New Directions in Information Matrix Testing: Eigenspectrum Tests

Richard M. Golden, Steven S. Henley, Halbert White and T. Michael Kashner

Abstract Model specification tests are essential tools for evaluating the appropriateness of probability models for estimation and inference. White (*Econometrica*, 50: 1–25, 1982) proposed that model misspecification could be detected by testing the null hypothesis that the Fisher information matrix (IM) Equality holds by comparing

R. M. Golden (✉)
Cognitive Science and Engineering,
School of Behavioral and Brain Sciences,
University of Texas at Dallas,
GR4.1 800 W. Campbell Rd.,
Richardson, TX 75080-3021, USA
e-mail: golden@utdallas.edu

S. S. Henley
Martingale Research Corporation,
101 E. Park Blvd., Suite 600,
Plano, TX 75074, USA
e-mail: stevenh@martingale-research.com

S. S. Henley and T. Michael Kashner
Department of Medicine, Loma Linda University School of Medicine,
Loma Linda, CA 92357, USA

H. White
Department of Economics,
University of California San Diego,
La Jolla, CA 92093-0508, USA
e-mail: hwhite@ucsd.edu

T. Michael Kashner
Department of Psychiatry, University of Texas Southwestern Medical Center,
Dallas, TX 75390, USA
e-mail: Michael.kashner@va.gov

T. Michael Kashner
Office of Academic Affiliations,
Department of Veterans Affairs,
Washington, D.C. 20420, USA

X. Chen and N. R. Swanson (eds.), *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, DOI: 10.1007/978-1-4614-1653-1_6,

linear functions of the Hessian to outer product gradient (OPG) inverse covariance matrix estimators. Unfortunately, a number of researchers have reported difficulties in obtaining reliable inferences using White's (*Econometrica*, 50: 1–25, 1982) original information matrix test (IMT). In this chapter, we extend White (*Econometrica*, 50: 1–25, 1982) to present a new generalized information matrix test (GIMT) theory and develop a new Adjusted Classical GIMT and five new Eigenspectrum GIMTs that compare nonlinear functions of the Hessian and OPG covariance matrix estimators. We then evaluate the level and power of these new GIMTs using simulation studies on realistic epidemiological data and find that they exhibit appealing performance on sample sizes typically encountered in practice. Our results suggest that these new GIMTs are important tools for detecting and assessing model misspecification, and thus will have broad applications for model-based decision making in the social, behavioral, engineering, financial, medical, and public health sciences.

Keywords Eigenspectrum · Goodness-of-fit · Information matrix test · Logistic regression · Specification analysis

1 Introduction

A correctly specified probability model has the property that it contains the probability distribution that generates the observed data. Model specification tests examine the null hypothesis that a researcher's probability model is correctly specified. If the researcher's model of the observed data is not correct (i.e., misspecified), then the interpretation of parameter estimates and the validity of inferences obtained from the resulting probability model may be suspect. Thus, to avoid misleading inferences, the effects of model specification must be considered. For example, in the social and medical sciences (e.g., Kashner et al. 2010), the incompleteness of behavioral and medical theories mandates the need for principled specification analysis methods that use empirical observations to assess quality of a particular theory. This situation, all too common in statistical modeling, provides considerable impetus for the development of improved model specification tests.

1.1 Model Misspecification

When viewed from a practical perspective, the problem of model misspecification is essentially unavoidable. Although ideally a correctly specified model is always preferable, in many fields of science such as econometrics, medicine, and psychology some degree of model misspecification is inevitable. Indeed, all probability models are abstractions of reality, so the issue of model misspecification is fundamentally an empirical issue that is dependent upon how the model will be developed and applied

in practice (e.g., Fisher 1922; White 1980, 1981, 1982, 1994; Begg and Lagakos 1990; Cox 1990; Lehmann 1990).

A variety of methods have been developed for the purpose of the assessment of model misspecification. For example, graphical residual diagnostics are useful for identifying the presence of model misspecification for the class of generalized linear models (e.g., Davison and Tsai 1992) and the larger class of exponential family nonlinear models (e.g., Wei 1998, Chap. 6). However, these methods require more subjective interpretations because results are expressed as measures of fit rather than as hypothesis tests. Moreover, specification tests such as chi-square goodness-of-fit tests (e.g., Hosmer et al. 1991, 1997) are not applicable in a straightforward manner when the observations contain continuous random variables. Link specification tests (Collett 2003; Hilbe 2009) are applicable for testing the assumption of linearity in the link function (e.g., logit), but are not designed to detect other types of model misspecification. Further, the applicability of these methods to more complex probability models such as hierarchical (e.g., Agresti 2002; Raudenbush and Bryk 2002), mixed (e.g., Verbeke and Lesaffre 1997), and latent variable (e.g., Gallini 1983; Arminger and Sobel 1990) models may not always be obvious.

