\dagger}(X_i) - \log g_{\theta^\dagger}(X_i) - nb| \le \sqrt{n}\}$. On the set E_1, we have an inequality of the integrand $e^{-\lambda^\dagger[\sum_{i=1}^n \log h_{\gamma^\dagger}(X_i) - \log g_{\theta^\dagger}(X_i) - nb]} \ge e^{-\lambda^\dagger\sqrt{n}}$. We plug this inequality back to (31) and obtain a lower bound for

$$I \ge e^{-\lambda^\dagger\sqrt{n}} Q^\dagger\left(\left\{\sum_{i=1}^n \log h_{\hat{\gamma}}(X_i) - \log g_{\hat{\theta}}(X_i) > nb\right\} \cap E_1\right). \tag{32}$$

For the rest of the proof, we develop a lower bound for the probability $Q^\dagger(\{\sum_{i=1}^n \log h_{\hat{\gamma}}(X_i) - \log g_{\hat{\theta}}(X_i) > nb\} \cap E_1)$. The maximum likelihood estimator $\hat{\gamma}$ satisfies the inequality

$$\sum_{i=1}^n \{\log h_{\hat{\gamma}}(X_i) - \log h_{\gamma^\dagger}(X_i)\} \ge 0. \tag{33}$$

Furthermore, with the aid of Rolle's theorem, there exists $\tilde{\theta}$ such that

$$\begin{aligned} &\sum_{i=1}^n \{\log g_{\hat{\theta}}(X_i) - \log g_{\theta^\dagger}(X_i)\} \\ &\quad = (\hat{\theta} - \theta^\dagger)^\top \sum_{i=1}^n \nabla_\theta \log g_{\theta^\dagger}(X_i) + \frac{1}{2}(\hat{\theta} - \theta^\dagger)^\top \sum_{i=1}^n \nabla_\theta^2 g_{\tilde{\theta}}(X_i)(\hat{\theta} - \theta^\dagger), \end{aligned} \tag{34}$$

where "∇_θ^2" denotes the Hessian matrices with respect to θ and "$\cdot$" denotes the inner product between vectors. (33) and (34) together give

$$\begin{aligned} &\sum_{i=1}^n \{\log h_{\hat{\gamma}}(X_i) - \log g_{\hat{\theta}}(X_i)\} - \sum_{i=1}^n \{\log h_{\gamma^\dagger}(X_i) - g_{\theta^\dagger}(X_i)\} \\ &\quad \ge -(\hat{\theta} - \theta^\dagger)^\top \sum_{i=1}^n \nabla_\theta \log g_{\theta^\dagger}(X_i) - \frac{1}{2}(\hat{\theta} - \theta^\dagger)^\top \sum_{i=1}^n \nabla_\theta^2 g_{\tilde{\theta}}(X_i)(\hat{\theta} - \theta^\dagger). \end{aligned} \tag{35}$$

We define $E_2 = \{(\hat{\theta} - \theta^\dagger)^\top \sum_{i=1}^n \nabla_\theta \log g_{\theta^\dagger}(X_i) \le \frac{\sqrt{n}}{4}\}$, $E_3 = \{\frac{1}{2}|\hat{\theta} - \theta^\dagger|^2 \times \sup_\theta \sum_{i=1}^n |\nabla_\theta^2 \log g_\theta(X_i)| \le \frac{\sqrt{n}}{4}\}$, and $E_4 = \{\frac{\sqrt{n}}{2} < \sum_{i=1}^n [\log h_{\gamma^\dagger}(X_i) - \log g_{\theta^\dagger}(X_i)] - nb \le \sqrt{n}\}$. Based on (35), we have that $(E_2 \cap E_3 \cap E_4) \subset \{\sum_{i=1}^n \log h_{\hat{\gamma}}(X_i) - \log g_{\hat{\theta}}(X_i) > nb\} \cap E_1$. We insert this to (31), and obtain that

$$I \ge e^{-\lambda^\dagger\sqrt{n}} Q^\dagger(E_2 \cap E_3 \cap E_4) \ge e^{-\lambda^\dagger\sqrt{n}}\{Q^\dagger(E_4) - Q^\dagger(E_2^c) - Q^\dagger(E_3^c)\}. \tag{36}$$

