econometrics

MDPI

Article

# Generalized Information Matrix Tests for Detecting Model Misspecification

**Richard M. Golden [1,*], Steven S. Henley [2,3,6], Halbert White [4,†] and T. Michael Kashner [3,5,6,7]**

1 School of Behavioral and Brain Sciences, GR4.1, 800 W. Campbell Rd., University of Texas at Dallas, Richardson, TX 75080, USA
2 Martingale Research Corporation, 101 E. Park Blvd., Suite 600, Plano, TX 75074, USA; stevenh@martingale-research.com
3 Department of Medicine, Loma Linda University School of Medicine, Loma Linda, CA 92357, USA
4 Department of Economics, University of California San Diego, La Jolla, CA 92093, USA
5 Office of Academic Affiliations (10A2D), Department of Veterans Affairs, 810 Vermont Ave. NW (10A2D), Washington, DC 20420, USA; michael.kashner@va.gov
6 Center for Advanced Statistics in Education, VA Loma Linda Healthcare System, Loma Linda, CA 92357, USA
7 Department of Psychiatry, University of Texas Southwestern Medical Center at Dallas, Dallas, TX 75390, USA
* Correspondence: golden@utdallas.edu; Tel.: +1-972-883-2423
† Halbert White sadly passed away before this article was published.

Academic Editors: Kerry Patterson and Marc S. Paolella
Received: 29 December 2015; Accepted: 26 October 2016; Published: 15 November 2016

**Abstract:** Generalized Information Matrix Tests (GIMTs) have recently been used for detecting the presence of misspecification in regression models in both randomized controlled trials and observational studies. In this paper, a unified GIMT framework is developed for the purpose of identifying, classifying, and deriving novel model misspecification tests for finite-dimensional smooth probability models. These GIMTs include previously published as well as newly developed information matrix tests. To illustrate the application of the GIMT framework, we derived and assessed the performance of new GIMTs for binary logistic regression. Although all GIMTs exhibited good level and power performance for the larger sample sizes, GIMT statistics with fewer degrees of freedom and derived using log-likelihood third derivatives exhibited improved level and power performance.

**Keywords:** asymptotic theory; Information Matrix Test; specification analysis; logistic regression; simulation study; information ratio; misspecification

**JEL Classification:** C12; C13; C15; C18; C52

## 1. Introduction

If a researcher's probability model of the observed data is not correctly specified, then the interpretation of its parameter estimates may not be valid, leading to incomplete or incorrect conclusions. Thus, whether a model is correctly specified must be considered when analyzing and interpreting data (e.g., [1,2]). This issue is critically important in econometrics as well as more general scientific inquiry. For example, in health economics, estimates of the impact of clinical treatments [3,4], care systems [5], and health policy interventions on health outcomes [6] are dependent on the underlying assumption that the model to be tested is correctly specified. Further, model misspecification testing is essential for statistical analysis of randomized control trials [7,8] and observational studies [9,10]. For these reasons, this paper introduces a unified framework for identifying, classifying, and developing a wide range of specification tests.

### *1.1. Information Matrix Test Methods for Detection of Model Misspecification*

Assume that the data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ observed in an experiment is a realization of a sequence of independent and identically distributed $d$-dimensional random vectors $\mathbf{X}_1, ..., \mathbf{X}_n$ with a common *data generating process density* $p_x$. Let $\mathcal{M} \equiv \{f(\mathbf{x}; \theta) : \theta \in \Theta\}$ denote a proposed probability model that is a collection of probability densities indexed by a $k$-dimensional parameter vector $\theta$. If $p_x \in \mathcal{M}$, so that $p_x(\mathbf{x}) = f(\mathbf{x}; \theta^*)$ *a.e.* for some $\theta^* \in \Theta$, then $\mathcal{M}$ is *correctly specified* with respect to $p_x$.

When $\mathcal{M}$ is correctly specified with respect to $p_x$, the inverse of the asymptotic covariance matrix of the maximum likelihood estimator $\hat{\theta}_n \equiv \text{argmax} \prod_{i=1}^{n} f(\mathbf{X}_i; \theta)$ is equal to both the *inverse Hessian covariance matrix* $\mathbf{A}^* \equiv -\nabla^2 E\{\log f(\mathbf{X}_i; \theta^*)\}$ and the *inverse Outer Product Gradient (OPG) covariance matrix* $\mathbf{B}^* \equiv E\left\{\nabla \log f(\mathbf{X}_i; \theta^*) (\nabla \log f(\mathbf{X}_i; \theta^*))^T\right\}$. This classic result is called the *Information Matrix Equality* (see [1,2], and Theorem 4 of this paper for relevant reviews).

Let $u : \mathcal{R}^k \to \mathcal{R}$. The notation $\nabla u$ refers to a $k$-dimensional column vector of functions called the *gradient* whose $i$th element is $\frac{\partial u}{\partial x_i}$, $i = 1, ..., k$. The notation $\nabla^2 u$ refers to a $k$-dimensional matrix-valued function which is called the *Hessian* of $u$. The element in the $i$th row and $j$th column of $\nabla^2 u$ is $\frac{\partial^2 u}{\partial x_i \partial x_j}$, $i, j = 1, ..., k$.

As described by White [1,2], the information matrix equality may be used as the basis for a test of model misspecification. White [1] proposed the Information Matrix Test (IMT) for testing the null hypothesis that the elements of the $k$-dimensional Hessian and $k$-dimensional Outer Product Gradient (OPG) inverse asymptotic covariance matrices (denoted by $\mathbf{A}^*$ and $\mathbf{B}^*$ respectively) are equal. That is, White [1] considered the null hypothesis: $H_o : vech(\mathbf{A}^* - \mathbf{B}^*) = \mathbf{0}_{k(k+1)/2}$ where $\mathbf{0}_{k(k+1)/2}$ denotes a $k(k + 1)/2$-dimensional column vector of zeros. Rejection of this null hypothesis thus implies a violation of the information matrix equality and thus the presence of model misspecification. Moreover, as noted by White [1], it may be helpful to also consider situations where the null hypothesis is "directional." If a directional null hypothesis is rejected, this implies $H_o : vech(\mathbf{A}^* - \mathbf{B}^*) = \mathbf{0}_{k(k+1)/2}$ is rejected (but the converse of this latter statement does not hold). White [1], in particular, discussed directional IMTs that have the form: $H_o : \mathbf{S}vech(\mathbf{A}^* - \mathbf{B}^*) = \mathbf{0}_r$ where the selection matrix $\mathbf{S} \in \mathcal{R}^{r \times (k(k+1)/2)}$ consists of $r$ rows of a $k(k + 1)/2$-dimensional identity matrix. In some cases directional IMTs may have more statistical power because they are designed to identify specific types of model misspecification.

For many years, the IMT approach has not been widely used outside of linear regression modeling because various instabilities (possibly associated with large degrees of freedom) of the test were observed. Chesher [11] and Lancaster [12] demonstrated how the calculation of the third derivatives of the log-likelihood function could be avoided for the full IMT, but the effectiveness of their approach was shown in some cases to exhibit unacceptable performance in logistic regression and linear regression [13–18].

### *1.2. Recent Developments in Information Matrix Test Theory*

An advance in the theory of information matrix testing was provided by Presnell and Boos [19] (also see, [20–22]), who introduced an IOS (in and out of sample) directional IMT and showed that it was effective in a variety of important situations through both theoretical analyses and simulation studies. More recently, Golden et al. [23] introduced a general unified theory for model specification testing based upon a nonlinear extension of White's [1] approach to specification testing. The new IMTs developed within the framework of Golden et al. [23] are called Generalized Information Matrix Tests (GIMT).

In particular, Golden et al. [23] discussed the problem of testing the null hypothesis that a smooth nonlinear *GIMT hypothesis function* $\mathbf{s} : \mathcal{R}^{k \times k} \times \mathcal{R}^{k \times k} \to \mathcal{R}^r$ of the Hessian and OPG inverse asymptotic covariance matrices is equal to an $r$-dimensional vector of zeros. That is, a GIMT tests the null hypothesis $H_o : \mathbf{s}(\mathbf{A}^*, \mathbf{B}^*) = \mathbf{0}_r$. Golden et al. [23] emphasized that different choices of

GIMT hypothesis function yield different types of directional and non-directional GIMT hypotheses. Although Golden et al. [23] did not provide explicit regularity conditions and a detailed analysis of their proposed general class of GIMTs, Golden et al. [23] introduced key formal definitions, provided an informal discussion of relevant theoretical results, and reported the results of a comprehensive simulation study of a realistic epidemiological analysis problem using logistic regression for six new GIMTs that exhibited appealing level and power performance. This approach for the detection of model misspecification has now been used in observational and randomized controlled trial studies [7–10].

Since the publication of Golden et al. [23], Cho and White [24] described an important class of non-directional GIMTs and showed that each of their three test statistics for model misspecification was asymptotically distributed as a squared Gaussian random variable under the null hypothesis. In addition, Cho and White [24] provided analyses of the power of their test statistics under local and global alternatives. Zhou et al. [25] proposed a non-directional GIMT statistic for the large important class of regression models where the distribution of the response variable conditioned upon the covariates is a member of the linear exponential family. Like Cho and White [24], they showed their misspecification test statistic has only a single degree of freedom and is asymptotically distributed as a squared Gaussian random variable under the null hypothesis. Huang and Prokhorov [26] also showed how the information matrix testing framework is useful for investigating goodness-of-fit using non-directional GIMT statistics for semi-parametric probability models that are specified by copulas. All of this previous work on GIMTs can be interpreted as special cases or variants of special cases of the general framework of Golden et al. [23] for finite-dimensional smooth probability models.

This paper provides a unified framework for addressing the detection of model misspecification using a variety of GIMT statistics for a large class of finite-dimensional smooth probability models. By presenting the details of the GIMT framework and explicitly presenting the relevant regularity assumptions, it establishes the foundation for supporting research into the further development of a large class of GIMTs as well as assisting in understanding the similarities and differences between different GIMTs in the existing published statistical literature.

Our paper is organized in the following manner. In Section 2, we provide the assumptions of the GIMT framework. In Section 3, we characterize the asymptotic distribution of a large family of GIMTs for a large class of finite-dimensional smooth probability models under the assumptions and definitions in Section 2. In Section 4, we investigate the performance of new GIMTs using simulation studies developed with respect to a particular logistic regression model intended to be representative of a commonly encountered problem of model misspecification detection. Conclusions are provided in Section 5.

## 2. GIMT Theoretical Framework: Definitions and Assumptions

In this section, we introduce the definitions and assumptions of our formal mathematical theory of Generalized Information Matrix Tests. In most practical applications, these assumptions are often satisfied for thrice continuously differentiable probability models with a fixed number of free parameters that have locally unique solutions. Throughout, it is assumed that observations are independent and identically distributed.

### *2.1. Data Generating Process*

Let $\mathcal{B}\left(\mathcal{R}^d\right)$ be the Borel $\sigma$-field generated by the open subsets of $\mathcal{R}^d$.

**Assumption 1.** ***Data Generating Process (DGP).*** *Let* $\mathbf{X}_i$, $i = 1, 2, \ldots$ *be a sequence of independent and identically distributed (i.i.d) random vectors where* $\mathbf{X}_i$ *has a common probability measure P on the measurable space* $\left(\mathcal{R}^d, \mathcal{B}\left(\mathcal{R}^d\right)\right)$ *with completion* $\left(\mathcal{R}^d, \mathcal{F}_0, P_0\right)$.

Let the triplet $(\Omega, \mathcal{F}_o, \mathcal{P}_o)$ be the *probability space* for the Data Generating Process (DGP).

In regression modeling applications, the first element of the $d$-dimensional real vector $\mathbf{x}_i$ (a realization of $\mathbf{X}_i$) may be a particular value of the outcome (dependent) variable for a regression model associated with the $i$th data record, the second element of $\mathbf{x}_i$ may be the number 1 for the purpose of introducing an intercept parameter, and the remaining elements of $\mathbf{x}_i$ may be particular values for the predictor variables associated with the $i$th data record, $i = 1, \ldots, n$.

Although Assumption 1 assumes that the observed data $\boldsymbol{X}_i$, $i = 1, 2, \ldots$ are *i.i.d.*, the theory presented here is also applicable to panel data analyses. For example, consider a situation where data are collected in a longitudinal study on a group of individuals over a period of time. Assume the observations across participants are assumed to be *i.i.d.*, but the observations for a particular participant are neither necessarily identically distributed nor independent. Let $\boldsymbol{X}_{it}$ denote the observation associated with the measurement of the $i$th participant in the study at time index $t$ for $t = 1, \ldots, T$ (where $T$ is a fixed finite number) and $i = 1, \ldots, n$. The theory described in this article is applicable to evaluating the degree to which a probability model can account for the observed data $\boldsymbol{X}_i \equiv \left[ \begin{array}{ccc} \boldsymbol{X}_{i,1} & \ldots & \boldsymbol{X}_{i,T} \end{array} \right]$, $i = 1, \ldots, n$.

The following assumption of absolute continuity is now introduced to permit alternative representations of $P_0$ in order to represent, construct, and manipulate probability densities for data generating processes involving data samples containing combinations of discrete and continuous random variables.

**Assumption 2.** ***Absolute Continuity.*** *Let $\nu_j\left(x_j\right)$ be a $\sigma$-finite measure on the measurable space $\left(\mathcal{R}, \mathcal{B}\left(\mathcal{R}\right)\right)$, $j = 1, \ldots, d$. Let $\nu \equiv \overset{d}{\underset{j=1}{\otimes}} \nu_j\left(x_j\right)$ be a $\sigma$-finite product measure on the measurable space $\left(\mathcal{R}^d, \mathcal{B}\left(\mathcal{R}^d\right)\right)$. Assume $P_0$ is absolutely continuous with respect to $\nu$.*

By the Radon-Nikodým Theorem, Assumption 2 guarantees the joint distribution of $\mathbf{X}_i$, $P_0$, may be represented using a *Radon-Nikodým density function*. The Radon-Nikodým density $p_x \equiv dP_0/d\nu$ is common to the *i.i.d.* random variables $\mathbf{X}_i$, $i = 1, \ldots, n$ on the measurable space $\left(\mathcal{R}^d, \mathcal{B}\left(\mathcal{R}^d\right)\right)$.

Assumption 2 allows the theoretical results developed here to be applicable to random vectors that contain both discrete and absolutely continuous components. If a random vector is a discrete random vector or an absolutely continuous random vector, then the Radon-Nikodým density becomes a probability mass function or an absolutely continuous probability density function and the associated measure theory notation may be avoided.

### *2.2. Probability Model*

Let $supp\ \boldsymbol{X}$ denote the support of $\mathbf{X}$.