1.2 Specification Analysis for Logistic Regression

Logistic regression modeling (Christensen 1997; Hosmer and Lemeshow 2000; Harrell 2001; Agresti 2002; Collett 2003; Hilbe 2009) is an important and widely used analysis tool in various fields; however, the number of available options for the assessment of model misspecification is relatively limited (see Sarkar and Midi 2010 for a review). Typically, the detection of model misspecification in logistic regression models is based upon direct comparison of the observed conditional frequencies of the response variable with predicted conditional probabilities (Hosmer et al. 1997). Unfortunately, the observed conditional frequencies of the response variable can only be compared with predicted conditional probabilities for a particular pattern of predictor variable values in a given data record. In practice, patterns of predictor variable values may rarely be repeated for more complex models involving either multiple categorical predictor variables or continuous-valued predictor variables. Because the number of distinct predictor patterns often increases as the number of records (i.e., sample size) increases, such applications of classical "fixed-cell asymptotic" results are problematic (e.g., Osius and Rojek 1992). To address this problem, "grouping" methods have been proposed that require artificially grouping similar, yet distinct predictor patterns (Bertolini et al. 2000; Archer and Lemeshow 2006).

A variety of test statistics that explicitly compare predicted probabilities with observed frequencies using grouping methods have been proposed, and include chi-square test methods (e.g., Hosmer and Lemeshow 1980; Tsiatis 1980; Hosmer et al. 1988, 1997; Copas 1989; Qin and Zhang 1997; Zhang 1999; Archer and Lemeshow 2006; Deng et al. 2009), sum-squared comparison methods (Copas 1989; Kuss 2002), and the closely related likelihood ratio test deviance-based comparison methods

(e.g., Hosmer and Lemeshow 2000, pp. 145–146; Kuss 2002). Without employing such grouping methods, the resulting test statistics associated with direct comparison of observed conditional frequencies and predicted conditional probabilities will have excessive degrees of freedom and thus poor power. However, when such grouping methods are applied, they may actually have the unintended consequence of redefining the probability model whose integrity is being evaluated (Hosmer et al. 1997).

One solution to dealing with the "grouping" problem is to introduce appropriate regularity conditions intended to characterize the asymptotic behavior of the test statistics while allowing the number of distinct predictor patterns to increase with the sample size (e.g., Osius and Rojek 1992). Another important solution to the "grouping" problem is to embed the probability model whose specification is being scrutinized within a larger probability model and then compare the predicted probabilities of both models (e.g., Stukel 1988). Other approaches have explored improved approximations to Pearson's goodness-of-fit statistic (McCullagh 1985; Farrington 1996). Yet, despite these approaches, the variety of methods available for assessing the presence of model misspecification is surprisingly limited, and these limitations are particularly striking in the context of logistic regression modeling (e.g., Sarkar and Midi 2010).

1.3 Information Matrix Test

White (1982; also see 1987, 1994) proposed a particular model specification test called the *information matrix test* (IMT). Unlike chi-square goodness-of-fit tests and graphical diagnostics, IMTs are based upon the theoretical expectation that the Hessian inverse covariance matrix estimator (derived from the Hessian of the log-likelihood function) and the outer product gradient (OPG) inverse covariance matrix estimator (derived from the first derivatives of the log-likelihood function) are asymptotically equivalent whenever the researcher's probability model is correctly specified. We define a *full IMT* as a statistical test that tests the null hypothesis of asymptotic equivalence of the Hessian and OPG asymptotic covariance matrix estimators.