For the rest of the proof, we develop upper bounds for $Q^\dagger(E_2^c)$ and $Q^\dagger(E_3^c)$ and a lower bound for $Q^\dagger(E_4)$. Because $\lambda^\dagger = \arg\inf_\lambda M_{g_{\theta_0}}(\theta^\dagger, \gamma^\dagger, \lambda)$, we have $\frac{\partial}{\partial\lambda} M_{g_{\theta_0}}(\theta^\dagger, \gamma^\dagger, \lambda^\dagger) = 0$. Consequently,

$$E^{Q^\dagger}(\log h_{\gamma^\dagger}(X) - \log g_{\theta^\dagger}(X) - b) = (M_{g_{\theta_0}}^\dagger)^{-1} \frac{\partial}{\partial\lambda} M_{g_{\theta_0}}(\theta^\dagger, \gamma^\dagger, \lambda^\dagger) = 0.$$

By the central limit theorem, there exists $\varepsilon_0 > 0$ such that $\liminf Q^{\dagger}(E_4) > \varepsilon_0$, a lower bound for $Q^{\dagger}(E_4)$. We need the following lemma, whose proof is given in the supplementary materials.

LEMMA 18. *Under settings of Theorem 10, we have $\gamma^{\dagger} = \bar{\gamma}$ and $\theta^{\dagger} = \bar{\theta}$.*

We now proceed to an upper bound of $Q^{\dagger}(E_2^c)$. We split the sum

$$
\begin{aligned}
&(\hat{\theta} - \theta^{\dagger})^{\top} \sum_{i=1}^{n} \nabla_{\theta} \log g_{\theta^{\dagger}}(X_i) \\
&\quad = (\hat{\theta} - \theta^{\dagger})^{\top} \sum_{i=1}^{n} [\nabla_{\theta} \log g_{\theta^{\dagger}}(X_i) - E^{Q^{\dagger}} \nabla_{\theta} g_{\theta^{\dagger}}(X_i)] \\
&\qquad + n(\hat{\theta} - \theta^{\dagger})^{\top} E^{Q^{\dagger}} \nabla_{\theta} g_{\theta^{\dagger}}(X).
\end{aligned} \tag{37}
$$

Note that $\hat{\theta} \in T_{\theta^{\dagger}}\Theta$. According to Assumption A6 and Lemma 18, we have that $(\hat{\theta} - \theta^{\dagger})^{\top} E^{Q^{\dagger}} \{\nabla_{\theta} g_{\theta^{\dagger}}(X)\} \le 0$. Therefore, (37) implies

$$
\begin{aligned}
&(\hat{\theta} - \theta^{\dagger})^{\top} \sum_{i=1}^{n} \nabla_{\theta} \log g_{\theta^{\dagger}}(X_i) \\
&\quad \le (\hat{\theta} - \theta^{\dagger})^{\top} \sum_{i=1}^{n} [\nabla_{\theta} \log g_{\theta^{\dagger}}(X_i) - E^{Q^{\dagger}} \nabla_{\theta} g_{\theta^{\dagger}}(X_i)].
\end{aligned} \tag{38}
$$

By the Chebyshev inequality and the fact $E(|\nabla_{\theta} \log g_{\theta^{\dagger}}(X)|^2) < \infty$, we have $n^{-\frac{3}{4}} \sum_{i=1}^{n} [\nabla_{\theta} \log g_{\theta^{\dagger}}(X_i) - E^{Q^{\dagger}} \nabla_{\theta} g_{\theta^{\dagger}}(X_i)] \to 0$ in probability $Q^{\dagger}$. By Slutsky's theorem and $\sqrt{n}(\hat{\theta} - \theta^{\dagger}) = O_{Q^{\dagger}}(1)$, we have

$$
\sqrt{n}(\hat{\theta} - \theta^{\dagger})^{\top} n^{-\frac{3}{4}} \sum_{i=1}^{n} \nabla_{\theta} \log g_{\theta^{\dagger}}(X_i) \to 0 \qquad \text{in probability } Q^{\dagger}.
$$