**Assumption 3.** ***Parametric Densities.*** *(i) Let $\Theta$ be a compact and non-empty subset of $\mathcal{R}^k$, $k \in \mathbb{N}$; (ii) Let $f : \mathcal{R}^d \times \Theta \to [0, \infty)$. For each $\boldsymbol{\theta}$ in $\Theta$, $f(\cdot; \theta)$ is a density with respect to $\nu$ and $f(\mathbf{x}; \cdot)$ is continuous on $\Theta$ for each $\boldsymbol{x} \in supp\ \boldsymbol{X}$; (iii) $\log f(\mathbf{x}; \cdot)$ is continuously differentiable on $\Theta$ for each $\boldsymbol{x} \in supp\ \boldsymbol{X}$; (iv) $\log f(\mathbf{x}; \cdot)$ is twice continuously differentiable on $\Theta$ for each $\boldsymbol{x} \in supp\ \boldsymbol{X}$; (v) $\log f(\mathbf{x}; \cdot)$ is thrice continuously differentiable on $\Theta$ for each $\boldsymbol{x} \in supp\ \boldsymbol{X}$.*

**Definition.** ***Probability Model.*** Let $f$ be defined as in Assumption 3(i) and Assumption 3(ii). Let $F : \mathcal{R}^d \times \Theta \to [0, 1]$ be defined such that for each $\theta$ in $\Theta$, $F(\cdot; \theta) : \mathcal{R}^d \to [0, 1]$ is the probability distribution for $\boldsymbol{X}$ specified by density $f(\cdot; \theta)$. The set $\mathcal{M} \equiv \left\{ F(\cdot; \theta) : \mathcal{R}^d \to [0, 1] | \theta \in \Theta \right\}$ is the *probability model on* $\boldsymbol{\Theta}$ specified by $f$.

**Definition.** ***Misspecified Model.*** The probability model $\mathcal{M}$ is *misspecified* when $P_0 \notin \mathcal{M}$, otherwise $\mathcal{M}$ is *correctly specified*.

### 2.3. Hypothesis Function

**Definition.** ***GIMT Hypothesis Function***. Let $\Upsilon$ be a compact and non-empty subset of $\mathcal{R}^{k\times k}$, $k \in \mathbb{N}$. A *Generalized Information Matrix Test (GIMT) Hypothesis function* $\mathbf{s}: \Upsilon \times \Upsilon \to \mathcal{R}^r$ has the property that if $\mathbf{A} = \mathbf{B}$, then $\mathbf{s}(\mathbf{A}, \mathbf{B}) = \mathbf{0}_r$ for every symmetric positive definite matrix $\mathbf{A} \in \Upsilon$ and for every symmetric positive definite matrix $\mathbf{B} \in \Upsilon$.

**Definition.** ***Nondirectional and directional GIMT Hypothesis Functions***. Let $\Upsilon$ be a compact and non-empty subset of $\mathcal{R}^{k\times k}$, $k \in \mathbb{N}$. A *nondirectional GIMT hypothesis function* $\mathbf{s}: \Upsilon \times \Upsilon \to \mathcal{R}^r$ has the property $\mathbf{A} = \mathbf{B}$ if and only if $\mathbf{s}(\mathbf{A}, \mathbf{B}) = \mathbf{0}_r$ for all $(\mathbf{A}, \mathbf{B}) \in \Upsilon \times \Upsilon$. A *directional GIMT hypothesis function* is a GIMT hypothesis function that is not nondirectional.

When $\mathbf{A}: \mathcal{R}^{m\times n} \to \mathcal{R}^{q\times r}$, let $\frac{d\mathbf{A}}{d\mathbf{B}} \equiv \frac{dvec(\mathbf{A}^T)}{dvec(\mathbf{B}^T)}$ when it exists (e.g., [27]; also see [28,29]). Let $\nabla\mathbf{s}: \Upsilon \times \Upsilon \to \mathcal{R}^{r\times 2k^2}$ be defined such that for all $\mathbf{A}, \mathbf{B} \in \Upsilon$: $\nabla\mathbf{s}(\mathbf{A}, \mathbf{B}) \equiv \left[ \begin{array}{cc} \frac{\partial \mathbf{s}(\cdot,\mathbf{B})}{\partial vec(\mathbf{A})} & \frac{\partial \mathbf{s}(\mathbf{A},\cdot)}{\partial vec(\mathbf{B})} \end{array} \right]$ when it exists.

**Assumption 4.** ***Hypothesis Function Regularity Conditions.*** *(i) Let $\Upsilon$ be a compact and non-empty subset of $\mathcal{R}^{k\times k}$, $k \in \mathbb{N}$. Let $\mathbf{s}: \Upsilon \times \Upsilon \to \mathcal{R}^r$ be continuous on $\Upsilon \times \Upsilon$; (ii) $\mathbf{A}^*$ and $\mathbf{B}^*$ are in the interior of $\Upsilon \subseteq \mathcal{R}^{k\times k}$; (iii) $\nabla\mathbf{s}$ exists and is continuous on $\Upsilon \times \Upsilon$; (iv) $\nabla\mathbf{s}^*$ has full row rank $r$ on $\Upsilon \times \Upsilon$.*

In practice, Assumption 4 provides a procedure for checking if the theory described here can be applied to a proposed GIMT hypothesis function.

**Definition.** ***Antisymmetric GIMT Hypothesis Function***. Let $\mathbf{s}: \Upsilon \times \Upsilon \to \mathcal{R}^r$ be a GIMT hypothesis function satisfying Assumption 4(i), Assumption 4(ii), and Assumption 4(iii). If, in addition, $\mathbf{s}(\mathbf{A}, \mathbf{B}) = -\mathbf{s}(\mathbf{B}, \mathbf{A})$ for all $(\mathbf{A}, \mathbf{B}) \in \Upsilon \times \Upsilon$, then $\mathbf{s}: \Upsilon \times \Upsilon \to \mathcal{R}^r$ is called an *antisymmetric GIMT hypothesis function*.

### 2.4. Notation

Let $\mathbf{g}(\mathbf{x}; \theta) \equiv -\nabla \log f(\mathbf{x}; \theta)$. Let $\overline{\mathbf{g}}_n(\theta) \equiv (1/n)\sum_{i=1}^{n} \mathbf{g}(\mathbf{X}_i; \theta)$. Let $\hat{\mathbf{g}}_n \equiv \overline{\mathbf{g}}_n\left(\hat{\theta}_n\right)$.

Let $\overline{\mathbf{A}}_n(\theta) \equiv -(1/n)\sum_{i=1}^{n} \nabla^2 \log f(\mathbf{X}_i; \theta)$. Let $\mathbf{A}(\theta) \equiv \nabla^2 l(\theta)$.

Let $\overline{\mathbf{B}}_n(\theta) \equiv (1/n)\sum_{i=1}^{n} \mathbf{g}(\mathbf{X}_i; \theta)(\mathbf{g}(\mathbf{X}_i; \theta))^T$.

Let $\mathbf{B}(\theta) \equiv \int \mathbf{g}(\mathbf{x}; \theta)(\mathbf{g}(\mathbf{x}; \theta))^T p_x(\mathbf{x})\, d\nu_x(\mathbf{x})$. Let $\hat{\mathbf{A}}_n \equiv \overline{\mathbf{A}}_n\left(\hat{\theta}_n\right)$. Let $\hat{\mathbf{B}}_n \equiv \overline{\mathbf{B}}_n\left(\hat{\theta}_n\right)$.

Let $\mathbf{A}^* = \mathbf{A}(\theta^*)$. Let $\mathbf{B}^* = \mathbf{B}(\theta^*)$. Let $\mathbf{d}_{x,\theta}(\mathbf{x}; \theta) \equiv \left[ \begin{array}{c} -vech\left(\nabla^2 \log f(\mathbf{x}; \theta)\right) \\ vech\left(\mathbf{g}(\mathbf{x}; \theta)(\mathbf{g}(\mathbf{x}; \theta))^T\right) \end{array} \right]$.

Let $\overline{\mathbf{d}}_n(\theta) = (1/n)\sum_{i=1}^{n} \mathbf{d}_{x,\theta}(\mathbf{X}_i; \theta)$. Let $\hat{\mathbf{d}}_n \equiv \overline{\mathbf{d}}_n\left(\hat{\theta}_n\right)$. Let $\overline{\mathbf{d}}_n^* \equiv \overline{\mathbf{d}}_n(\theta^*)$.

Let $\mathbf{d}(\theta) \equiv \left[ \begin{array}{c} vech(\mathbf{A}(\theta)) \\ vech(\mathbf{B}(\theta)) \end{array} \right]$. Let $\mathbf{d}^* \equiv \mathbf{d}(\theta^*)$.

Let the notation $\mathbf{I}_k$ denote a $k$-dimensional identity matrix.

Let the *duplication matrix* $\mathcal{D}_k: \mathcal{R}^{k(k+1)/2} \to \mathcal{R}^{k^2}$ be defined such that: $\mathcal{D}_k vech(\mathbf{A}) = vec(\mathbf{A})$ and the *inverse duplication matrix* $\mathcal{D}_k^{\dagger}: \mathcal{R}^{k^2} \to \mathcal{R}^{k(k+1)/2}$ be defined such that: $\mathcal{D}_k^{\dagger} vec(\mathbf{A}) = vech(\mathbf{A})$.

Let $\mathcal{D}_k^{\otimes} \equiv \mathbf{I}_2 \otimes \mathcal{D}_k$ and let $\mathcal{D}_k^{\otimes\dagger} \equiv \mathbf{I}_2 \otimes \mathcal{D}_k^{\dagger}$.

Let $\nabla\mathbf{d}: \Theta \to \mathcal{R}^{k(k+1)}$ where $\nabla\mathbf{d} \equiv \left[ \begin{array}{c} \frac{dvech(\mathbf{A})}{d\Theta} \\ \frac{dvech(\mathbf{B})}{d\Theta} \end{array} \right] = \mathcal{D}_k^{\otimes\dagger} \left[ \begin{array}{c} \frac{d\mathbf{A}}{d\Theta} \\ \frac{d\mathbf{B}}{d\Theta} \end{array} \right]$.

Let $\nabla \bar{\mathbf{d}}_n(\theta) \equiv (1/n)\sum_{i=1}^{n} \nabla \mathbf{d}_{x,\theta}(\mathbf{X}_i;\theta)$. Let $\nabla \hat{\mathbf{d}}_n \equiv \nabla \bar{\mathbf{d}}_n\left(\hat{\theta}_n\right)$. Let $\nabla \mathbf{d}^* \equiv \nabla \mathbf{d}(\theta^*)$.

### 2.5. Regularity Conditions

The following Assumption 5 uses a matrix version of the standard definition of dominated by an integrable function (see Appendix A).

**Assumption 5.** ***Domination Conditions***

*(i)(a)* $\log f(\mathbf{x};\theta)$ *is dominated on* $\Theta$ *with respect to* $p_x$*;*
*(i)(b)* $\frac{d\log f(\mathbf{x};\theta)}{d\theta}$ *is dominated on* $\Theta$ *with respect to* $p_x$*;*
*(i)(c)* $\mathbf{g}(\mathbf{x};\theta)(\mathbf{g}(\mathbf{x};\theta))^T$ *is dominated on* $\Theta$ *with respect to* $p_x$*;*
*(i)(d)* $\frac{d^2\log f(\mathbf{x};\theta)}{d\theta^2}$ *is dominated on* $\Theta$ *with respect to* $p_x$*;*
*(ii)(a)* $\frac{d(\mathbf{d}_{x,\theta}(\mathbf{x};\theta))}{d\theta}$ *is dominated on* $\Theta$ *with respect to* $p_x$*;*
*(ii)(b)* $\mathbf{d}_{x,\theta}(\mathbf{x};\theta)(\mathbf{d}_{x,\theta}(\mathbf{x};\theta))^T$ *is dominated on* $\Theta$ *with respect to* $p_x$*;*
*(ii)(c)* $\mathbf{g}(\mathbf{x};\theta)(\mathbf{d}_{x,\theta}(\mathbf{x};\theta))^T$ *is dominated on* $\Theta$ *with respect to* $p_x$*;*
*(iii) There exists a finite positive number K such that for all* $\mathbf{x} \in supp\ \mathbf{X}$ *and for all* $\theta \in \Theta$ *:* $\left|\frac{f(\mathbf{x};\theta)}{p_x(\mathbf{x})}\right| \leq K$.

Assumption 5 identifies specific regularity conditions that are used here to ensure that relevant expectations exist, that integral and differentiation operators can be interchanged, and that relevant laws of large numbers are applicable.

Assumption 5(i) is used to ensure that the conclusions of Theorems 2, 3, 4, 5, 6, and 7 hold. These theorems characterize the asymptotic distribution of the quasi-maximum likelihood estimator. Assumption 5(ii) is also required to ensure that the conclusions of Theorems 6 and 7 hold which characterize the asymptotic distribution of $\mathbf{s}\left(\hat{\mathbf{A}}_n, \hat{\mathbf{B}}_n\right)$.

A sufficient but not necessary condition for both Assumption 5(i) and Assumption 5(ii) to hold is that $\log f$ is thrice continuously differentiable on the compact set $\Theta$, measurable in its first argument (e.g., piecewise continuous), and the support of $\boldsymbol{X}$ is bounded. The assumption that the support of $\boldsymbol{X}$ is bounded is satisfied, for example, by observational data consisting of discrete random variables. Assumptions 5(i) and Assumption 5(ii) more generally are satisfied for many commonly used finite-dimensional parametric smooth probability models for observational data modeled as combinations of both discrete and absolutely continuous random variables.

Assumption 5(iii) in conjunction with Assumptions 5(i) and Assumption 5(ii) is used in Theorem 4 to ensure that: (1) when $\mathbf{A}^* \neq \mathbf{B}^*$ this corresponds to the case of model misspecification; and (2) the correctly specified probability model implies that $\mathbf{A}^* = \mathbf{B}^*$. Thus, Assumption 5(iii) is important for ensuring the proper semantic interpretation of a GIMT result (see Proposition 1 and Theorem 4). In addition, Assumption 5(iii) in conjunction with Assumption 5(i) and Assumption 5(ii) is also used to ensure that the Lancaster-Chesher approximation holds (see Theorem 8), which provides a method for constructing GIMTs without computing the third derivative of the negative log-likelihood function.

Assumption 5(iii) can be interpreted as stating that the density $f(\mathbf{x};\theta)$ in the probability model and the data generating process density $p_x(\mathbf{x})$ cannot be too dissimilar. A sufficient but not necessary condition for satisfying Assumption 5(iii) would be that there exists two finite positive numbers $K_1$ and $K_2$ such that for all $\theta \in \Theta$ and for all $\boldsymbol{x} \in$ supp $\boldsymbol{X}$: $f(\mathbf{x};\theta) < K_1$ and $p_x(\mathbf{x}) > K_2$. Although Assumption 5(iii) could be formulated in a slightly more general manner, we use this more specialized version for expository reasons.