An important virtue of the IMT method is that it is applicable in a straightforward manner to a broad class of probability models. This includes not only linear and nonlinear regression models, but also even more complex models such as: limited dependent variables models (e.g., Maddala 1999; Greene 2003), exponential family nonlinear models (e.g., Wei 1998), generalized linear models (e.g., McCullagh and Nelder 1989), generalized additive models (e.g., Hastie and Tibshirani 1986, 1990), hierarchical models (e.g., Agresti 2002; Raudenbush and Bryk 2002), mixed models (e.g., Verbeke and Lesaffre 1997), latent variable models (e.g., Gallini 1983; Arminger and Sobel 1990), conditional random fields (e.g., Winkler 1991), and time series models (e.g., Hamilton 1994; White 1994; Box et al. 2008; Tsay 2010). However, despite the broad applicability of the IMT, the majority of the research in the

development and evaluation of IMTs has focused on linear regression (Hall 1987; Taylor 1987; Davidson and MacKinnon 1992, 1998), logistic regression (Aparicio and Villanua 2001; Zhang 2001), probit (Davidson and MacKinnon 1992, 1998; Stomberg and White 2000; Dhaene and Hoorelbeke 2004), and Tobit (Horowitz 1994, 2003) modeling.

1.4 Empirical Performance of the Information Matrix Test

Although theoretically attractive, the IMT has not been widely used to detect model misspecification. In particular, some researchers have found the full IMT (White 1982) both analytically and computationally burdensome because its derivation and computation require third derivatives of the log-likelihood. To address this problem, Chesher (1983) and Lancaster (1984) demonstrated how the calculation of the third derivatives of the log-likelihood function could be avoided for the full IMT by showing that when the OPG and Hessian inverse covariance matrix estimators are asymptotically equivalent, the third derivatives of the log-likelihood may be expressed in terms of the first and second derivatives of the log-likelihood. This particular version of the White (1982) full IMT is commonly referred to as the *OPG IMT*. Unfortunately, OPG full IMTs were subsequently found to exhibit poor performance in various simulation studies for logistic regression (Aparicio and Villanua 2001) and linear regression (Taylor 1987; Davidson and MacKinnon 1992; Dhaene and Hoorelbeke 2004). This prompted some researchers (Davidson and MacKinnon 1992, 1998; Stomberg and White 2000; Dhaene and Hoorelbeke 2004) to re-evaluate the original formulation by White (1982), which involves explicit analytical computation of the third derivatives of the log-likelihood function.

In a series of simulation studies, researchers (e.g., Orme 1990; Stomberg and White 2000) have demonstrated that both the original White (1982) formulation and the OPG-IMT method exhibit relatively erratic performance and require excessively large sample sizes to ensure that the test statistic behaves properly. This led a number of researchers (e.g., Davidson and MacKinnon 1992; Stomberg and White 2000; Aparicio and Villanua 2001; Dhaene and Hoorelbeke 2004) to suggest that the erratic behavior of the full IMT for linear regression is due to excessive test statistic variance, since the degrees of freedom of the full IMT increase as a quadratic function of the number of free parameters of the probability model.

Further, researchers (Taylor 1987; Orme 1990; Horowitz 1994, 2003) have provided empirical evidence that the poor level performance of the OPG IMT is due to failure to incorporate the third derivatives of the log-likelihood functions as originally recommended by White (1982). Stomberg and White (2000) have shown demonstrable improvements using a bootstrapped version of the full IMT, but this method requires substantial computational resources.

1.5 Nondirectional and Directional Tests

A "nondirectional IMT" examines the null hypothesis that the Hessian and OPG covariance matrix estimators are asymptotically equivalent. White's (1982) Classical Full IMT is an example of a nondirectional information matrix test. If the null hypothesis of a nondirectional information test is false, it directly follows from Fisher's Information Matrix Equality that the probability model is misspecified.

A "directional IMT" compares functions of the OPG and Hessian covariance matrix estimators for the purpose of identifying specific types of model misspecification, rather than implementing a full covariance matrix estimator comparison. Two potential advantages of directional tests are: (1) gaining important insights regarding how to improve the quality of a misspecified model by identifying specific aspects of a model that appear to be correctly or incorrectly specified, and (2) better level performance and greater statistical power in the detection of model misspecification. White (1982) explicitly emphasized that improved specification testing performance and specific specification tests could be obtained through the use of directional information matrix tests. Nonetheless, as previously described, the majority of research has focused upon the full IMT rather than on particular directional versions of the full IMT as recommended by White (1982).