Thus, $\lim_{n\to\infty} Q^{\dagger}\{(\hat{\theta} - \theta^{\dagger})^{\top} \sum_{i=1}^{n} [\nabla_{\theta} \log g_{\theta^{\dagger}}(X_i) - E^{Q^{\dagger}} \nabla_{\theta} g_{\theta^{\dagger}}(X_i)] > \frac{\sqrt{n}}{4}\} = 0$, which, together with (38), implies

$$
\lim_{n\to\infty} Q^{\dagger}\left((\hat{\theta} - \theta^{\dagger})^{\top} \sum_{i=1}^{n} \nabla_{\theta} \log g_{\theta^{\dagger}}(X_i) > \frac{\sqrt{n}}{4} \right) = 0.
$$

Hence, $Q^{\dagger}(E_2^c) \to 0$ as $n \to \infty$. We establish an upper bound of $Q^{\dagger}(E_3^c)$ using a similar method. Again by the Chebyshev inequality we have

$$
n^{-\frac{5}{4}} \sum_{i=1}^{n} \sup_{\theta} |\nabla_{\theta}^{2} \log g_{\theta}(X_i)| \to 0 \qquad \text{in probability } Q^{\dagger}.
$$

According to Slutsky's theorem and $\sqrt{n}(\hat{\theta}-\theta^{\dagger}) = O_{Q^{\dagger}}(1)$, we have

$$n|\hat{\theta}-\theta^{\dagger}|^2 \times n^{-\frac{5}{4}} \sum_{i=1}^{n} \sup_{\theta} |\nabla_{\theta}^2 \log g_{\theta}(X_i)| \xrightarrow{d} 0.$$

Consequently,

$$\begin{aligned}\lim_{n\to\infty} Q^{\dagger}\left(|\hat{\theta}-\theta^{\dagger}|^2 \sup_{\theta} \sum_{i=1}^{n} |\nabla_{\theta}^2 \log g_{\theta}(X_i)| > \frac{\sqrt{n}}{4}\right) \\ \le \lim_{n\to\infty} Q^{\dagger}\left(|\hat{\theta}-\theta^{\dagger}|^2 \sum_{i=1}^{n} \sup_{\theta} |\nabla_{\theta}^2 \log g_{\theta}(X_i)| > \frac{\sqrt{n}}{4}\right) = 0.\end{aligned}$$

Therefore, $Q^{\dagger}(E_3^c) \to 0$ as $n \to \infty$. We combine the results for $Q^{\dagger}(E_2^c)$, $Q^{\dagger}(E_3^c)$, $Q^{\dagger}(E_4)$ and (36) to get $I \ge \frac{\varepsilon_0}{2} e^{-\lambda^{\dagger}\sqrt{n}}$ for n sufficiently large. Combining this with (30), we arrive at $P_{g_{\theta_0}}(\mathrm{LR}_n > e^{nb}) \ge e^{-n(1+o(1))\rho^{\dagger}_{g_{\theta_0}}}$. We complete the proof by combining the lower bound and upper bound for $P_{g_{\theta_0}}(\mathrm{LR}_n > 1)$.

10. Proof of Theorem 13.

The proof is similar to that of Theorem 10. Throughout the proof, we will use κ as a generic notation to denote large and not-so-important constants whose value may vary from place to place. Similarly, we use ε as a generic notation for small positive constants. The proof of the theorem consists of establishing upper and lower bounds for the probability $P_{\beta^0}(\mathrm{LR}_n \ge 1) = P_{\beta^0}(\sup_{\gamma} \inf_{\beta} \sum_{i=1}^{n} [\log h_i(Y_i, \gamma) - \log g_i(Y_i, \beta)] \ge 0)$.

Upper bound. Similar to (27),

$$P_{\beta^0}(\mathrm{LR}_n \ge 1) \le P_{\beta^0}\left\{\sup_{\gamma} \sum_{i=1}^{n} [\log h_i(Y_i, \gamma) - \log g_i(Y_i, \beta_n^{\dagger})] \ge 0\right\}.$$

By definitions of h_i and g_i, we have

$$\begin{aligned}&\log h_i(Y_i, \gamma) - \log g_i(Y_i, \beta_n^{\dagger}) \\ &\quad = [\gamma^T Z^{(i)} Y_i - b(\gamma^T Z^{(i)})] - [\beta_n^{\dagger T} X^{(i)} Y_i - b(\beta_n^{\dagger T} X^{(i)})].\end{aligned}$$