The *negative average log-likelihood* is defined as:

$$\bar{l}_n(\theta) \equiv -n^{-1}\sum_{i=1}^{n} \log f(\mathbf{X}_i;\theta).$$

When it exists, the unique global minimizer of $\bar{l}_n(\theta)$ is called the *quasi-maximum likelihood estimate* $\hat{\theta}_n$ rather than a *maximum likelihood estimate* to allow for the possibility that $f$ may be misspecified [1].

The *negative expected log-likelihood* is defined as:

$$l(\theta) \equiv -\int p_x(\mathbf{x}) \log f(\mathbf{x};\theta)\, d\nu(\mathbf{x}).$$

A global minimizer of $l(\theta)$ is called the *pseudo-true parameter value* $\theta^*$ because of the possibility that $f$ may be misspecified. If there exists a $\theta_o$ such that $f(\cdot;\theta_0) = p_x\ \nu(\mathbf{x})$ almost everywhere, then $\theta_o$ is called a *true parameter value*.

**Assumption 6.** ***Uniqueness.*** *(i) For some $\theta^* \in \Theta$, $l$ has a unique minimum at $\theta^*$; (ii) $\theta^*$ is interior to $\Theta$.*

Let $H_0 : \mathbf{s}(\mathbf{A}^*, \mathbf{B}^*) = \mathbf{0}_r$ be a particular GIMT null hypothesis specified by a given GIMT hypothesis function $\mathbf{s}$. Our ultimate goal is to construct a statistical test for testing the GIMT null hypothesis $H_0 : \mathbf{s}(\mathbf{A}^*, \mathbf{B}^*) = \mathbf{0}_r$ by characterizing the asymptotic behavior of the test statistic $\hat{\mathbf{s}}_n \equiv \mathbf{s}\left(\hat{\mathbf{A}}_n, \hat{\mathbf{B}}_n\right)$. Note that the GIMT hypothesis function test statistic $\hat{\mathbf{s}}_n \equiv \mathbf{s}\left(\hat{\mathbf{A}}_n, \hat{\mathbf{B}}_n\right)$ is an estimator of $\mathbf{s}^* \equiv \mathbf{s}(\mathbf{A}^*, \mathbf{B}^*)$ (see Theorem 6).

Let $\nabla \mathbf{s}^* \equiv \nabla \mathbf{s}(\mathbf{A}^*, \mathbf{B}^*)$. Let $\nabla \hat{\mathbf{s}}_n \equiv \nabla \mathbf{s}\left(\hat{\mathbf{A}}_n, \hat{\mathbf{B}}_n\right)$.

Let $\delta^*(\mathbf{X}_i) \equiv \nabla \mathbf{s}^* \mathcal{D}_k^{\otimes}\left(\mathbf{d}_{x,\theta}(\mathbf{X}_i;\theta^*) - \nabla \mathbf{d}^*(\mathbf{A}^*)^{-1}\mathbf{g}(\mathbf{X}_i;\theta^*) - \mathbf{d}^*\right)$.

Given appropriate regularity conditions, it will be shown (see Theorems 6 and 7) that the asymptotic covariance matrix of $n^{1/2}(\hat{\mathbf{s}}_n - \mathbf{s}^*)$ is the *GIMT asymptotic covariance matrix*

$$\Sigma_s^* \equiv \int \delta^*(\mathbf{x}_i)\left(\delta^*(\mathbf{x}_i)\right)^T p_x(\mathbf{x}_i)\, d\nu(\mathbf{x}_i), \tag{1}$$

which may be estimated by

$$\hat{\Sigma}_s^n \equiv (1/n)\sum_{i=1}^{n} \hat{\delta}_n(\mathbf{X}_i)\left(\hat{\delta}_n(\mathbf{X}_i)\right)^T, \tag{2}$$

where $\hat{\delta}_n(\mathbf{X}_i) \equiv \nabla \hat{\mathbf{s}}_n \mathcal{D}_k^{\otimes}\left(\mathbf{d}_{x,\theta}\left(\mathbf{X}_i;\hat{\theta}_n\right) - \nabla \hat{\mathbf{d}}_n(\hat{\mathbf{A}}_n)^{-1}\mathbf{g}\left(\mathbf{X}_i;\hat{\theta}_n\right) - \hat{\mathbf{d}}_n\right)$.

**Assumption 7.** ***Positive Definiteness.*** *(i) $\mathbf{A}^*$ is positive definite; (ii) $\mathbf{B}^*$ is positive definite; and (iii) $\sum_s^*$ is positive definite.*

Assumption 7(i) is a sufficient but not necessary condition for the quasi-maximum likelihood estimate to be a strict local minimizer. Assumption 7(ii) is used in order to apply the Multivariate Central Limit Theorem to characterize the asymptotic distribution of the quasi-maximum likelihood estimates. Assumption 7(iii) is used in order to apply the Multivariate Central Limit Theorem to obtain the asymptotic distribution of the GIMT statistic $\hat{\mathbf{s}}_n$. Violation of Assumption 7 is analogous to the presence of multicollinearity in classical linear regression modeling.

Assumptions 6, 7(i), and 7(ii) are often checked in practice by checking if the infinity norm of $\hat{\mathbf{g}}_n$ is sufficiently small and that the condition numbers of $\hat{\mathbf{A}}_n$ and $\hat{\mathbf{B}}_n$ are not excessively large. In addition, it is necessary to check that the condition number of an estimator of $\sum_s^*$ denoted by $\hat{\sum}_s^n$ (see Equation (2)) is not excessively large. Note that Assumption 4(iv) is a necessary condition for $\sum_s^*$ to be positive definite. If the magnitude of the asymptotic covariance matrix of the selection test statistic $\hat{\mathbf{s}}_n \equiv \mathbf{s}\left(\hat{\mathbf{A}}_n, \hat{\mathbf{B}}_n\right)$, $\sum_s^*$, is not finite or $\sum_s^*$ is singular, then Assumption 7(iii) fails.

## 3. GIMT Theoretical Framework: Theorems and Formulas

In this section, a brief review of relevant results from classical asymptotic theory is provided (Theorems 1, 2, 3, 4, 5, 8) in conjunction with our new results in Theorems 6 and 7. Proofs of all theorems and propositions are provided in the Appendix A.

### *3.1. Classical Results*

**Theorem 1.** ***Estimator Measurability ([30], Lemma 2).*** *Assume that Assumptions 1, 2, 3(i), and 3(ii) hold. Let $P_0^n$ be the joint distribution of $\mathbf{X}_1, ..., \mathbf{X}_n$. Then for each $n = 1, 2, \ldots$, there exists a measurable function $\hat{\theta}_n : \mathcal{R}^{dn} \to \Theta$ and an element, $B_n$, of $\left(\mathcal{B}\left(\mathcal{R}^d\right)\right)^n$ with $P_0^n(B_n) = 1$ such that for all $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in B_n$:*

$$\bar{l}_n\left(\hat{\theta}_n\left(\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}\right)\right) = \min_{\theta \in \Theta} \bar{l}_n(\theta).$$

**Theorem 2.** ***Estimator Consistency ([31], Theorem 2.1).*** *Assume Assumptions 1, 2, 3(i), 3(ii), 5(i)a, and 6 hold. Then as $n \to \infty$, $\hat{\theta}_n \to \theta^*$ with probability one.*

**Theorem 3.** ***Estimator Asymptotic Distribution ([1], Theorem 3.2; also see [32]).*** *Assume Assumptions 1, 2, 3(i), 3(ii), 3(iii), 3(iv), 5(i), 6, 7(i), and 7(ii) hold. As $n \to \infty$, $\sqrt{n}\left(\hat{\theta}_n - \theta^*\right)$ converges in distribution to a zero-mean Gaussian random vector with non-singular covariance matrix $\mathbf{C}^* \equiv (\mathbf{A}^*)^{-1} \mathbf{B}^* (\mathbf{A}^*)^{-1}$.*

**Theorem 4.** ***Contrapositive Information Matrix Equality ([1], Theorem 3.3).*** *Assume Assumptions 1, 2, 3(i), 3(ii), 3(iii), 3(iv), 5, and 6 hold. If $\mathbf{A}^* \neq \mathbf{B}^*$, then the probability model $\mathcal{M} \equiv \left\{F(\cdot; \theta) : \mathcal{R}^d \to [0,1] \,|\, \theta \in \Theta\right\}$ is misspecified.*

Theorem 4 is the contrapositive statement of the familiar information matrix equality that states that if a smooth regular probability model is correctly specified, then $\mathbf{A}^* = \mathbf{B}^*$. The contrapositive statement implies that a difference between $\mathbf{A}^*$ and $\mathbf{B}^*$ indicates the presence of model misspecification.

Moreover, if the information matrix equality is violated (i.e., $\mathbf{A}^* \neq \mathbf{B}^*$), then the asymptotic distribution of the quasi-maximum likelihood estimator is still Gaussian centered at $\theta^*$ but its asymptotic covariance matrix is $\mathbf{C}^* \equiv (\mathbf{A}^*)^{-1} \mathbf{B}^* (\mathbf{A}^*)^{-1}$. In this case, the standard formulas for estimating the asymptotic covariance matrix of the maximum likelihood estimators based upon estimating either $(\mathbf{A}^*)^{-1}$ or $(\mathbf{B}^*)^{-1}$ are not appropriate. Thus, detecting that $\mathbf{A}^* \neq \mathbf{B}^*$ is not only useful for detecting model misspecification but also detects situations where the sandwich covariance matrix estimator $\hat{\mathbf{C}}_n \equiv \left(\hat{\mathbf{A}}_n\right)^{-1} \hat{\mathbf{B}}_n \left(\hat{\mathbf{A}}_n\right)^{-1}$ should be used to ensure an asymptotically unbiased estimate of $\mathbf{C}^* \equiv (\mathbf{A}^*)^{-1} \mathbf{B}^* (\mathbf{A}^*)^{-1}$ is obtained. This is important in applications when one encounters predictive, yet misspecified, models. For example, a linear regression model may have small residual errors yet the residual error term is not Gaussian.

Let $\overline{\mathbf{C}}_n(\theta) \equiv \left(\overline{\mathbf{A}}_n\right)^{-1} \overline{\mathbf{B}}_n(\theta) \left(\overline{\mathbf{A}}_n(\theta)\right)^{-1}$. Let $\hat{\mathbf{C}}_n \equiv \overline{\mathbf{C}}_n\left(\hat{\theta}_n\right)$.

**Theorem 5.** ***Consistent QMLE Covariance Matrix Estimators (e.g., [1]).*** *Assume Assumptions 1, 2, 3(i), 3(ii), 3(iii), 3(iv), 5(i), 6, 7(i) and 7(ii) hold. Then, with probability one as $n \to \infty$: $\hat{\mathbf{B}}_n \to \mathbf{B}^*$, $\left(\hat{\mathbf{B}}_n\right)^{-1} \to (\mathbf{B}^*)^{-1}$, $\hat{\mathbf{A}}_n \to \mathbf{A}^*$, $\left(\hat{\mathbf{A}}_n\right)^{-1} \to (\mathbf{A}^*)^{-1}$, $\hat{\mathbf{C}}_n \to \mathbf{C}^*$, and $\left(\hat{\mathbf{C}}_n\right)^{-1} \to (\mathbf{C}^*)^{-1}$.*

### *3.2. GIMT Statistic Asymptotic Behavior*

**Theorem 6.** ***GIMT Statistic Consistency.*** *Assume Assumptions 1, 2, 3, 4(i), 4(ii), 4(iii), 5(i), and 6 hold. Then as $n \to \infty$, $\hat{\mathbf{s}}_n \to \mathbf{s}^*$ with probability one. If, in addition, Assumptions 5(ii) and 7(iii) hold, then with probability one $\hat{\sum}_s^n \to \sum_s^*$ and $\left(\hat{\sum}_s^n\right)^{-1} \to \left(\sum_s^*\right)^{-1}$ as $n \to \infty$.*

The asymptotic distribution of $\hat{\mathbf{s}}_n \equiv \mathbf{s}\left(\hat{\mathbf{A}}_n, \hat{\mathbf{B}}_n\right)$ is described in the next theorem. Strategies for estimating $\sum_s^*$ are discussed at the end of this section.

**Theorem 7.** ***Generalized Information Matrix Wald Test.*** *Assume Assumptions 1, 2, 3, 4, 5(i), 5(ii), 6, 7 hold with respect to a GIMT hypothesis function* $\mathbf{s}: \Upsilon \times \Upsilon \rightarrow \mathcal{R}^r$ *and probability model* $\mathcal{M}$. *Let* $\hat{\mathcal{W}}_n \equiv n\left(\hat{\mathbf{s}}_n\right)^T \left(\hat{\Sigma}_s^n\right)^{-1} \left(\hat{\mathbf{s}}_n\right)$. *If* $H_0: \mathbf{s}^* = \mathbf{0}_r$, *then* $\hat{\mathcal{W}}_n \xrightarrow{d} \chi_r^2$ *as* $n \rightarrow \infty$. *If* $H_0: \mathbf{s}^* = \mathbf{0}_r$ *is false, then* $\hat{\mathcal{W}}_n \rightarrow \infty$ *as* $n \rightarrow \infty$ *w.p.1.*

Using a Wald test approach, Theorem 7 establishes that the GIMT $p$-value will be consistently estimated under the null hypothesis $H_0: \mathbf{s}^* = \mathbf{0}_r$ thus allowing us to bound Type 1 errors by chosen significance levels. Under the alternative hypothesis $H_a: \mathbf{s}^* \neq \mathbf{0}_r$, Theorem 7 ensures that the Type 2 error goes to zero as the sample size increases with probability one.

From Theorem 4 and the definition of a GIMT Hypothesis Function $\mathbf{s}: \Upsilon \times \Upsilon \rightarrow \mathcal{R}^r$, it follows that $\mathbf{s}\left(\mathbf{A}^*, \mathbf{B}^*\right) \neq \mathbf{0}_r$ implies the presence of model misspecification. This statement follows immediately from the definition of a GIMT hypothesis function and the conclusion of Theorem 4. It is formally presented because of its semantic importance.

**Proposition 1.** ***Interpretation of GIMT Null and Alternative Hypotheses.*** *Suppose the Assumptions of Theorem 4 hold. Let* $\mathbf{s}$ *be a GIMT hypothesis function. (i) If* $\mathcal{M}$ *is correctly specified, then* $H_0: \mathbf{s}^* = \mathbf{0}_r$; *(ii) If* $H_0: \mathbf{s}^* = \mathbf{0}_r$ *is false, then* $\mathcal{M}$ *is misspecified.*

Proposition 1 states that for either a directional or nondirectional GIMT, evidence supporting the rejection of the null hypothesis $H_0: \mathbf{s}^* = \mathbf{0}_r$ is also evidence supporting the presence of model misspecification. Note, however, the assertion that $H_0: \mathbf{s}^* = \mathbf{0}_r$ is true does not necessarily imply correct model specification.