Directional tests also may, in some cases, provide improved statistical power if such tests are appropriately designed. However, despite the advantages of directional specification testing, little theoretical or empirical research has been conducted to more thoroughly explore directional IMTs as viable alternatives to White's (1982) nondirectional Classical Full IMT. Such insights may also be helpful for suggesting specific modifications to a researcher's model to improve its quality. Although, nondirectional tests are useful for overall assessments of model misspecification, but directional tests provide insights into which properties of a model are sensitive to the effects of model misspecification.

Prior research on directional versions of the full IMT has focused upon the detection of skewness, kurtosis, and heteroskedasticity in linear regression models, with a few notable exceptions (i.e., Henley et al. 2001, 2004; Alonso et al. 2008). For example, Bera and Lee (1993; also see Hall 1987; Chesher and Spady 1991) have shown how to derive directional information matrix tests for linear regression models using White's (1982) theoretical framework. These directional information matrix tests were shown to be mathematically equivalent (see White 1982; Hall 1987; Chesher and Spady 1991; Bera and Lee 1993 for relevant reviews) to commonly used statistical tests for checking for the presence of autoregressive conditional heteroskedasticity as well as checking for normality in the residual errors.

1.6 Logistic Regression Modeling IMTs

The IMT method is particularly attractive in the context of logistic regression modeling because it does not require the use of grouping mechanisms, and the degrees of freedom are solely dependent upon the number of free parameters in the model

rather than the degree to which the predictor patterns in the data set are replicated. However, the application of IMTs to the problem of the detection of misspecification in categorical regression (Agresti 2002) and, in particular, logistic regression modeling (Hosmer and Lemeshow 2000; Hilbe 2009) is less common (but see Orme 1988; Aparicio and Villanua 2001; Zhang 2001; Kuss 2002), despite the major role that logistic regression plays in applied statistical analysis (Christensen 1997; Hosmer and Lemeshow 2000; Harrell 2001; Agresti 2002; Collett 2003; Hilbe 2009).

1.7 Generalized Information Matrix Test Theory

In this chapter, we introduce the essential ideas of our Generalized Information Matrix Test (GIMT) theory (Henley et al. 2001, 2004, 2008). GIMT theory includes the IMTs previously discussed in the literature, as well as a larger class of directional and nondirectional IMTs. We apply GIMT theory to develop six specific new GIMTs. We begin with a new version of the original $k(k+1)/2$ degrees of freedom White (1982) Classical Full IMT, called the "Adjusted Classical GIMT", which is applicable to a k parameter model. In addition, we explore information matrix testing by introducing and empirically evaluating five new Information Matrix Tests based upon comparing specific nonlinear functions of the eigenspectra of the Hessian and OPG covariance matrices (rather than their inverses) developed by Henley et al. (2001, 2004, 2008). The first of these directional tests is the k-degree of freedom "Log Eigenspectrum GIMT" based on the null hypothesis that the k eigenvalues of the Hessian and OPG covariance matrices are the same. The one-degree of freedom "Log Determinant GIMT" tests the null hypothesis that the products of the eigenvalues of the Hessian and OPG covariance matrices are identical. Log Determinant GIMTs are exceptionally sensitive to small differences in the eigenstructures. The "Log Trace GIMT" is a one-degree of freedom GIMT that tests the null hypothesis that the sums of the eigenvalues of the Hessian and OPG covariance matrices are identical. Log Trace GIMTs focus on differences in the major principal components of the Hessian and OPG covariance matrices. The fourth eigenspectrum test is the two-degree of freedom "Generalized Variance GIMT" that tests the composite null hypothesis that the Log Determinant and Log Trace GIMTs' null hypotheses hold. In particular, the Generalized Variance GIMT exploits the complementary features of the Log Trace and Log Determinant GIMTs, since the Log Determinant GIMT is sensitive to small differences in the entire eigenspectrum of the Hessian and OPG covariance matrices, while the Log Trace GIMT tends to focus on the larger eigenvalues. Finally, if the Hessian and OPG covariance matrices are identical, then the Hessian covariance matrix multiplied by the inverse of the OPG covariance matrix will be the identity matrix. This observation suggests a fifth type of GIMT called the "Log Generalized Akaike Information Criterion (GAIC) GIMT" for examining the average relative deviation between the eigenspectra of the Hessian and OPG covariance matrices. The Log GAIC GIMT, like the Log Determinant and Log Trace

GIMTs, is also a one-degree of freedom test sensitive to small differences in the eigenstructures of the Hessian and OPG covariance matrices.