Consequently, we have

$$P_{\beta^0}(\mathrm{LR}_n \ge 1) \le P_{\beta^0}\left(\left(\frac{1}{n}\sum_{i=1}^{n} Z^{(i)} Y_i, \frac{1}{n}\sum_{i=1}^{n} X^{(i)} Y_i\right) \in A_n\right), \tag{39}$$

where $A_n = \{(s_1, s_2) \in R^p \times R^q : \sup_\gamma[\gamma^T s_2 - \frac{1}{n}\sum_{i=1}^n b(\gamma^T Z^{(i)})] \geq [\beta_n^{\dagger T} s_1 - \frac{1}{n}\sum_{i=1}^n b(\beta_n^{\dagger T} X^{(i)})]\}$. We consider the change of measure

$$\begin{aligned}\frac{dQ^\dagger}{dP} = \exp\Bigg\{&\lambda_n^\dagger \sum_{i=1}^n (\gamma_n^{\dagger T} Z^{(i)} Y_i - \beta_n^{\dagger T} X^{(i)} Y_i) \\ &- \sum_{i=1}^n [b((\beta^0)^T X^{(i)} + \lambda_n^\dagger\{\gamma_n^{\dagger T} Z^{(i)} - \beta_n^{\dagger T} X^{(i)}\}) - b((\beta^0)^T X^{(i)})]\Bigg\}.\end{aligned} \tag{40}$$

From (39), $P_{\beta^0}(\mathrm{LR}_n \geq 1) \leq E^{Q^\dagger}[\frac{dP}{dQ^\dagger}; (\frac{1}{n}\sum_{i=1}^n Z^{(i)} Y_i, \frac{1}{n}\sum_{i=1}^n X^{(i)} Y_i) \in A_n]$. This inequality, together with (40), gives

$$\begin{aligned}P_{\beta^0}(\mathrm{LR}_n \geq 1) \leq\, & e^{\sum_{i=1}^n \{b[(\beta^0)^T X^{(i)} + \lambda_n^\dagger\{\gamma_n^{\dagger T} Z^{(i)} - \beta_n^{\dagger T} X^{(i)}\}] - b[(\beta^0)^T X^{(i)}]\}} \\ & \times E^{Q^\dagger}\Bigg[e^{-\lambda_n^\dagger \sum_{i=1}^n (\gamma_n^{\dagger T} Z^{(i)} Y_i - \beta_n^{\dagger T} X^{(i)} Y_i)}; \\ & \left(\frac{1}{n}\sum_{i=1}^n Z^{(i)} Y_i, \frac{1}{n}\sum_{i=1}^n X^{(i)} Y_i\right) \in A_n\Bigg].\end{aligned} \tag{41}$$

The next lemma, whose proof is in the supplementary materials, shows a property of $\beta_n^\dagger$ and A_n.

LEMMA 19. *For all* $(s_1, s_2) \in A_n$, $\gamma^{\dagger T} s_2 - \frac{1}{n}\sum_{i=1}^n b(\gamma^{\dagger T} Z^{(i)}) \geq \beta_n^{\dagger T} s_1 - \frac{1}{n}\sum_{i=1}^n b(\beta_n^{\dagger T} X^{(i)})$.

By Lemma 19, the right-hand side of (41) is further bounded above by

$$\begin{aligned}&P_{\beta^0}(\mathrm{LR}_n \geq 1) \\ &\quad \leq \exp\Bigg\{\sum_{i=1}^n [b((\beta^0)^T X^{(i)} + \lambda_n^\dagger\{\gamma_n^{\dagger T} Z^{(i)} - \beta_n^{\dagger T} X^{(i)}\}) - b((\beta^0)^T X^{(i)})] \\ &\quad\quad - \lambda_n^\dagger \sum_{i=1}^n [b(\gamma_n^{\dagger T} Z^{(i)}) - b(\beta_n^{\dagger T} X^{(i)})]\Bigg\} \\ &\quad\quad \times Q^\dagger\left[\left(\frac{1}{n}\sum_{i=1}^n Z^{(i)} Y_i, \frac{1}{n}\sum_{i=1}^n X^{(i)} Y_i\right) \in A_n\right].\end{aligned} \tag{42}$$