### *3.3. GIMT Covariance Matrix Estimators*

A *non-directional GIMT covariance matrix estimator* $\ddot{\Sigma}_s^n$ is defined as an estimator with the following two properties: (i) $\ddot{\Sigma}_s^n \rightarrow \Sigma_s^*$ as $n \rightarrow \infty$ with probability one when $\mathbf{A}^* = \mathbf{B}^*$; and (ii) $\ddot{\Sigma}_s^n$ converges to a positive definite matrix as $n \rightarrow \infty$ with probability one regardless of whether the probability model is correctly specified. Property (ii) is analogous to Assumption 7(iii) and can be empirically checked by examining the condition number of the GIMT covariance matrix estimator $\ddot{\Sigma}_s^n$.

Let the *Lancaster-Chesher 3rd Derivative Formula* $\ddot{\nabla}\mathbf{d}: \Theta \rightarrow \mathcal{R}^{k(k+1)\times k}$ be defined such that:

$$\ddot{\nabla}\mathbf{d}_n(\theta) = \mathcal{D}_k^{\otimes\dagger} \begin{bmatrix} \frac{d\overline{\mathbf{B}}_n(\theta)}{d\theta} + n^{-1}\sum_{i=1}^{n} vec\left(\mathbf{A}\left(\mathbf{X}_i;\theta\right) - \mathbf{B}\left(\mathbf{X}_i;\theta\right)\right)\left(\mathbf{g}\left(\mathbf{X}_i;\theta\right)\right)^T \\ \frac{d\overline{\mathbf{B}}_n(\theta)}{d\theta} \end{bmatrix}, \tag{3}$$

where

$$\frac{d\overline{\mathbf{B}}_n(\theta)}{d\theta} = n^{-1}\sum_{i=1}^{n}\left[\left(\mathbf{A}\left(\mathbf{X}_i;\theta\right)\otimes\mathbf{g}\left(\mathbf{X}_i;\theta\right)\right) + \left(\mathbf{g}\left(\mathbf{X}_i;\theta\right)\otimes\mathbf{A}\left(\mathbf{X}_i;\theta\right)\right)\right]. \tag{4}$$

The formulas for the GIMT covariance matrix estimator require computation of both the second and third derivatives of the negative log-likelihood function, which are represented in Equations (1) and (2) by the formula $\nabla\mathbf{d}^*$. Theorem 8 shows that the formula $\ddot{\nabla}\hat{\mathbf{d}}_n \equiv \ddot{\nabla}\mathbf{d}_n\left(\hat{\theta}_n\right)$, which uses only first and second derivatives of the negative average log-likelihood, may be used to asymptotically approximate $\nabla\mathbf{d}^*$ for the purpose of avoiding calculation of negative average log-likelihood third derivatives.

**Theorem 8.** ***Lancaster-Chesher Estimator (see [12]).*** *Assume Assumptions 1, 2, 3, 5(i)a, 5(i)c, 5(i)d, 5(ii)a, 5(ii)c, 5(iii), and 6 hold with respect to a GIMT hypothesis function* $\mathbf{s}: \Upsilon \times \Upsilon \rightarrow \mathcal{R}^r$ *and probability model* $\mathcal{M}$. *If* $\mathcal{M}$ *is correctly specified, then with probability one* $\ddot{\nabla}\mathbf{d}_n\left(\hat{\theta}_n\right) \rightarrow \nabla\mathbf{d}^*$ *as* $n \rightarrow \infty$.

Theorem 8 provides an additional mechanism for constructing alternative and possibly computationally convenient covariance matrix estimators for estimating $\Sigma_s^*$ when the null hypothesis

that the model is correctly specified holds. In particular, the formula $\ddot{\nabla}\mathbf{d}_n\left(\hat{\theta}_n\right)$ is substituted for $\nabla\hat{\mathbf{d}}_n$ in Equation (2) to obtain a real symmetric matrix with non-negative eigenvalues called the *Lancaster-Chesher covariance matrix estimator*. If the null hypothesis that the model is correctly specified is false, then the Lancaster-Chesher covariance matrix estimator simply needs to converge to any finite positive definite matrix. This latter assumption can be empirically checked by examining the condition number of the Lancaster-Chesher covariance matrix estimator.

We now provide formulas for a variety of different types of non-directional GIMT covariance matrix estimators. First note that when the probability model is correctly specified, the contrapositive of Theorem 4 in conjunction with Theorem 5 implies that $\hat{\mathbf{A}}^{-1} = \hat{\mathbf{B}}^{-1} = \hat{\mathbf{C}}$. Thus, one can use either the OPG inverse Hessian estimator $\hat{\mathbf{B}}^{-1}$ or the sandwich inverse Hessian estimator $\hat{\mathbf{C}}_n$ as alternative estimators for the inverse Hessian estimator $\hat{\mathbf{A}}^{-1}$ in (2). Second, if the GIMT selection function $\mathbf{s}$ is anti-symmetric and $\mathbf{A}^* = \mathbf{B}^*$, then it follows that the term $(\nabla\mathbf{s}^*)\,\mathcal{D}_k^{\otimes}\mathbf{d}^* = \mathbf{0}_r$ so that the *centering term* $\mathbf{d}^*$ in (1) can be set equal to $\mathbf{0}_r$. Thus, an alternative estimator of $\mathbf{d}^*$ that can be used instead of the *centering term estimator* $\hat{\mathbf{d}}_n$ in (2) is simply a vector of zeros. These two methods yield six different non-directional GIMT covariance matrix estimators.

Six additional GIMT covariance matrix estimators can be obtained by using the Lancaster-Chesher estimator $\ddot{\nabla}\hat{\mathbf{d}}_n$ (defined above) as an alternative estimator for the third derivative negative average log-likelihood estimator $\nabla\hat{\mathbf{d}}_n$. The Lancaster-Chesher estimator $\ddot{\nabla}\hat{\mathbf{d}}_n$ has the computational advantage relative to $\nabla\hat{\mathbf{d}}_n$ that only the first and second derivatives of the negative log-likelihood are used. However, previous empirical studies have suggested that the use of the Lancaster-Chesher estimator $\ddot{\nabla}\hat{\mathbf{d}}_n$ instead of the third-derivative negative average log-likelihood estimator $\nabla\hat{\mathbf{d}}_n$ may degrade performance in some cases (e.g., [13,15–18]).

### *3.4. Adjusted GIMT Hypothesis Functions*

Assumption 7(iii) requires that $\Sigma_s^*$ is a positive definite matrix. The GIMT selection function $\mathbf{s}$ may have the property that the $r$-dimensional matrix $\Sigma_s^*$ is singular with rank $g$ where $g < r$ so that Assumption 7(iii) fails. However, it is often possible to replace the original GIMT hypothesis function $\mathbf{s} : \Upsilon \times \Upsilon \to \mathcal{R}^r$ with an *alternative "adjusted" GIMT hypothesis function* $\mathbf{s}' : \Upsilon \times \Upsilon \to \mathcal{R}^g$ that tests a similar null hypothesis yet has the properties that: (i) the resulting asymptotic covariance matrix of $n^{1/2}\mathbf{s}_n'$ is nonsingular; and (ii) rejection of $H_0 : \mathbf{s}'(\mathbf{A}^*, \mathbf{B}^*) = \mathbf{0}_r$ implies rejection of $H_0 : \mathbf{s}(\mathbf{A}^*, \mathbf{B}^*) = \mathbf{0}_r$.

**Proposition 2.** ***Adjusted GIMT Hypothesis Function Properties.*** *Let $\Sigma_s^*$ be an $r$-dimensional GIMT asymptotic covariance matrix for GIMT hypothesis function $\mathbf{s} : \Upsilon \times \Upsilon \to \mathcal{R}^r$ such that Assumption 7(iii) holds. Let the $g$ rows of the rank $g$ matrix $\mathbf{T} \in \mathcal{R}^{g\times r}$ be $r$-dimensional orthonormal eigenvectors of $\Sigma_s^*$ $(r > g \geq 1)$ for GIMT hypothesis function $\mathbf{s}$. Define an alternative GIMT hypothesis function $\mathbf{s}' \equiv \mathbf{Ts}$ whose respective $g$-dimensional GIMT asymptotic covariance matrix is $\Sigma_T^* = \mathbf{T}\Sigma_s^*\mathbf{T}^T$. (i) If $H_0 : \mathbf{s}'(\mathbf{A}^*, \mathbf{B}^*) = \mathbf{0}_g$ is false, then $H_0 : \mathbf{s}(\mathbf{A}^*, \mathbf{B}^*) = \mathbf{0}_r$ is false; (ii) The $g$-dimensional GIMT asymptotic covariance matrix, $\Sigma_T^*$, for $\mathbf{s}'$ is finite and positive definite.*

The matrix $\mathbf{T}$ in Proposition 1 is called the *adjusted GIMT hypothesis projection matrix*. The proof of Proposition 2(i) follows from the observation that if $|\mathbf{s}|^2 = \mathbf{0}_r$, then $|\mathbf{s}'|^2 = \mathbf{0}_g$. Proposition 2(ii) follows from the observation that $\Sigma_T^* = \mathbf{T}\Sigma_s^*\mathbf{T}^T$ is non-singular by the construction of $\mathbf{T}$ and Assumption 7(iii).

## 4. Simulation Studies

As discussed, some previously published information matrix tests for model misspecification have demonstrated good level and power performance (e.g., [19,23,24]). These tests may be viewed with respect to the GIMT framework presented here. The theoretical framework presented in Sections 2 and 3 provides an important perspective in understanding the similarities and differences among existing misspecification tests within a unified framework. Further, these prior published empirical

studies support the value of the GIMT framework by showing that GIMTs with good level and power performance can be constructed.

However, the GIMT framework in Sections 2 and 3 is also valuable for developing entirely new GIMTs for a large class of probability models in a straightforward manner through the use of Theorems 6 and 7. To illustrate our approach to the construction and evaluation of such GIMTs, we show how Theorems 6 and 7 can be used to derive five new GIMTs. Although an important goal of these derivations was an interest in developing useful tests for model misspecification, a major reason for deriving five additional GIMTs was to demonstrate the flexibility and generality of the unified GIMT theory developed in Sections 2 and 3 .

Next, simulation studies of the level and power performance of the new GIMTs are provided to examine the performance of the GIMTs for some specific empirical examples. This particular logistic regression modeling problem is intended to be representative of a commonly encountered situation where a relevant predictor in a regression model is not properly recoded and an irrelevant predictor is included. The simulation studies were not intended to be comprehensive but rather were designed to empirically demonstrate how the general GIMT theory (Sections 2 and 3) can be used to develop a wide range of misspecification tests. For comparison purposes, the Adjusted Classical GIMT originally proposed by Golden et al. [23] was included as a sixth GIMT in the simulation studies.

### *4.1. Generalized Information Matrix Tests*

#### 4.1.1. Adjusted Classical GIMT (Directional) [23]

Suppose one desires to test the classical full Information Matrix Test hypothesis $H_0 : \mathbf{A}^* = \mathbf{B}^*$. Let $\Sigma_s^*$ be the $r$-dimensional GIMT asymptotic covariance matrix associated with this GIMT. Note that $r = k(k+1)/2$ may be relatively large. Assume, however, that $\Sigma_s^*$ only has rank $g$ where $g < r$. Because $\Sigma_s^*$ is not of full rank, the asymptotic theory developed here cannot be directly applied since Assumption 7(iii) is violated. However, following the discussion in Proposition 1, let $\mathbf{T} \in \mathcal{R}^{g \times r}$ be a matrix with full row rank defined such that the $g$ rows of $\mathbf{T}$ are $r$-dimensional orthonormal eigenvectors of $\Sigma_s^*$ ($r > g \geq 1$). Then, instead of testing the null hypothesis $H_0 : \mathbf{A}^* = \mathbf{B}^*$ associated with the classical full non-directional Information Matrix Test [1], the null hypothesis $H_0 : \mathbf{T}vech(\mathbf{A}^*) = \mathbf{T}vech(\mathbf{B}^*)$ is tested using the GIMT hypothesis function $\mathbf{s}$ defined such that: $\mathbf{s}(\mathbf{A}^*, \mathbf{B}^*) = \mathbf{T}vech(\mathbf{A}^* - \mathbf{B}^*)$. The GIMT associated with this hypothesis function is called the *Adjusted Classical GIMT (Directional).* Golden et al. [23] provided further discussion of this GIMT and showed that it had good level and power properties using simulation studies of a realistic epidemiological data analysis problem.

#### 4.1.2. Fisher Spectra GIMT (Directional)

The *Fisher Spectra GIMT (Directional)* is a new $k$-degree of freedom test specified by the GIMT hypothesis function $\mathbf{s}$ defined such that:

$$\mathbf{s}(\mathbf{A}^*, \mathbf{B}^*) = diag\left((\mathbf{A}^*)^{-1}\mathbf{B}^*\right) - \mathbf{1}_k,$$

which tests the null hypothesis $H_o : \mathbf{s}(\mathbf{A}^*, \mathbf{B}^*) = \mathbf{0}_k$. The notation $\mathbf{1}_k$ denotes a $k$-dimensional column vector of ones. The $diag : \mathcal{R}^{k \times k} \to \mathcal{R}^k$ is defined such that $diag\left((\mathbf{A}^*)^{-1}\mathbf{B}^*\right)$ is a column vector of the on-diagonal elements of $(\mathbf{A}^*)^{-1}\mathbf{B}^*$. The degrees of freedom of this test are equal to the number of free parameters in the model. When the Information Matrix Equality holds, then $(\mathbf{A}^*)^{-1}\mathbf{B}^*$ will be the identity matrix and this GIMT tests the null hypothesis that the $k$ on-diagonal elements of $(\mathbf{A}^*)^{-1}\mathbf{B}^*$ are all equal to one. Note that the Fisher Spectra GIMT tests that the eigenvalues of the two matrices are the same but does not test the null hypothesis that the two matrices have the same eigenvectors. The Fisher Spectra GIMT presented here is similar to the Copula Eigenvalue Test [33]; however, the test statistic is different because the Fisher Spectra GIMT was not developed within a copula framework.