We then provide a series of simulation studies to investigate the level and power properties of the new Eigenspectrum GIMTs and the Adjusted Classical GIMT. Our simulation studies are intended to achieve three specific objectives. First, we evaluate the reliability of the large sample approximations for estimating Type I error probabilities (level) for the Adjusted Classical GIMT and our five new Eigenspectrum GIMTs. Second, we evaluate the level-power performance of the new Eigenspectrum GIMTs relative to the Adjusted Classical GIMT. Finally, we evaluate the applicability of the new GIMTs to detect model misspecification in representative, realistic epidemiological data.

2 Theory

2.1 Information Matrix Equality

In what follows, we do not give formal results. For the most part, the necessary theory can already be found in White (1982, 1994). We use the following notation. Let the d-dimensional real column vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be realizations of the *i.i.d.* random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$ having support $\mathcal{R}^d$. Let the parameter space $\boldsymbol{\Theta} \subseteq \mathcal{R}^k$ be a compact set with non-empty interior. Let $f : \mathcal{X} \times \boldsymbol{\Theta} \rightarrow [0, \infty)$ be defined such that $f(\cdot\,; \boldsymbol{\theta})$ is a Radon-Nikodým density for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Let $f(\mathbf{x}_i; \boldsymbol{\theta})$ denote the likelihood of an observation $\boldsymbol{x}_i$ for parameter vector $\boldsymbol{\theta}$. Let $\bar{\mathbf{B}}_n = n^{-1} \sum_{i=1}^n \mathbf{B}_i$ where $\mathbf{B}_i = \mathbf{g}_i \mathbf{g}_i^T$ and $\mathbf{g}_i \equiv -\nabla_{\boldsymbol{\theta}} \log f(\mathbf{X}_i; \cdot)$. Let $\bar{\mathbf{A}}_n = n^{-1} \sum_{i=1}^n \mathbf{A}_i$ where $\mathbf{A}_i \equiv -\nabla^2_{\boldsymbol{\theta}} \log f(\mathbf{X}_i; \cdot)$. Let $\mathbf{A}$ and $\mathbf{B}$ denote the respective expected values of $\bar{\mathbf{A}}_n$ and $\bar{\mathbf{B}}_n$ (when they exist). Suppose the *maximum likelihood estimator* $\hat{\boldsymbol{\theta}}_n$, which maximizes the *likelihood function* $\prod_{i=1}^n f(\mathbf{X}_i; \boldsymbol{\theta})$, converges almost surely to $\boldsymbol{\theta}^* \in \text{int}\, \Theta$. Let $\mathbf{A}^* \equiv \mathbf{A}(\boldsymbol{\theta}^*)$ and $\mathbf{B}^* \equiv \mathbf{B}(\boldsymbol{\theta}^*)$. Let $\hat{\mathbf{A}}_n \equiv \bar{\mathbf{A}}_n(\hat{\boldsymbol{\theta}}_n)$ and $\hat{\mathbf{B}}_n \equiv \bar{\mathbf{B}}_n(\hat{\boldsymbol{\theta}}_n)$. We say the model is *correctly specified* if there exists $\boldsymbol{\theta}_0$ such that $f(\cdot; \boldsymbol{\theta}_0)$ is the true density of $\mathbf{X}_i$. In this case, it holds under general conditions that $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$. The GIMT is based upon the critical observation that under correct specification, the Fisher Information Matrix equality holds, that is, $\mathbf{A}^* = \mathbf{B}^*$ (e.g., White 1982, 1994). This hypothesis may be tested by comparing $\hat{\mathbf{A}}_n$ and $\hat{\mathbf{B}}_n$. Rejecting the null hypothesis that $\mathbf{A}^* = \mathbf{B}^*$, thus indicates the presence of model misspecification. In this situation, the *classic Hessian covariance matrix estimator* $\hat{\mathbf{A}}_n^{-1}$ and *classic OPG covariance matrix* estimator $\hat{\mathbf{B}}_n^{-1}$ for $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\right)$ are inconsistent and the *robust estimator* $\hat{\mathbf{C}}_n \equiv \hat{\mathbf{A}}_n^{-1} \hat{\mathbf{B}}_n \hat{\mathbf{A}}_n^{-1}$ (e.g., Huber 1967; White 1982, 1994; Golden 1996) is consistent and should be used instead.