Because $Q^\dagger[(\frac{1}{n}\sum_{i=1}^n Z^{(i)} Y_i, \frac{1}{n}\sum_{i=1}^n X^{(i)} Y_i) \in A_n] \leq 1$, we arrive at

$$\begin{aligned}&P_{\beta^0}(\mathrm{LR}_n \geq 1) \\ &\quad \leq \exp\Bigg\{\sum_{i=1}^n [b((\beta^0)^T X^{(i)} + \lambda_n^\dagger\{\gamma_n^{\dagger T} Z^{(i)} - \beta_n^{\dagger T} X^{(i)}\}) - b((\beta^0)^T X^{(i)})]\end{aligned}$$

$$-\lambda_n^\dagger \sum_{i=1}^{n}\bigl[b\bigl(\gamma_n^{\dagger T} Z^{(i)}\bigr) - b\bigl(\beta_n^{\dagger T} X^{(i)}\bigr)\bigr]\Biggr\},$$

which, by definition of $\widetilde{\rho}_n^\dagger$, equals $e^{-n\widetilde{\rho}_n^\dagger}$. Thus, $P_{\beta^0}(\mathrm{LR}_n \geq 1) \leq e^{-n\widetilde{\rho}_n^\dagger}$.

Lower bound. First, $\sum_{i=1}^n \log h_i(Y_i, \gamma_n^\dagger) \geq \sup_\beta \sum_{i=1}^n \log g_i(Y_i, \beta)$ implies $\sup_\gamma \sum_{i=1}^n \log h_i(Y_i, \gamma) \geq \sup_\beta \sum_{i=1}^n \log g_i(Y_i, \beta)$. Thus, a lower bound for $P_{\beta^0}(\mathrm{LR}_n \geq 1)$ is $P_{\beta^0}(\sum_{i=1}^n \log h_i(Y_i, \gamma_n^\dagger) - \sup_\beta \sum_{i=1}^n \log g_i(Y_i, \beta) \geq 0)$, which, by the definition of $Q^\dagger$ in (40), equals

$$\begin{aligned}
&\exp\Biggl\{\sum_{i=1}^{n}\bigl[b\bigl((\beta^0)^T X^{(i)} + \lambda_n^\dagger\{\gamma_n^{\dagger T} Z^{(i)} - \beta_n^{\dagger T} X^{(i)}\}\bigr) - b\bigl((\beta^0)^T X^{(i)}\bigr)\bigr]\Biggr\}\\
&\qquad \times E^{Q^\dagger}\bigl[e^{-\lambda_n^\dagger \sum_{i=1}^n (\gamma_n^{\dagger T} Z^{(i)} Y_i - \beta_n^{\dagger T} X^{(i)} Y_i)}; E\bigr],
\end{aligned}\tag{43}$$

where $E = \{\sum_{i=1}^n \gamma_n^{\dagger T} Z^{(i)} Y_i - \hat{\beta}_n^T X^{(i)} Y_i - b(\gamma_n^{\dagger T} X^{(i)}) + b(\hat{\beta}_n^T X^{(i)}) \geq 0\}$ and $\hat{\beta}_n = \arg\sup_\beta\{\sum_{i=1}^n \beta^T X^{(i)} Y_i - b(\beta X^{(i)})\}$. Because $\exp\{\sum_{i=1}^n [b((\beta^0)^T X^{(i)} + \lambda_n^\dagger\{\gamma_n^{\dagger T} Z^{(i)} - \beta_n^{\dagger T} X^{(i)}\}) - b((\beta^0)^T X^{(i)})] - \lambda_n^\dagger [b(\gamma_n^{\dagger T} Z^{(i)}) - b(\beta_n^{\dagger T} X^{(i)})]\} = e^{-n\widetilde{\rho}_n^\dagger}$, we have

$$P_{\beta^0}(\mathrm{LR}_n \geq 1) \geq e^{-n\widetilde{\rho}_n} \times J,\tag{44}$$