4.1.3. Robust Log GAIC GIMT (Directional)

The *Robust Log GAIC GIMT (Directional)* is a new 1-degree of freedom test specified by the GIMT hypothesis function **s** defined such that:

$$\mathbf{s}\left(\mathbf{A}^{*},\mathbf{B}^{*}\right)=\log\left(\left(\frac{1}{k}\right)trace\left(\left(\mathbf{A}^{*}\right)^{-1}\mathbf{B}^{*}\right)\right),$$

which tests the null hypothesis $H_o : \mathbf{s}\left(\mathbf{A}^{*},\mathbf{B}^{*}\right) = 0$. If the null hypothesis of this test is rejected, then not only does it indicate the presence of model misspecification, it mandates that one uses misspecification-robust estimation methods such as the sandwich estimator [1,32] and misspecification-robust model selection criteria such as the Generalized Akaike Information Criterion (GAIC) [34–36]. The GAIC that is defined by the formula: $GAIC = 2n\hat{l}_n + 2trace\left(\left(\mathbf{A}^{*}\right)^{-1}\mathbf{B}^{*}\right)$ is an unbiased estimator of the expected value of the log-likelihood measure $2n\hat{l}_n$ (e.g., see Appendix of [35]). Note that the Log GAIC GIMT tests the same null hypothesis as the IOS IMT described by Presnell and Boos [19] (also see [20–22]); however, the test statistic is the logarithm of the IOS IMT statistic.

4.1.4. Robust Log GAIC Ratio GIMT (Directional)

The 1-degree of freedom Composite Log GAIC Ratio GIMT is specified by the GIMT hypothesis function **s** is defined such that:

$$\mathbf{s}\left(\mathbf{A}^{*},\mathbf{B}^{*}\right)=\log\left(\frac{trace\left(\left(\mathbf{A}^{*}\right)^{-1}\mathbf{B}^{*}\right)}{trace\left(\left(\mathbf{B}^{*}\right)^{-1}\mathbf{A}^{*}\right)}\right),$$

which tests the null hypothesis $H_o : \mathbf{s}\left(\mathbf{A}^{*},\mathbf{B}^{*}\right) = 0$. The Robust Log GAIC Ratio GIMT (Directional) tests a null hypothesis similar to the null hypotheses associated with the group of non-directional 1 degree of freedom GIMTs discussed by Cho and Phillips [37] that compares the arithmetic mean and harmonic mean of the eigenvalues of the matrix $\left(\mathbf{A}^{*}\right)^{-1}\mathbf{B}^{*}$. It is also closely related to the IOS IMT discussed by Presnell and Boos [19] (also see [20–22]).

4.1.5. Composite Log GAIC GIMT (Nondirectional)

Lemma 1 of [37] shows that $\mathbf{A}^{*} = \mathbf{B}^{*}$ if and only if $trace\left(\left(\mathbf{A}^{*}\right)^{-1}\mathbf{B}^{*}\right) = k$ and $trace\left(\left(\mathbf{B}^{*}\right)^{-1}\mathbf{A}^{*}\right) = k$. This result provides a justification for a new type of GIMT called the 2-degree of freedom *Composite Log GAIC GIMT (Non-Directional).* The Composite Log GAIC GIMT specified by the GIMT hypothesis function **s** is defined such that:

$$\mathbf{s}\left(\mathbf{A}^{*},\mathbf{B}^{*}\right)=\begin{bmatrix}\log\left(\left(\frac{1}{k}\right)trace\left(\left(\mathbf{A}^{*}\right)^{-1}\mathbf{B}^{*}\right)\right)\\ \log\left(\left(\frac{1}{k}\right)trace\left(\left(\mathbf{B}^{*}\right)^{-1}\mathbf{A}^{*}\right)\right)\end{bmatrix},$$

which tests the null hypothesis $H_o : \mathbf{s}\left(\mathbf{A}^{*},\mathbf{B}^{*}\right) = \mathbf{0}_2$. The Composite Log GAIC GIMT (Non-Directional) tests a null hypothesis similar to the null hypotheses associated with the group of non-directional 1 degree of freedom GIMTs discussed by Cho and Phillips [37] that compare the arithmetic mean and harmonic mean of the eigenvalues of the matrix $\left(\mathbf{A}^{*}\right)^{-1}\mathbf{B}^{*}$.

4.1.6. Composite GAIC GIMT (Non-Directional)

The *Composite GAIC GIMT (Non-Directional)* tests exactly the same null hypothesis as the Composite Log GAIC GIMT but does not include the log transformation. The Composite GAIC GIMT specified by the GIMT hypothesis function **s** is defined such that:

$$\mathbf{s}\left(\mathbf{A}^*, \mathbf{B}^*\right) = \begin{bmatrix} \left(\frac{1}{k}\right) trace\left(\left(\mathbf{A}^*\right)^{-1}\mathbf{B}^*\right) - 1 \\ \left(\frac{1}{k}\right) trace\left(\left(\mathbf{B}^*\right)^{-1}\mathbf{A}^*\right) - 1 \end{bmatrix},$$

which tests the null hypothesis $H_o : \mathbf{s}\left(\mathbf{A}^*, \mathbf{B}^*\right) = \mathbf{0}_2$. Cho and Phillips [37] have proposed the magnitude of the Composite Log GAIC GIMT as a 1-degree of freedom non-directional GIMT. Note that this GIMT is also closely related to the IOS test of [19].

*4.2. Methods*

4.2.1. Simulated Data Generating Processes

The level and power performance of the six GIMTs are tested using simulation methods described in [23]. First, five data samples, consisting of 1000, 2000, 4000, 8000, and 16,000 exemplars respectively, were created by random sampling a value $x_1$ from a uniform density on the interval $[-1, 1]$ and sampling a value of $x_2$ from a binomial density. A response variable for each exemplar was randomly generated from the predictor $x_1$ using the "true" data generating process specified by the logistic regression model:

$$\log\left(\frac{p\left(y=1\right)}{p\left(y=0\right)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 {x_1}^2 + \beta_3 {x_1}^3, \tag{5}$$

defined by the true coefficient values: $\beta_0 = -1.98$, $\beta_1 = 4.03$, $\beta_2 = 1.73$, $\beta_3 = 1.15$.

The response variable $y$ is assigned to a value of one if the computed probability is greater than 0.5, and zero otherwise. Note that the four-parameter regression model in (5) is thus called the *correctly specified model*, and is used to re-estimate the true coefficient values using the "true" data generating process that is generated using (5).

We also modeled the same binary response variable in the simulated datasets using an "incorrectly" specified model specified by Equation (6):

$$\log\left(\frac{p\left(y=1\right)}{p\left(y=0\right)}\right) = \beta_0 + \beta_1 {x_1}^3 + \beta_2\sqrt{|x_1|} + \beta_3 x_2. \tag{6}$$

Notice that the parametric forms for the correctly (Equation (5)) and incorrectly (Equation (6)) specified models are the same, except that the incorrectly specified model (Equation (6)) omits $x_1$ and $x_1^2$, and includes an "irrelevant predictor", $x_2$, and an incorrect transformation, $\sqrt{|x_1|}$.

Assume a large dataset is constructed by sampling from the data generating process specified by the model in (5). In the correctly specified case when the parameters of the model in (5) are estimated using the dataset generated by the model in (5), the resulting estimators for $\hat{\mathbf{A}}_n$ and $\hat{\mathbf{B}}_n$ are very similar in magnitude, indicating a lack of evidence of misspecification. On the other hand, in the misspecified case when the parameters of the model in (6) are estimated using the dataset generated by the model in (5), the resulting estimators for $\hat{\mathbf{A}}_n$ and $\hat{\mathbf{B}}_n$ are quite different, evidencing misspecification (see Theorem 4 of this paper).

In practice, researchers often choose the model that best fits the observed data using in-sample (training data) and out-of-sample (test data) log-likelihood based measures. Two models, however, can have equivalent fits to the observed data using either in-sample ($2n\hat{l}_n$) or out-of-sample (GAIC) model fit measures, yet one of the models can be correctly specified while the other model is not. The data generating process and models used in the simulation studies described here are designed to illustrate this important situation.

The model in (5) when fitted to the dataset generated by (5) had approximately the same in-sample fit ($2n\hat{l}_n = 9295.69$) as the in-sample fit ($2n\hat{l}_n = 9295.94$) obtained when model (6) was fitted to the dataset generated by (5). Further, the Discrepancy Risk Model Selection Test [38–43] did not show a significant difference in the model fits for the models in (5) and (6) ($Z = 0.003$, $p = 0.997$).

In addition, using the GAIC [35,36,44] which estimates the out-sample (test data) model fit, the model in (5) when fitted to the dataset generated by (5) had approximately the same out-of-sample fit (GAIC = 9303.6) as the out-sample fit (GAIC = 9305.6) of model (6) to the data set generated by (5). The Discrepancy Risk Model Selection Test [38–43] showed no significant difference in GAIC model fits ($Z = 0.028$, $p = 0.98$). Thus, despite the presence of model misspecification, both the misspecified model and the correctly specified model provide observationally equivalent fits to the observed data, underscoring the importance of checking for model misspecification.

#### 4.2.2. Estimation of Type 1 and Type 2 Error Rates

To evaluate the level and power performance of the six GIMTs, we estimate the percentage of times that each GIMT incorrectly rejected the null hypothesis in the correctly specified case (or GIMT level), and correctly rejected the null hypothesis in the misspecified case (GIMT power). Since the data were simulated from a known data generating process, the computation of these statistics is straightforward.

Throughout these simulation studies, a MLE was defined as a set of parameter values such that the sup norm of the gradient of the negative average log-likelihood evaluated at the MLE was less than $1 \times 10^{-8}$. Further, we avoided fitting models to degenerate simulated data by omitting samples with condition numbers greater than $4.5 \times 10^{14}$ to insure numerical stability. The condition number is defined as the maximum eigenvalue divided by the minimum eigenvalue of the inverse of the Hessian covariance matrix estimator. Each simulation was run until $m = 10{,}000$ simulated data samples of size $n$ was reached. The sample sizes $n$ for the simulated data represented 6.25%, 12.5%, 25%, 50%, and 100% of the original 16,000-member sample.

### *4.3. Results and Discussion*

#### 4.3.1. Type 1 Error Performance

Tables 1 and 2 provide estimated Type 1 errors (i.e., estimated $p$-values using Theorem 7 and Equation (2)) computed using 10,000 simulated data samples for a sample size of $n = 16{,}000$. Empirical level (observed Type 1 error rates) are for pre-specified (nominal) significance levels: 0.01, 0.025, 0.05, and 0.10. The average number of times the null hypothesis was incorrectly rejected by a GIMT in a simulation run was used to estimate the Type 1 error rate. The standard error of the number of times the null hypothesis was incorrectly rejected was defined as the bootstrap sampling error. The average number of times the null hypothesis was incorrectly accepted by a GIMT in a simulation run was used to estimate the Type 2 error rate.

The $p$-values estimated in Table 1 are based upon the exact formula for the GIMT test statistic provided in Equation (2), which uses the third derivative of the log-likelihood function. Table 2 provides estimates of the Type 1 error rate using formulas that do not require the use of third derivatives of the log-likelihood function by using the Lancaster-Chesher third derivative approximation (see Theorem 8) for the Hessian covariance matrix estimator obtained by substituting the formula $\ddot{\nabla}\hat{\mathbf{d}}_n$ as defined in Equations (3) and (4) for $\nabla\hat{\mathbf{d}}_n$ in Equation (2).

Level performance in Tables 1 and 2 was evaluated using the Mean Absolute Deviation (MAD), which is defined as the average absolute deviation between an estimated $p$-value and its theoretical expected asymptotic value. Directional GIMTs showed better performance (MAD = 0.013) than non-directional GIMTs (MAD = 0.44). In addition, the Lancaster-Chesher third derivative approximation method (Table 2) showed better performance (MAD = 0.034) than the analytic third derivative method (Table 1) (MAD = 0.055) for non-directional GIMTs. Level performance for directional GIMTs derived using the Lancaster-Chesher third derivative approximation method

(MAD = 0.017) were comparable to directional GIMTs derived using the analytic third derivative method (MAD = 0.0084).

**Table 1.** Type 1 error performance of GIMTs using the analytic third derivative formula for pre-specified (nominal) significance levels: 0.01, 0.025, 0.05, and 0.10. Level performance for the directional GIMTs was better than level performance for the non-directional GIMTs. Bootstrap simulation standard errors are shown in parentheses. Computed values are for 10,000 simulated data samples for sample size $n$ = 16,000. $df$ = degrees of freedom.

| Generalized Information Matrix Test (GIMT) | Test Type | $p$ = 0.01 | $p$ = 0.025 | $p$ = 0.05 | $p$ = 0.10 |
|---|---|---|---|---|---|
| Adjusted Classical ($\leq$10 df) | Directional | 0.0136 (0.0012) | 0.0308 (0.0017) | 0.0550 (0.0023) | 0.1059 (0.0031) |
| Composite GAIC (2 df) | Non-Directional | 0.0830 (0.0027) | 0.1014 (0.0030) | 0.1225 (0.0032) | 0.1546 (0.0036) |
| Composite Log GAIC (2 df) | Non-Directional | 0.0564 (0.0023) | 0.0742 (0.0026) | 0.0930 (0.0029) | 0.1219 (0.0032) |
| Fisher Spectra (4 df) | Directional | 0.0205 (0.0014) | 0.0337 (0.0018) | 0.0584 (0.0023) | 0.1035 (0.0030) |
| Robust Log GAIC (1 df) | Directional | 0.0185 (0.0013) | 0.0360 (0.0018) | 0.0618 (0.0024) | 0.1144 (0.0031) |
| Robust Log GAIC Ratio (1 df) | Directional | 0.0158 (0.0012) | 0.0335 (0.0018) | 0.0590 (0.0023) | 0.1135 (0.0031) |

**Table 2.** Type 1 error performance of GIMTs using the Lancaster-Chesher third derivative approximation for pre-specified (nominal) significance levels: 0.01, 0.025, 0.05, and 0.10. Like the third derivative method in Table 1, level performance for the directional GIMTs was better than level performance for the non-directional GIMTs. Further, for non-directional GIMTs, level performance of the Lancaster-Chesher third derivative approximation for the non-directional GIMTs was better than using third derivative GIMTs. Bootstrap simulation standard errors are shown in parentheses. Computed values are for 10,000 simulated data samples for sample size $n$ = 16,000. $df$ = degrees of freedom.

| Generalized Information Matrix Test (GIMT) | Test Type | $p$ = 0.01 | $p$ = 0.025 | $p$ = 0.05 | $p$ = 0.10 |
|---|---|---|---|---|---|
| Adjusted Classical ($\leq$10 df) | Directional | 0.0085 (0.0009) | 0.0195 (0.0014) | 0.0409 (0.0020) | 0.0916 (0.0029) |
| Composite GAIC (2 df) | Non-Directional | 0.0662 (0.0024) | 0.0821 (0.0026) | 0.1006 (0.0029) | 0.1259 (0.0032) |
| Composite Log GAIC (2 df) | Non-Directional | 0.0403 (0.0019) | 0.0498 (0.0021) | 0.0646 (0.0023) | 0.0884 (0.0027) |
| Fisher Spectra (4 df) | Directional | 0.0071 (0.0008) | 0.0161 (0.0012) | 0.0264 (0.0015) | 0.0535 (0.0021) |
| Robust Log GAIC (1 df) | Directional | 0.0045 (0.0006) | 0.0138 (0.0011) | 0.0236 (0.0014) | 0.0622 (0.0023) |
| Robust Log GAIC Ratio (1 df) | Directional | 0.0032 (0.0005) | 0.0097 (0.0009) | 0.0285 (0.0016) | 0.0588 (0.0022) |

The improved Type 1 error estimation performance of the directional GIMTs may be due to the fact that the directional GIMT statistics had fewer degrees of freedom and thus reduced variance. One possible explanation for the good level performance of the Lancaster-Chesher third derivative approximation method is that this method uses assumptions that hold under the null hypothesis to derive an alternative GIMT covariance matrix estimator without calculating third derivatives. In simulation studies where the null hypothesis of correct model specification holds, key large sample assumptions of the Lancaster-Chesher third derivative approximation method are satisfied

by construction. This suggests that, in some cases, for the purpose of estimating Type 1 errors, the Lancaster-Chesher method may be appropriate for large sample sizes. On the other hand, Taylor [18] has provided examples where the size properties of the Lancaster-Chesher method are poor.