2.2 The Null Hypothesis for a Generalized IMT

Let $\Upsilon^{k\times k} \subseteq \mathcal{R}^{k\times k}$ be a compact set that contains $\mathbf{A}^*$ and $\mathbf{B}^*$ in its interior. Let $\mathbf{s} : \Upsilon^{k\times k} \times \Upsilon^{k\times k} \to \mathcal{R}^r$ be continuously differentiable in both of its matrix arguments where r is a positive integer less than or equal to $k(k+1)/2$. The function $\mathbf{s}$ is called a *Generalized Information Matrix Test (GIMT) Hypothesis Function* when it satisfies the condition that: For every $\mathbf{A}, \mathbf{B} \in \Upsilon^{k\times k}$, if $\mathbf{A} = \mathbf{B}$, then $\mathbf{s}(\mathbf{A}, \mathbf{B}) = \mathbf{0}_r$. Throughout this chapter, we assume that the GIMT Hypothesis function $\mathbf{s} : \Upsilon^{k\times k} \times \Upsilon^{k\times k} \to \mathcal{R}^r$ is a continuously differentiable function of both its arguments and that $\frac{d\mathbf{s}(\mathbf{A}(\boldsymbol{\theta}),\mathbf{B}(\boldsymbol{\theta}))}{d\boldsymbol{\theta}}$ evaluated at $\boldsymbol{\theta}^*$ has full row rank r. It will also be convenient to let $\mathbf{s}^* \equiv \mathbf{s}(\mathbf{A}^*, \mathbf{B}^*)$.

A GIMT is defined as a test statistic $\hat{\mathbf{s}}_n \equiv \mathbf{s}\left(\hat{\mathbf{A}}_n, \hat{\mathbf{B}}_n\right)$ that tests the null hypothesis:

$$H_0 : \mathbf{s}^* = \mathbf{0}_r.$$

We distinguish between "nondirectional" and "directional" GIMT hypothesis functions. A GIMT hypothesis function $\mathbf{s}$ is called *nondirectional* when $\mathbf{s}$ has the property that: For every $\mathbf{A}, \mathbf{B} \in \Upsilon^{k\times k}$, $\mathbf{A} = \mathbf{B}$, *if and only if* $\mathbf{s}(\mathbf{A}, \mathbf{B}) = \mathbf{0}_r$. Otherwise, the GIMT hypothesis function $\mathbf{s}$ is called *directional*.

2.3 Asymptotic Behavior of the Generalized IMT Statistic

We now define the *Generalized Information Matrix Test (GIMT)* statistic:

$$\hat{\mathcal{W}}_n \equiv n\left(\hat{\mathbf{s}}_n\right)^T \hat{\sum}_{n,\mathbf{s}}^{-1}\left(\hat{\mathbf{s}}_n\right). \tag{1}$$

where the estimator $\hat{\sum}_{n,\mathbf{s}}^{-1}$ is an estimator of the asymptotic covariance matrix of $n^{1/2}\hat{\mathbf{s}}_n$, $\sum_{\mathbf{s}}^{-1}(\boldsymbol{\theta}^*)$.

Under standard regularity conditions, $\hat{\mathcal{W}}_n$ has a chi-squared distribution with r degrees of freedom when the null hypothesis $H_0 : \mathbf{s}^* = \mathbf{0}_r$ holds. Let $\hat{\mathbf{g}}_i \equiv \mathbf{g}_i\left(\hat{\boldsymbol{\theta}}_n\right)$, $\mathbf{d}_i \equiv \begin{bmatrix} \mathbf{vec}(\mathbf{A}_i(\boldsymbol{\theta})) \\ \mathbf{vec}(\mathbf{B}_i(\boldsymbol{\theta})) \end{bmatrix}$, and $\nabla\bar{\mathbf{d}}_n(\boldsymbol{\theta}) \equiv n^{-1}\sum_{i=1}^n \nabla\mathbf{d}_i(\boldsymbol{\theta})$. The covariance matrix estimator $\hat{\sum}_{n,\mathbf{s}}$ is given by:

$$\hat{\sum}_{n,\mathbf{s}} \equiv \left[\frac{\partial\mathbf{s}}{\partial\mathbf{A}}\left(\hat{\mathbf{A}}_n\right)\ \frac{\partial\mathbf{s}}{\partial\mathbf{B}}\left(\hat{\mathbf{B}}_n\right)\right]^T \hat{\mathbf{Q}}_n \left[\frac{\partial\mathbf{s}}{\partial\mathbf{A}}\left(\hat{\mathbf{A}}_n\right)\ \frac{\partial\mathbf{s}}{\partial\mathbf{B}}\left(\hat{\mathbf{B}}_n\right)\right]$$

where $\hat{\mathbf{Q}}_n$ is computed from $\mathbf{d}_i$, $\hat{\mathbf{A}}_n$, $\nabla\bar{\mathbf{d}}_n$, $\mathbf{g}_i$ and $\hat{\boldsymbol{\theta}}_n$ following the approach of White (1982).