where $J = E^{Q^\dagger}[e^{-\lambda_n^\dagger [\sum_{i=1}^n \gamma_n^{\dagger T} Z^{(i)} Y_i - \beta_n^{\dagger T} X^{(i)} Y_i - b(\gamma_n^{\dagger T} X^{(i)}) + b(\beta_n^{\dagger T} X^{(i)})]}; E]$. We proceed to establishing a lower bound of J. We consider two events $E_1 = \{\frac{\sqrt{n}}{2} < \sum_{i=1}^n \gamma_n^{\dagger T} Z^{(i)} Y_i - \beta_n^{\dagger T} X^{(i)} Y_i - b(\gamma_n^{\dagger T} X^{(i)}) + b(\beta_n^{\dagger T} X^{(i)}) \leq \sqrt{n}\}$ and $E_2 = \{\sum_{i=1}^n [\hat{\beta}_n^T X^{(i)} Y_i - \beta_n^{\dagger T} X^{(i)} Y_i - b(\hat{\beta}_n^T X^{(i)}) + b(\beta_n^{\dagger T} X^{(i)})] \leq \frac{\sqrt{n}}{2}\}$. Because E_1 together with E_2 implies E, we have $E \supset E_1 \cap E_2$. Consequently, $J \geq E^{Q^\dagger}[e^{-\lambda_n^\dagger [\sum_{i=1}^n \gamma_n^{\dagger T} Z^{(i)} Y_i - \beta_n^{\dagger T} X^{(i)} Y_i - b(\gamma_n^{\dagger T} X^{(i)}) + b(\beta_n^{\dagger T} X^{(i)})]}; E_1 \cap E_2]$. Notice that on the set E_1, $\sum_{i=1}^n \gamma_n^{\dagger T} Z^{(i)} Y_i - \beta_n^{\dagger T} X^{(i)} Y_i - b(\gamma_n^{\dagger T} X^{(i)}) + b(\beta_n^{\dagger T} X^{(i)}) \leq \sqrt{n}$. Therefore,

$$J \geq e^{-\lambda_n^\dagger \sqrt{n}} Q^\dagger(E_1 \cap E_2) \geq e^{-\lambda_n^\dagger \sqrt{n}}\bigl(Q^\dagger(E_1) - Q^\dagger(E_2^c)\bigr).\tag{45}$$

We provide an upper bound for $Q^\dagger(E_1)$ and a lower bound for $Q^\dagger(E_2^c)$.

LEMMA 20. *Let $v_n = \mathrm{Var}^{Q^\dagger}(\sum_{i=1}^n \gamma_n^{\dagger T} Z^{(i)} Y_i - \beta_n^{\dagger T} X^{(i)} Y_i - b(\gamma_n^{\dagger T} X^{(i)}) + b(\beta_n^{\dagger T} X^{(i)}))$, then $v_n = O(n)$ as $n \to \infty$. Furthermore, we have*

$$\mathcal{L}\Biggl(v_n^{-\frac{1}{2}}\Biggl[\sum_{i=1}^{n} \gamma_n^{\dagger T} Z^{(i)} Y_i - \beta_n^{\dagger T} X^{(i)} Y_i - b\bigl(\gamma_n^{\dagger T} X^{(i)}\bigr) + b\bigl(\beta_n^{\dagger T} X^{(i)}\bigr)\Biggr]\Biggr) \to N(0,1).$$

Here, $\mathcal{L}(\cdot)$ denotes the law of random variables and $N(0,1)$ is the distribution of standard normal.

This lemma, to be proved in the supplementary materials, shows that there exists a constant $\varepsilon > 0$ such that

$$Q^\dagger(E_1) \geq \varepsilon. \tag{46}$$

For $Q^\dagger(E_2)$, let $u(\mu,\beta) = (\beta - \beta_n^\dagger)^T \mu - \sum_{i=1}^n [b(\beta^T X^{(i)}) - b(\beta_n^\dagger X^{(i)})]$ and $v(\mu) = \sup_\beta u(\mu,\beta)$.

LEMMA 21. *Let* $\mu^\dagger = \sum_{i=1}^n b'(\lambda_n^\dagger(\gamma_n^{\dagger T} Z^{(i)} - \beta_n^\dagger X^{(i)}) + (\beta^0)^T X^{(i)})X^{(i)}$, *then* $v(\mu)$ *is twice continuously differentiable around* $\mu^\dagger$, *with* $v(\mu^\dagger) = 0$ *and* $\nabla v(\mu^\dagger) = 0$. *Moreover,* $\nabla^2 v(\mu) = \{\sum_{i=1}^n b''[\beta(\mu)^T X^{(i)}] X^{(i)} X^{(i)T}\}^{-1}$, *where* $\beta(\mu) = \arg\sup_\beta u(\mu,\beta)$.