#### 4.3.2. Level-Power Analyses

The level-power performance of the new GIMTs were investigated by examining how the estimated Type 1 and Type 2 errors varied as a function of test significance level. In particular, for a range of possible significance levels, the estimated power (i.e., percent correct rejections) and estimated Type 1 error (i.e., percent incorrect rejections) can be calculated to obtain a Receiver Operating Characteristic (ROC) curve [14,45–47]. The Area under the ROC (AUROC) is measure of discrimination performance. An AUROC = 1.0 indicates perfect discrimination performance and an AUROC = 0.5 indicates chance discrimination performance [45–47]. Although discrimination performance can vary dramatically as a function of test problem difficulty, this paradigm is useful for comparing discrimination performance of different GIMT statistics with respect to a particular test problem.

Figure 1 shows the Level-power for GIMTs using the analytic 3rd derivative for the inverse Hessian matrix estimator by sample size. With respect to the chosen test problem described in the text, these GIMTs obtain nearly perfect performance in correct rejection of the null hypothesis and correct acceptance of the null hypothesis when the sample size in this simulation study exceeds 4000 exemplars.

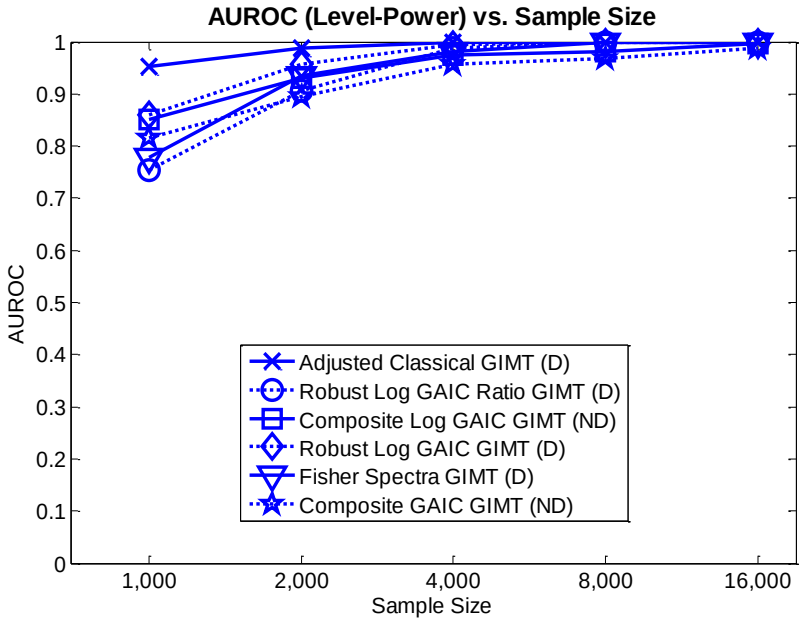


**Figure 1.** Level-power for GIMTs using the analytic 3rd derivative formula is characterized by Area Under the Receiver Operating Characteristic curve (AUROC) as a function of sample size. With respect to the chosen test problem, these GIMTs obtain nearly perfect performance in correct rejection of the null hypothesis and correct acceptance of the null hypothesis when the sample size in this simulation study exceeds 4000 exemplars. Each data point in the above graph was generated from 10,000 bootstrap data samples.

Figure 2 shows the Level-power for GIMTs using the Lancaster-Chesher 3rd derivative approximation. With respect to the chosen test problem these GIMTs obtain excellent performance in correct rejection of the null hypothesis and correct acceptance of the null hypothesis when the sample size in this simulation study is near 16,000 exemplars. However, while the Adjusted Classical GIMT evidences excellent performance across sample sizes in all cases, the other GIMTs show

poor Level-Power performance below 15,000 exemplars with the Lancaster-Chesher 3rd derivative approximation. In addition, with the exception of the Adjusted Classical GIMT, there is not a clear difference in performance between the directional and non-directional tests. These results are consistent with the observations of previous investigators regarding the power performance of the Lancaster-Chesher method (e.g., [14,15,17,18]).

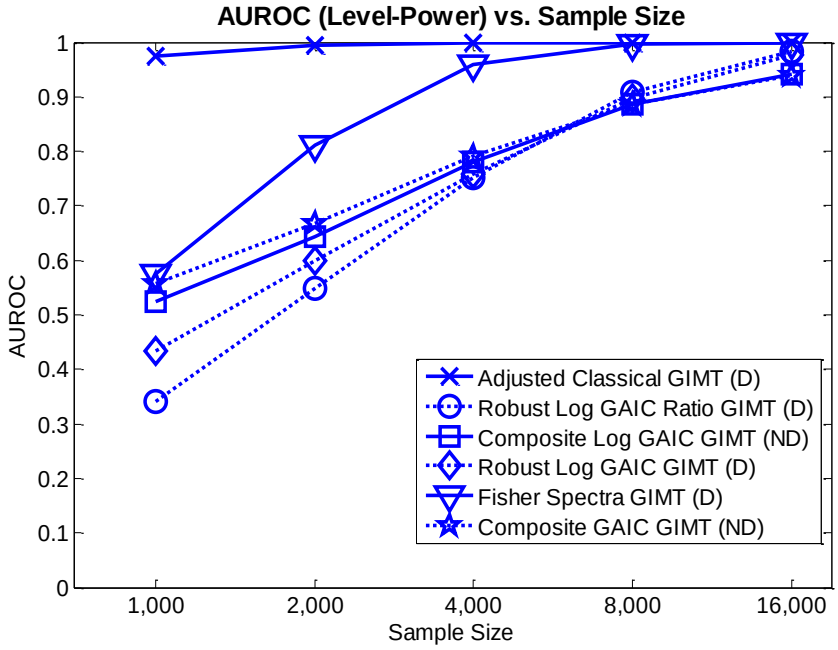


**Figure 2.** Level-power for GIMTs using the Lancaster-Chesher 3rd derivative approximation is characterized by Area Under the Receiver Operating Characteristic curve (AUROC) as a function of sample size. With respect to the chosen test problem these GIMTs obtain excellent performance in correct rejection of the null hypothesis and correct acceptance of the null hypothesis when the sample size in this simulation study is near 16,000 exemplars. While the Adjusted Classical GIMT evidences excellent performance across sample sizes, the other GIMTs show poor Level-Power performance below 15,000 exemplars. Each data point in the above graph was generated from 10,000 bootstrap data samples.

## 5. Conclusions

This paper formally introduces a unified framework for specification testing that is applicable to a wide range of smooth probability models including, for example, the class of generalized linear models (e.g., [48–50]), linear and nonlinear regression (e.g., [51,52]), structural equation models with or without latent variables (e.g., [53,54]), and hierarchical linear models (e.g., [55]). The essential idea is based upon the Contrapositive of the Information Matrix Equality (Theorem 4), which asserts that observed differences between the inverse Hessian covariance matrix estimator $\hat{\mathbf{A}}_n$ and the inverse OPG covariance matrix estimator $\hat{\mathbf{B}}_n$ are indicators of the presence of model misspecification.

Theorem 6 provided explicit conditions for ensuring that $\hat{\mathbf{s}}_n$ converges with probability one to $\mathbf{s}(\mathbf{A}^*, \mathbf{B}^*)$ as $n \to \infty$. Theorem 7 provided explicit conditions for showing that if the null hypothesis $H_0 : \mathbf{s}(\mathbf{A}^*, \mathbf{B}^*) = \mathbf{0}_r$ holds, then a Wald test statistic can be constructed that has an asymptotic chi-squared distribution with $r$ degrees of freedom. If, however, the null hypothesis $H_0 : \mathbf{s}(\mathbf{A}^*, \mathbf{B}^*) = \mathbf{0}_r$ is false, then that same Wald test statistic asymptotically converges to infinity with probability one. Proposition 1 asserts that: (1) if the probability model is correctly specified then $H_0 : \mathbf{s}(\mathbf{A}^*, \mathbf{B}^*) = \mathbf{0}_r$, and (2) if $H_0 : \mathbf{s}(\mathbf{A}^*, \mathbf{B}^*) = \mathbf{0}_r$ is false then the probability model is misspecified.

In the simulation studies, each of the new directional and non-directional GIMTs exhibited excellent level-power performance using the third derivative formulas for the GIMT covariance matrix

estimator. However, performance in estimating the Type 1 error rate varied for different GIMTs, indicating the importance of simulation studies for characterizing the performance of new GIMTs derived within the GIMT framework. In fact, the performance of the directional GIMTs was better than the non-directional GIMTs. The simulation studies also showed that the level-power performance of the GIMTs declined with smaller sample sizes for the Lancaster-Chesher third derivative approximation formula. In addition, the appealing level-power performance of the Adjusted Classical GIMT for both the true third derivative and Lancaster-Chesher third derivative approximation suggests that additional research into the development of GIMTs with adjusted covariance matrices as described in Proposition 2 is merited. It is also important to emphasize that the alternative model used in the above power analyses was chosen such that its fit to the observed data was comparable to the fit of the "true" model that generated the data.

In summary, the simulation studies illustrate a general methodology for using the GIMT framework to derive and evaluate new model misspecification tests. We showed that it is possible for an incorrectly specified model to appear to fit the data well, while testing positive for model misspecification (i.e., reject the null hypothesis that the model is correctly specified). To reach proper statistical inferences when interpreting estimates of the parameters to a fitted model, it is critical to consider both model fit and model specification.

In conclusion, a unified GIMT framework has been presented for identifying, classifying, and developing information matrix type statistical tests for the detection of model misspecification for smooth finite-dimensional probability models. This GIMT framework provides a practical and powerful methodology for the development of both directional and non-directional GIMTs for a wide range of smooth probability models. Furthermore, unlike some existing methods for specification testing in logistic regression modeling, the degrees of freedom of the GIMT test statistic do not increase as a function of the number of distinct patterns of predictor variable values, suggesting that GIMTs will have good level and power performance [51,56–58]. In the real world, it is inevitable that model misspecification will manifest itself in different ways for different probability models and in different situations. Accordingly, it is desirable to have a variety of tests for assessing model misspecification as some tests may be more appropriate than others in detecting the presence of model misspecification in different situations.

**Acknowledgments:** This research was made possible by grants from the National Institute of General Medical Sciences (NIGMS) (R43GM114899, PI: S.S. Henley; R43GM106465, PI: S.S. Henley), the National Institute of Mental Health (NIMH) (R43MH105073, PI: S.S. Henley), the National Cancer Institute (NCI) (R44CA139607, PI: S.S. Henley), and the National Institute on Alcohol Abuse and Alcoholism (NIAAA) (R43/R44AA013768, PI: S.S. Henley; R43/R44AA013351, PI: S.S. Henley) under the Small Business Innovation Research (SBIR) program. The authors wish to gratefully acknowledge this support. This paper reflects the authors' views and not necessarily the opinions or views of the NIGMS, NIMH, NCI, or the NIAAA.

**Author Contributions:** The GIMT mathematical framework was developed by Richard M. Golden and Halbert White in collaboration with Steven S. Henley and T. Michael Kashner. Richard M. Golden and Steven S. Henley developed the GIMT algorithms. The simulation studies were designed and implemented by Steven S. Henley, Richard M. Golden, and T. Michael Kashner. Halbert White did not have the opportunity to review the final version of this manuscript due to his untimely passing. Hal was a great friend and colleague who is very much missed.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A. Proofs of Theorems and Propositions

The following matrix definition of dominated by an integrable function is used in the statement and proofs of the theorems in this paper. It is provided for completeness.

**Definition.** ***Dominated by an Integrable Function.*** Let $\mathbf{X}$ be a random d-dimensional real vector defined on a complete probability space $(\Omega, \Im, P)$, where P has Radon-Nikodým density $p$ with respect to a σ-finite measure $\nu_x$. Let $\Theta \subset \mathcal{R}^r$ be a compact set, $r \in \mathbb{N}$. Let $\mathbf{Q} : \mathcal{R}^d \times \Theta \rightarrow \mathcal{R}^{m \times n}$ be a function defined such that each element of $\mathbf{Q}(\mathbf{x}, \cdot)$ is continuous on $\Theta$ for all $\boldsymbol{x} \in supp\ \boldsymbol{X}$, and each element

of $\mathbf{Q}(\cdot, \theta)$ is measurable for all $\theta \in \Theta$. Suppose there exists a function $K : \mathcal{R}^d \to \mathcal{R}^+$ such that each element, $q_{ij}$ of $\mathbf{Q}$: $\left|q_{ij}(\mathbf{x}, \theta)\right| \leq K(\mathbf{x})$ for all $\theta \in \Theta$ and for all $\boldsymbol{x} \in supp\ \boldsymbol{X}$. Also assume that the expected value of $K(\mathbf{X})$ with respect to $p$ is finite. Then $\mathbf{Q}$ is *dominated by an integrable function K on* $\Theta$ *with respect to* $p$.

In some cases, we will abbreviate the statement "dominated by an integrable function $K$ on $\Theta$ with respect to $p$" to the statement "dominated on $\Theta$ with respect to $p$".

In addition, the Dominated Convergence Theorem (e.g., [30], Theorem 2; [2], Theorem A.2.1), Slutsky's Theorem (e.g., [59], p. 19), Mean Value Theorem (e.g., [60], p. 80), Uniform Law of Large Numbers (e.g., [2], Theorem A.2.2), and Multivariate Central Limit Theorem (e.g., [59], Theorem B, p. 28) are used throughout the following discussion.