The proof of this lemma is in the supplementary materials. According to Lemma 21 and Taylor expansion of $v(\mu)$ around $\mu^\dagger$, we have

$$\left\{v(\mu) \geq \frac{\sqrt{n}}{2}\right\} \subset \left\{\frac{1}{2}\|\mu - \mu^\dagger\|^2 \|\nabla^2 v(\mu^\dagger)\|_2 \geq \frac{\sqrt{n}}{2}\right\}, \tag{47}$$

where $\|\cdot\|_2$ is the spectral norm of matrices. By Lemma 21 and Assumptions A10 and A11, $\|\nabla^2 v(\mu^\dagger)\|_2 = O(n^{-1})$. Therefore, (47) implies $\{v(\mu) \geq \frac{\sqrt{n}}{2}\} \subset \{\|\mu - \mu^\dagger\| \geq \varepsilon n^{\frac{3}{4}}\}$. Since $E_2^c = \{v(\sum_{i=1}^n X^{(i)} Y_i) \geq \frac{\sqrt{n}}{2}\}$, we have $Q^\dagger(E_2^c) \leq Q^\dagger(\|\sum_{i=1}^n X^{(i)} Y_i - \mu^\dagger\| \geq \varepsilon n^{\frac{3}{4}})$. This, together with Chebyshev's inequality, implies $Q^\dagger(E_2^c) \leq (\varepsilon^{-2} n^{-\frac{3}{2}}) E^{Q^\dagger} \|\sum_{i=1}^n X^{(i)} Y_i - \mu^\dagger\|^2$. Because $E^{Q^\dagger} \|\sum_{i=1}^n X^{(i)} Y_i - \mu^\dagger\|^2 = O(n)$, we have $Q^\dagger(E_2^c) \to 0$ as $n \to \infty$. Combining this result with (45) and (46), we arrive at a lower bound $J \geq \frac{\varepsilon}{2} e^{-\lambda_n^\dagger \sqrt{n}}$. This lower bound of J together with (44) give a lower bound $P(\mathrm{LR}_n \geq 1) \geq \frac{\varepsilon}{2} e^{-n\widetilde{\rho}_n^\dagger - \lambda_n^\dagger \sqrt{n}}$. Under Assumption A9, $\widetilde{\rho}_n^\dagger \geq \inf_\gamma \sup_\lambda \widetilde{\rho}_n(\beta^0, \gamma, \lambda) \geq \delta_1$, so $\lambda_n^\dagger \sqrt{n} = o(1) n \widetilde{\rho}_n^\dagger$. Thus, we have $P_{\beta^0}(\mathrm{LR}_n \geq 1) \geq e^{-n\widetilde{\rho}_n^\dagger (1+o(1))}$. We complete the proof by combining the lower and upper bounds for $P_{\beta^0}(\mathrm{LR}_n \geq 1)$.

Acknowledgment. The authors thank the Associate Editor and two referees for their helpful comments and suggestions.

SUPPLEMENTARY MATERIAL

Supplement to "Chernoff index for Cox test of separate parametric families" (DOI: 10.1214/16-AOS1532SUPP; .pdf). In the supplement [Li, Liu and Ying (2017)], we present proofs of Corollary 6, Lemmas 17, 18 19, 20 and 21.

REFERENCES

ADLER, R. J., BLANCHET, J. H. and LIU, J. (2012). Efficient Monte Carlo for high excursions of Gaussian random fields. *Ann. Appl. Probab.* **22** 1167–1214. MR2977989

ANDRESEN, A. and SPOKOINY, V. (2014). Critical dimension in profile semiparametric estimation. *Electron. J. Stat.* **8** 3077–3125. MR3301302

ARCONES, M. A. (2006). Large deviations for M-estimators. *Ann. Inst. Statist. Math.* **58** 21–52. MR2281205

BAHADUR, R. R. (1960). Stochastic comparison of tests. *Ann. Math. Stat.* **31** 276–295. MR0116413

BAHADUR, R. R. (1967). Rates of convergence of estimates and test statistics. *Ann. Math. Stat.* **38** 303–324. MR0207085

BERRINGTON DE GONZÁLEZ, A. and COX, D. R. (2007). Interpretation of interaction: A review. *Ann. Appl. Stat.* **1** 371–385. MR2415740

BRAGANCA PEREIRA, B. (2005). Separate families of hypotheses. In *Encyclopedia of Biostatistics*.