**Proof of Theorem 1.** See Lemma 2 of [30]. Q.E.D.

**Proof of Theorem 2.** See Theorem 2.1 of [61]. Q.E.D.

**Proof of Theorem 3.** See Theorem 3.2. of [1]. Q.E.D.

**Proof of Theorem 4.** See Theorem 3.3. of [1]. Q.E.D.

**Proof of Theorem 5.** See proof of Theorem 3.3 of [31]. Q.E.D.

**Proof of Theorem 6.** The proof follows the approach to the proof of Theorem 4.1 in [1]. Let $\bar{\mathbf{s}}_n^* = \left(\overline{\mathbf{A}}_n\left(\theta^*\right), \overline{\mathbf{B}}_n\left(\theta^*\right)\right)$.

Using Assumptions 3, 4(i), 4(ii), 4(iii), and the Mean Value Theorem:

$$\bar{\mathbf{s}}_n^* = \mathbf{s}^* + \nabla \ddot{\mathbf{s}}_n \mathcal{D}_k^{\otimes}\left(\overline{\mathbf{d}}_n\left(\theta^*\right) - \mathbf{d}^*\right), \tag{A1}$$

where $\nabla \ddot{\mathbf{s}}_n$ is a matrix defined such that the $m$th row of $\nabla \mathbf{s}$ is evaluated at: $\left(\ddot{\mathbf{A}}_n, \ddot{\mathbf{B}}_n\right)^{(m)} \equiv \lambda_m\left(\overline{\mathbf{A}}_n\left(\theta^*\right), \overline{\mathbf{B}}_n\left(\theta^*\right)\right) + (1-\lambda_m)\left(\mathbf{A}^*, \mathbf{B}^*\right)$ for some $\lambda_m \in (0,1)$, $m = 1, \ldots, r$.

Using Assumptions 3, 4(i), 4(ii), and 4(iii), and the Mean Value Theorem:

$$\hat{\mathbf{s}}_n = \bar{\mathbf{s}}_n^* + \nabla \widehat{\mathbf{s}}_n \mathcal{D}_k^{\otimes} \nabla \widehat{\mathbf{d}}_n\left(\hat{\theta}_n - \theta^*\right), \tag{A2}$$

where $\nabla \widehat{\mathbf{s}}_n \mathcal{D}_k^{\otimes} \nabla \widehat{\mathbf{d}}_n$ is a matrix constructed by evaluating the $m$th row of the matrix-valued function $\nabla \mathbf{s}_n \mathcal{D}_k^{\otimes} \nabla \overline{\mathbf{d}}_n \equiv \left[\nabla \mathbf{s}\left(\overline{\mathbf{A}}_n\left(\theta\right), \overline{\mathbf{B}}_n\left(\theta\right)\right)\right]^T \mathcal{D}_k^{\otimes} \nabla \overline{\mathbf{d}}_n\left(\theta\right)$ at $\widehat{\theta}_n \equiv \gamma_m \hat{\theta}_n + (1-\gamma_m)\,\theta^*$ for some $\gamma_m \in (0,1)$, $m = 1, \ldots, r$.

Substituting (A1) into (A2) gives:

$$\hat{\mathbf{s}}_n - \mathbf{s}^* = \nabla \ddot{\mathbf{s}}_n \mathcal{D}_k^{\otimes}\left(\overline{\mathbf{d}}_n\left(\theta^*\right) - \mathbf{d}^*\right) + \nabla \widehat{\mathbf{s}}_n \mathcal{D}_k^{\otimes} \nabla \widehat{\mathbf{d}}_n\left(\hat{\theta}_n - \theta^*\right). \tag{A3}$$

In addition, $\nabla \widehat{\mathbf{s}}_n = \nabla \mathbf{s}^* + o_p(1)$ because $\nabla \mathbf{s}$ is continuous by Assumptions 4(i), 4(ii), 4(iii) and $\widetilde{\mathbf{A}}_n \to \mathbf{A}^*$ w.p.1 and $\widetilde{\mathbf{B}}_n \to \mathbf{B}^*$ w.p.1 using Theorem 5. By Assumptions 3, 4, 5, and the Uniform Law of Large Numbers, $\nabla \widehat{\mathbf{s}}_n \mathcal{D}_k^{\otimes} \nabla \widehat{\mathbf{d}}_n = \nabla \mathbf{s}^* \mathcal{D}_k^{\otimes} \nabla \mathbf{d}^* + o_p(1)$. Thus, (A3) can be rewritten as:

$$\begin{aligned} \hat{\mathbf{s}}_n - \mathbf{s}^* \quad = & \nabla \mathbf{s}^* \mathcal{D}_k^{\otimes}\left(\overline{\mathbf{d}}_n\left(\theta^*\right) - \mathbf{d}^*\right) + \nabla \mathbf{s}^* \mathcal{D}_k^{\otimes} \nabla \mathbf{d}^*\left(\hat{\theta}_n - \theta^*\right) \\ & + o_p(1)\left(\overline{\mathbf{d}}_n\left(\theta^*\right) - \mathbf{d}^*\right) + o_p(1)\left(\hat{\theta}_n - \theta^*\right) \end{aligned}. \tag{A4}$$

Assumptions 1, 2, 3(i), 3(ii), 5(i)a, and 6 with Theorem 2 imply that $\hat{\theta}_n \to \theta^*$ and Assumptions 1, 2, 3(i), 3(ii), 3(iii), 3(iv), and 5(i) with the Law of Large Numbers imply that $\overline{\mathbf{d}}_n(\theta^*) \to \mathbf{d}^*$ as $n \to \infty$ with probability one. Thus, the right-hand side of (A4) approaches zero as $n \to \infty$ with probability one. The last part of Theorem 6 asserts that $\hat{\sum}_s^n \to \sum_s^*$ and $\left(\hat{\sum}_s^n\right)^{-1} \to \left(\sum_s^*\right)^{-1}$ with probability one, which follows from Assumptions 1, 2, 3, 4, 5(i), 5(ii), 6, 7 and the Uniform Law of Large Numbers. Q.E.D.

**Proof of Theorem 7.** The proof follows the approach to the proof of Theorem 4.1 in [1]. Using Assumptions 3(i), 3(ii), 3(iii), 3(iv) and expand: $n^{-1}\sum_{i=1}^{n}\mathbf{g}(\mathbf{X}_i;\theta)$ about $\theta^*$ and evaluate at $\hat{\theta}_n$ using the Mean Value Theorem to obtain:

$$\hat{\mathbf{g}}_n = n^{-1}\sum_{i=1}^{n}\mathbf{g}(\mathbf{X}_i;\theta^*) + \overline{\mathbf{A}}_n\left(\ddot{\theta}_n\right)\left[\hat{\theta}_n - \theta^*\right] \tag{A5}$$

where $\ddot{\theta}_n$ lies on the chord connecting $\hat{\theta}_n$ and $\theta^*$. By Assumptions 1, 2, 3(i), 3(ii), 5(i)(a), and 6, and Theorem 2: $\hat{\theta}_n \to \theta^*$ with probability one as $n \to \infty$ which implies with the Uniform Law of Large Numbers: $\hat{\mathbf{g}}_n \to \mathbf{0}_q$ with probability one using Assumption 5(i)(b) and $\overline{\mathbf{A}}_n\left(\widehat{\theta}_n\right) \to \mathbf{A}^*$ using with probability one where $\mathbf{A}^*$ is positive definite by Assumption 7(i). By Slutsky's Theorem, (A5), rearranging terms, and multiplying by $n^{1/2}$ we then have:

$$n^{1/2}\left(\hat{\theta}_n - \theta^*\right) = -n^{1/2}\left(\mathbf{A}^*\right)^{-1}\left(n^{-1}\sum_{i=1}^{n}\mathbf{g}(\mathbf{X}_i;\theta^*)\right) + o_p(1). \tag{A6}$$

Multiplying (13) by $n^{1/2}$ and substituting (A6) into Equation (A4) gives:

$$\begin{array}{c} n^{1/2}\left(\hat{\mathbf{s}}_n - \mathbf{s}^*\right) = \nabla\mathbf{s}^*\mathcal{D}_k^{\otimes}\left(n^{1/2}\left(\overline{\mathbf{d}}_n(\theta^*) - \mathbf{d}^*\right) - \nabla\mathbf{d}^* n^{1/2}\left(\mathbf{A}^*\right)^{-1}\left(n^{-1}\sum_{i=1}^{n}\mathbf{g}(\mathbf{X}_i;\theta^*)\right)\right) \\ -\nabla\mathbf{s}^*\mathcal{D}_k^{\otimes}\nabla\mathbf{d}^* o_p(1) + o_p(1)\, n^{1/2}\left(\overline{\mathbf{d}}_n(\theta^*) - \mathbf{d}^*\right) + o_p(1)\, n^{1/2}\left(\hat{\theta}_n - \theta^*\right) \end{array}. \tag{A7}$$

Assumptions 1, 2, 3(i), 3(ii), 3(iii), 3(iv), 5(i), 6, 7(i), and 7(ii) with Theorem 3 imply that $n^{1/2}\left(\hat{\theta}_n - \theta^*\right) = O_p(1)$. Assumptions 1, 2, 3(i), 3(ii), 3(iii), 3(iv), 5(i)c, 5(i)d, and 5(ii)b in conjunction with the Dominated Convergence Theorem imply that the variance of $|d_{x,\theta}(\mathbf{X}_i,\theta^*) - \mathbf{d}^*|$, $VAR\{|d_{x,\theta}(\mathbf{X}_i,\theta^*) - \mathbf{d}^*|\}$, is finite. Thus, by the Markov Inequality, $\left|n^{1/2}\left(\overline{\mathbf{d}}_n(\theta^*) - \mathbf{d}^*\right)\right| = O_p(1)$. Thus, the last three terms on the right-hand side of (A7) converge to zero with probability one.

Pre-multiplying (A7) by the positive definite matrix $\left(\Sigma_s^*\right)^{-1/2}$ gives:

$$\begin{array}{c} n^{1/2}\left(\Sigma_s^*\right)^{-1/2}\left(\hat{\mathbf{s}}_n - \mathbf{s}^*\right) = \\ n^{1/2}\left(\Sigma_s^*\right)^{-1/2}\nabla\mathbf{s}^*\mathcal{D}_k^{\otimes} n^{-1}\sum_{i=1}^{n}\left(\mathbf{d}_{x,\theta}(\mathbf{X}_i;\theta^*) - \nabla\mathbf{d}^*\left(\mathbf{A}^*\right)^{-1}\mathbf{g}(\mathbf{X}_i;\theta^*) - \mathbf{d}^*\right) + o_p(1) \end{array}. \tag{A8}$$

From the definition of $\Sigma_s^*$ and the assumption that $\Sigma_s^*$ is positive definite (see Assumption 7(iii)), the assumption that $\nabla\mathbf{s}^*$ has full row rank (see Assumption 4), and the Multivariate Central Limit Theorem, it follows that the first term on the right-hand side of (A8) converges in distribution to a zero mean $r$-dimensional multivariate Gaussian random vector $\mathcal{Z}_r$ with an identity covariance matrix. By Slutsky's theorem, the right-hand side of (A8) also converges to $\mathcal{Z}_r$ in distribution and is thus bounded in probability.

If $\mathbf{s}^* = \mathbf{0}_r$, then by (A8) $\hat{\mathcal{W}}_n^* \equiv n\left(\hat{\mathbf{s}}_n\right)^T\left(\sum_s^*\right)^{-1}\left(\hat{\mathbf{s}}_n\right)$ converges in distribution to the sum of the squares of $r$ asymptotically normally distributed random variables (e.g., [59], p. 4) and thus has an asymptotic chi-square distribution with $r$ degrees of freedom. If $\mathbf{s}^* \neq \mathbf{0}_r$, then $\hat{\mathcal{W}}_n^* \to n\left(\mathbf{s}^*\right)^T\left(\sum_s^*\right)^{-1}\left(\mathbf{s}^*\right)$ with probability one and thus $\hat{\mathcal{W}}_n^* \to \infty$ with probability one.

Finally, note that since from (A8) $n^{1/2}\left(\hat{\mathbf{s}}_n - \mathbf{s}^*\right)$ converges in distribution and thus is bounded in probability, and since from Theorem 6 $\sum_s^* - \hat{\sum}_s^n = o_p(1)$ we have

$$\hat{\mathcal{W}}_n = \hat{\mathcal{W}}_n^* + n\left(\hat{\mathbf{s}}_n - \mathbf{s}^*\right)^T o_p(1)\left(\hat{\mathbf{s}}_n - \mathbf{s}^*\right) = \hat{\mathcal{W}}_n^* + o_p(1)\left|n^{1/2}\left(\hat{\mathbf{s}}_n - \mathbf{s}^*\right)\right|^2 = \hat{\mathcal{W}}_n^* + o_p(1)\,O_p(1),$$

it follows from Slutsky's Theorem that $\hat{\mathcal{W}}_n^*$ and $\hat{\mathcal{W}}_n$ have the same asymptotic distribution. Q.E.D.

**Proof of Theorem 8.** The proof follows the approach of [12].

$$\begin{array}{l}\frac{d}{d\theta}\left[\int \mathbf{A}(\mathbf{x};\theta) f(\mathbf{x};\theta) d\nu(\mathbf{x})\right] = \\ \quad \int \left[\frac{d\mathbf{A}(\mathbf{x};\theta)}{d\theta} f(\mathbf{x};\theta) + vec\left(\mathbf{A}(\mathbf{x};\theta)\right)\left(\frac{d\log f(\mathbf{x};\theta)}{d\theta}\right) f(\mathbf{x};\theta)\right] d\nu(\mathbf{x})\end{array} \tag{A9}$$

$$\begin{array}{l}\frac{d}{d\theta}\left[\int \mathbf{B}(\mathbf{x};\theta) f(\mathbf{x};\theta) d\nu(\mathbf{x})\right] = \\ \quad \int \left[\frac{d\mathbf{B}(\mathbf{x};\theta)}{d\theta} f(\mathbf{x};\theta) + vec\left(\mathbf{B}(\mathbf{x};\theta)\right)\left(\frac{d\log f(\mathbf{x};\theta)}{d\theta}\right) f(\mathbf{x};\theta)\right] d\nu(\mathbf{x})\end{array} \tag{A10}$$

Differentiation under the integral operator is permitted by Assumptions 3 and 5, and the Dominated Convergence Theorem. Differentiate both sides of $\int f(\mathbf{x};\theta)\, d\nu(\mathbf{x}) = 1$ three times and use (A9) and (A10) to obtain:

$$\begin{array}{l}\int \frac{d\mathbf{A}(\mathbf{x};\theta)}{d\theta} f(\mathbf{x};\theta)\, d\nu(\mathbf{x}) = \int \frac{d\mathbf{B}(\mathbf{x};\theta)}{d\theta} f(\mathbf{x};\theta)\, d\nu(\mathbf{x}) - \\ \quad \int vec\left(\mathbf{A}(\mathbf{x};\theta) - \mathbf{B}(\mathbf{x};\theta)\right)\left(\frac{d\log f(\mathbf{x};\theta)}{d\theta}\right) f(\mathbf{x};\theta)\, d\nu(\mathbf{x})\end{array} \tag{A11}$$

where

$$\frac{d\mathbf{B}(\mathbf{x};\theta)}{d\theta} = \frac{d}{d\theta}\mathbf{g}(\mathbf{x};\theta)\,\mathbf{g}(\mathbf{x};\theta)^T = \left(\mathbf{A}(\mathbf{x};\theta)\otimes \mathbf{g}(\mathbf{x};\theta)\right) + \left(\mathbf{g}(\mathbf{x};\theta)\otimes \mathbf{A}(\mathbf{x};\theta)\right) \tag{A12}$$

If the probability model is correctly specified there exists a $\theta^*$ such that for all $\boldsymbol{x} \in \text{supp}\, \boldsymbol{X}$: $f(\mathbf{x};\theta^*) = p_x(\mathbf{x})\ \nu(\mathbf{x})$-almost everywhere. Let $\ddot{\nabla}\mathbf{d}(\theta) \equiv \begin{bmatrix} \int \frac{dvech(\mathbf{A}(\mathbf{x};\theta))}{d\theta} f(\mathbf{x};\theta)\, d\nu(\mathbf{x}) \\ \int \frac{dvech(\mathbf{B}(\mathbf{x};\theta))}{d\theta} f(\mathbf{x};\theta)\, d\nu(\mathbf{x}) \end{bmatrix}$.