CHERNOFF, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Stat.* **23** 493–507. MR0057518

COX, D. R. (1961). Tests of separate families of hypotheses. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 105–123. Univ. California Press, Berkeley, CA. MR0131927

COX, D. R. (1962). Further results on tests of separate families of hypotheses. *J. Roy. Statist. Soc. Ser. B* **24** 406–424. MR0156409

COX, D. R. (2013). A return to an old paper: 'Tests of separate families of hypotheses'. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 207–215. MR3021385

DAVIDSON, R. and MACKINNON, J. G. (1981). Several tests for model specification in the presence of alternative hypotheses. *Econometrica* **49** 781–793. MR0619482

FINE, J. P. (2002). Comparing nonnested Cox models. *Biometrika* **89** 635–647. MR1929168

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability* (*Berkeley, Calif.*, 1965/66), *Vol. I*: *Statistics* 221–233. Univ. California Press, Berkeley, CA. MR0216620

KALLENBERG, W. C. M. (1983). Intermediate efficiency, theory and examples. *Ann. Statist.* **11** 170–182. MR0684874

LI, X. and LIU, J. (2015). Rare-event simulation and efficient discretization for the supremum of Gaussian random fields. *Adv. in Appl. Probab.* **47** 787–816. MR3406608

LI, X., LIU, J. and YING, Z. (2017). Supplement to "Chernoff Index for Cox Test of Separate Parametric Families." DOI:10.1214/16-AOS1532SUPP.

LIU, J., LU, J. and ZHOU, X. (2015). Efficient rare event simulation for failure problems in random media. *SIAM J. Sci. Comput.* **37** A609–A624. MR3317386

LIU, J. and XU, G. (2012). Some asymptotic results of Gaussian random fields with varying mean functions and the associated processes. *Ann. Statist.* **40** 262–293. MR3014307

LIU, J. and XU, G. (2014a). On the conditional distributions and the efficient simulations of exponential integrals of Gaussian random fields. *Ann. Appl. Probab.* **24** 1691–1738. MR3211008

LIU, J. and XU, G. (2014b). Efficient simulations for the exponential integrals of Hölder continuous Gaussian random fields. *ACM Trans. Model. Comput. Simul.* **24** Art. 9, 24. MR3195228

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, London. MR3223057

PESARAN, M. H. (1974). On the general problem of model selection. *Rev. Econ. Stud.* **41** 153–171.

PESARAN, M. H. (1984). Asymptotic power comparisons of tests of separate parametric families by Bahadur's approach. *Biometrika* **71** 245–252. MR0767152

PESARAN, M. H. and DEATON, A. S. (1978). Testing non-nested nonlinear regression models. *Econometrica* **46** 677–694. MR0501711

RUKHIN, A. L. (1993). Bahadur efficiency of tests of separate hypotheses and adaptive test statistics. *J. Amer. Statist. Assoc.* **88** 161–165. MR1212484

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York. MR0595165

SPOKOINY, V. (2012). Parametric estimation. Finite sample theory. *Ann. Statist.* **40** 2877–2909. MR3097963

VUONG, Q. H. (1989). Likelihood ratio tests for model selection and nonnested hypotheses. *Econometrica* **57** 307–333. MR0996939

WHITE, H. (1982a). Maximum likelihood estimation of misspecified models. *Econometrica* **50** 1–25. MR0640163

WHITE, H. (1982b). Regularity conditions for Cox's test of nonnested hypothesis. *J. Econometrics* **19** 301–318. MR0672058

X. LI
SCHOOL OF STATISTICS
UNIVERSITY OF MINNESOTA
224 CHURCH STREET SE
MINNEAPOLIS, MINNESOTA 55455
USA
E-MAIL: lixx1766@umn.edu

J. LIU
Z. YING
DEPARTMENT OF STATISTICS
COLUMBIA UNIVERSITY
1255 AMSTERDAM AVENUE
NEW YORK, NEW YORK 10027
USA
E-MAIL: jcliu@stat.columbia.edu
E-MAIL: zying@stat.columbia.edu