Substituting $\theta^*$ into (A11) and (A12) gives the result that $\nabla\mathbf{d}^* = \ddot{\nabla}\mathbf{d}(\theta^*)$ when the probability model is correctly specified. The result $\ddot{\nabla}\mathbf{d}_n \to \ddot{\nabla}\mathbf{d}$ as $n \to \infty$ with probability one then follows from using the Uniform Law of Large Numbers. The result $\ddot{\nabla}\mathbf{d}_n\left(\hat{\theta}_n\right) \to \ddot{\nabla}\mathbf{d}(\theta^*)$ follows from $\mathbf{d}_n \to \ddot{\nabla}\mathbf{d}$ and the result of Theorem 2, which is $\hat{\theta}_n \to \theta^*$ with probability one. Q.E.D.

## References


1. White, H. Maximum Likelihood Estimation of Misspecified Models. *Econometrica* **1982**, *50*, 1–25. [CrossRef]
2. White, H. *Estimation, Inference, and Specification Analysis*; Cambridge University Press: New York, NY, USA, 1994.
3. Kashner, T.M.; Henley, S.S.; Golden, R.M.; Rush, A.J.; Jarrett, R.B. Assessing the preventive effects of cognitive therapy following relief of depression: A methodological innovation. *J. Affect. Disord.* **2007**, *104*, 251–261. [CrossRef] [PubMed]
4. Kashner, T.M.; Rosenheck, R.; Campinell, A.B.; Suris, A.; Crandall, R.; Garfield, N.J.; Lapuc, P.; Pyrcz, K.; Soyka, T.; Wicker, A. Impact of work therapy on health status among homeless, substance-dependent veterans: A randomized controlled trial. *Arch. Gen. Psychiatry* **2002**, *59*, 938–944. [CrossRef] [PubMed]
5. Kashner, T.M.; Carmody, T.J.; Suppes, T.; Rush, A.J.; Crismon, M.L.; Miller, A.L.; Toprac, M.; Madhukar, T. Catching up on health outcomes: The Texas Medication Algorithm Project. *Health Serv. Res.* **2003**, *38*, 311–331. [CrossRef] [PubMed]
6. Kashner, T.M.; Henley, S.S.; Golden, R.M.; Byrne, J.M.; Keitz, S.A.; Cannon, G.W.; Chang, B.K.; Holland, G.J.; Aron, D.C.; Muchmore, E.A.; et al. Studying the Effects of ACGME Duty Hours Limits on Resident

Satisfaction: Results From VA Learners' Perceptions Survey. *Acad. Med.* **2010**, *85*, 1130–1139. [CrossRef] [PubMed]
7. Henley, S.S.; Kashner, T.M.; Golden, R.M.; Westover, A.N. Response to letter regarding "A systematic approach to subgroup analyses in a smoking cessation trial". *Am. J. Drug Alcohol Abuse* **2016**, *42*, 112–113. [CrossRef] [PubMed]
8. Westover, A.N.; Kashner, T.M.; Winhusen, T.M.; Golden, R.M.; Henley, S.S. A Systematic Approach to Subgroup Analyses in a Smoking Cessation Trial. *Am. J. Drug Alcohol Abuse* **2015**, *41*, 498–507. [CrossRef] [PubMed]
9. Brakenridge, S.C.; Henley, S.S.; Kashner, T.M.; Golden, R.M.; Paik, D.; Phelan, H.A.; Cohen, M.; Sperry, J.L.; Moore, E.E.; Minei, J.P.; et al. Comparing Clinical Predictors of Deep Venous Thrombosis vs. Pulmonary Embolus After Severe Blunt Injury: A New Paradigm for Post-Traumatic Venous Thromboembolism? *J. Trauma Acute Care Surg.* **2013**, *74*, 1231–1238. [CrossRef] [PubMed]
10. Brakenridge, S.C.; Phelan, H.A.; Henley, S.S.; Golden, R.M.; Kashner, T.M.; Eastman, A.E.; Sperry, J.L.; Harbrecht, B.G.; Moore, E.E.; Cuschieri, J.; et al. Early blood product and crystalloid volume resuscitation: Risk association with multiple organ dysfunction after severe blunt traumatic injury. *J. Trauma* **2011**, *71*, 299–305. [CrossRef] [PubMed]
11. Chesher, A. The information matrix test: Simplified calculation via a score test interpretation. *Econ. Lett.* **1983**, *13*, 45–48. [CrossRef]
12. Lancaster, T. The Covariance Matrix of the Information Matrix Test. *Econometrica* **1984**, *52*, 1051–1054. [CrossRef]
13. Aparicio, T.; Villanua, I. The asymptotically efficient version of the information matrix test in binary choice models. A study of size and power. *J. Appl. Stat.* **2001**, *28*, 167–182. [CrossRef]
14. Davidson, R.; MacKinnon, J.G. Graphical Methods for Investigating the Size and Power of Hypothesis Tests. *Manch. Sch.* **1998**, *66*, 1–26. [CrossRef]
15. Davidson, R.; MacKinnon, J.G. A New Form of the Information Matrix Test. *Econometrica* **1992**, *60*, 145–157. [CrossRef]
16. Dhaene, G.; Hoorelbeke, D. The information matrix test with bootstrap-based covariance matrix estimation. *Econ. Lett.* **2004**, *82*, 341–347. [CrossRef]
17. Stomberg, C.; White, H. *Bootstrapping the Information Matrix Test*; Discussion Paper; Department of Economics, University of California: San Diego, CA, USA, 2000.
18. Taylor, L.W. The Size Bias of White's Information Matrix Test. *Econ. Lett.* **1987**, *24*, 63–67. [CrossRef]
19. Presnell, B.; Boos, D.D. The IOS Test for Model Misspecification. *J. Am. Stat. Assoc.* **2004**, *99*, 216–227. [CrossRef]
20. Capanu, M.; Presnell, B. Misspecification tests for binomial and beta-binomial models. *Stat. Med.* **2008**, *27*, 2536–2554. [CrossRef] [PubMed]
21. Capanu, M. Tests of Misspecification for Parametric Models. University of Florida, 2005. Available online: http://etd.fcla.edu/UF/UFE0010943/capanu_m.pdf (accessed on 1 June 2016).
22. Zhang, S.; Song, P.X.K.; Shi, D.; Zhou, Q.M. Information ratio test for model misspecification on parametric structures in stochastic diffusion models. *Comput. Stat. Data Anal.* **2012**, *56*, 3975–3987. [CrossRef]
23. Golden, R.M.; Henley, S.S.; White, H.; Kashner, T.M. New Directions in Information Matrix Testing: Eigenspectrum Tests. In *Causality, Prediction, and Specification Analysis: Recent Advances and Future Directions: Essays in Honor of Halbert L. White, Jr. (Festschrift Hal White Conference)*; Chen, X., Swanson, N.R., Eds.; Springer: New York, NY, USA, 2013; pp. 145–178.
24. Cho, J.S.; White, H. Testing the Equality of Two Positive-Definite Matrices with Application to Information Matrix Testing. In *Essays in Honor of Peter C. B. Phillips*; Chang, Y., Fomby, T.B., Park, J.Y., Eds.; Emerald Group Publishing Limited: Bingley, UK, 2014; pp. 491–556.
25. Zhou, Q.M.; Song, P.X.K.; Thompson, M.E. Information Ratio Test for Model Misspecification in Quasi-Likelihood Inference. *J. Am. Stat. Assoc.* **2012**, *107*, 205–213. [CrossRef]
26. Huang, W.; Prokhorov, A. A Goodness-of-Fit Test for Copulas. *Econom. Rev.* **2014**, *33*, 751–771. [CrossRef]
27. Marlow, W.H. *Mathematics for Operations Research*; Dover Publications: Mineola, NY, USA, 2012.
28. Magnus, J.R. On the concept of matrix derivative. *J. Multivar. Anal.* **2010**, *101*, 2200–2206. [CrossRef]
29. Magnus, J.R.; Neudecker, H. *Matrix Differential Calculus with Applications in Statistics and Econometrics*; John Wiley & Sons: New York, NY, USA, 1999.

30. Jennrich, R.I. Asymptotic Properties of Non-linear Least Squares Estimators. *Ann. Math. Stat.* **1969**, *40*, 633–643. [CrossRef]
31. White, H. Consequences and detection of misspecified nonlinear regression models. *J. Am. Stat. Assoc.* **1981**, *76*, 419–433. [CrossRef]
32. Huber, P. The Behavior of Maximum Likelihood Estimates under Non-Standard Conditions. In *Proceedings Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1*; University of California Press: Berkeley, CA, USA, 1967; pp. 221–233.
33. Prokhorov, A.; Schepsmeier, U.; Zhu, Y. *Generalized Information Matrix Tests for Copulas, Working Paper*; University of Sydney Business School, Discipline of Business Analytics: Sydney, Australia, 2015.
34. Bozdogan, H. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* **1987**, *52*, 345–370. [CrossRef]
35. Linhart, H.; Zucchini, W. *Model Selection*; Wiley: New York, NY, USA, 1986.
36. Takeuchi, K. Distribution of information statistics and a criterion of model fitting for adequacy of models. *Math. Sci.* **1976**, *153*, 12–18.
37. Cho, J.; Phillips, P. Testing Equality of Covariance Matrices via Pythagorean Means. 2014. Available online: http://ssrn.com/abstract=2533002 (accessed on 1 June 2016).
38. Golden, R.M. Statistical tests for comparing possibly misspecified and nonnested models. *J. Math. Psychol.* **2000**, *44*, 153–170. [CrossRef] [PubMed]
39. Golden, R.M. Discrepancy risk model selection test theory for comparing possibly misspecified or nonnested models. *Psychometrika* **2003**, *68*, 229–249. [CrossRef]
40. Henley, S.S.; Golden, R.M.; Kashner, T.M.; White, H. *Exploiting Hidden Structures in Epidemiological Data: Phase II Project*; NIH/NIAAA: Plano, TX, USA, 2000.
41. Henley, S.S.; Golden, R.M.; Kashner, T.M.; White, H.; Paik, D. *Robust Classification Methods for Categorical Regression: Phase II Project*; National Cancer Institute: Plano, TX, USA, 2008.
42. Henley, S.S.; Golden, R.M.; Kashner, T.M.; White, H.; Katz, R.D. *Model Selection Methods for Categorical Regression: Phase I Project*; NIH/NIAAA: Plano, TX, USA, 2003.
43. Vuong, Q.H. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **1989**, *57*, 307–333. [CrossRef]
44. Bozdogan, H. Akaike's Information Criterion and Recent Developments in Information Complexity. *J. Math. Psychol.* **2000**, *44*, 62–91. [CrossRef] [PubMed]
45. Fawcett, T. An introduction to ROC analysis. *Pattern Recogn. Lett.* **2006**, *27*, 861–874. [CrossRef]
46. Pepe, M.S. *The Statistical Evaluation of Medical Tests for Classification and Prediction*; Oxford University Press: Oxford, UK, 2004.
47. Wickens, T.D. *Elementary Signal Detection Theory*; Oxford University Press: New York, NY, USA, 2002.
48. Hastie, T.; Tibshirani, R. *Generalized Additive Models*; Chapman and Hall: New York, NY, USA, 1990.
49. McCullagh, P.; Nelder, J.A. *Generalized Linear Models*; Chapman and Hall: London, UK; New York, NY, USA, 1989.
50. Wei, B. *Exponential Family Nonlinear Models*; Springer: New York, NY, USA, 1998.
51. Hosmer, D.W.; Lemeshow, S. *Applied Logistic Regression*; Wiley: New York, NY, USA, 1989.
52. Harrell, F.E. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*; Springer: New York, NY, USA, 2001.
53. Arminger, G.; Sobel, M.E. Pseudo-maximum likelihood estimation of mean and covariance structures with missing data. *J. Am. Stat. Assoc.* **1990**, *85*, 195–203. [CrossRef]
54. Gallini, J. Misspecifications that can result in path analysis structures. *Appl. Psychol. Meas.* **1983**, *7*, 125–137. [CrossRef]
55. Raudenbush, S.W.; Bryk, A.S. *Hierarchical Linear Models: Applications and Data Analysis Methods*; Sage Publications, Inc.: Thousand Oaks, CA, USA, 2002.
56. Hosmer, D.W.; Lemeshow, S. A goodness-of-fit test for the multiple logistic regression model. *Commun. Stat.* **1980**, *A10*, 1043–1069. [CrossRef]
57. Hosmer, D.W.; Lemeshow, S.; Klar, J. Goodness-of-Fit Testing for Multiple Logistic Regression Analysis when the Estimated Probabilities are Small. *Biom. J.* **1988**, *30*, 1–14. [CrossRef]
58. Hosmer, D.W.; Taber, S.; Lemeshow, S. The importance of assessing the fit of logistic regression models: A case study. *Am. J. Public Health* **1991**, *81*, 1630–1635. [CrossRef] [PubMed]

59. Serfling, R.J. *Approximation Theorems of Mathematical Statistics*; Wiley-Interscience: New York, NY, USA, 1980.
60. White, H. *Asymptotic Theory for Econometricians*, Revised Edition; Academic Press: New York, NY, USA, 2001.
61. White, H. Using least squares to approximate unknown regression functions. *Int. Econ. Rev.* **1980**, *21*, 149–170. [CrossRef]