When the r-dimensional matrix $\sum_{\mathbf{s}}(\boldsymbol{\theta}^*)$ is singular and has rank g where $0 < g < r$, it is often possible to replace the original GIMT hypothesis function $\mathbf{s} : \Upsilon^{k\times k} \times \Upsilon^{k\times k} \rightarrow \mathcal{R}^r$ with an *alternative "adjusted" GIMT hypothesis function* $\tilde{\mathbf{s}} : \Upsilon^{k\times k} \times \Upsilon^{k\times k} \rightarrow \mathcal{R}^g$ that tests a similar null hypothesis yet has the property that the resulting asymptotic covariance matrix of $n^{1/2}\tilde{\mathbf{s}}_n$ is nonsingular. Let the *adjusted hypothesis projection matrix* $\mathbf{T}$ be a rectangular matrix with g rows and r columns with full row rank. Then, a decision indicating the "adjusted" null hypothesis $\tilde{H}_0 : \mathbf{T}\mathbf{s}^* = \mathbf{0}_g$ should be rejected also implies that the original null hypothesis $\mathbf{H}_0 : \mathbf{s}^* = \mathbf{0}_r$ should be rejected as well. Note that the adjusted null hypothesis projects the original GIMT hypothesis function from the original r-dimensional space into a g-dimensional subspace. Let $\tilde{\mathbf{s}}_n \equiv \mathbf{T}\hat{\mathbf{s}}_n$. Let $\tilde{\sum}_{n,\mathbf{s}} \equiv \mathbf{T}\hat{\sum}_{n,\mathbf{s}}\mathbf{T}^T$. Then $\tilde{\mathcal{W}}_n \equiv n\,(\tilde{\mathbf{s}}_n)^T \tilde{\sum}_{n,\mathbf{s}}^{-1}(\tilde{\mathbf{s}}_n)$ is called an "adjusted" GIMT, having g degrees of freedom (rather than r degrees of freedom) and testing the null hypothesis: $H_0 : \mathbf{T}\mathbf{s}^* = \mathbf{0}_g$.

Finally, although calculation of $\nabla\mathbf{d}_i(\boldsymbol{\theta})$ requires using the derivative of $\mathbf{A}_i$, which requires third derivatives of the log-likelihood, one can use the Lancaster-Chesher formula for $\nabla\mathbf{d}_i(\boldsymbol{\theta})$, denoted $\ddot{\nabla}\mathbf{d}_i(\boldsymbol{\theta})$. This avoids third derivatives by expressing $\nabla\mathbf{d}_i(\boldsymbol{\theta})$ in terms of the first and second derivatives of the log-likelihood function when the null hypothesis that the model is correctly specified holds (Lancaster 1984; also see Chesher 1983).

Thus, this yields six distinct GIMT statistics that can be used to test a single null hypothesis specified by a given GIMT Hypothesis function. When the GIMT null hypothesis holds either $\hat{\mathbf{B}}_n^{-1}$ or $\hat{\mathbf{C}}_n$ may be used instead of $\hat{\mathbf{A}}_n^{-1}$ to calculate $\hat{\mathbf{Q}}_n$. Furthermore, the assumption that the GIMT null hypothesis holds permits the use of the Lancaster-Chesher formula $\ddot{\nabla}\mathbf{d}_i(\boldsymbol{\theta})$ to avoid explicitly computing the third derivatives of the log-likelihood function (i.e., $\nabla\mathbf{d}_i(\boldsymbol{\theta})$). A *Hessian-GIMT statistic* corresponds to the case denoted by $\left\{\left(\hat{\mathbf{A}}_n\right)^{-1}, \nabla\mathbf{d}_i(\boldsymbol{\theta})\right\}$ where $\left(\hat{\mathbf{A}}_n\right)^{-1}$ is estimated by the Hessian covariance matrix estimator. An *OPG-GIMT statistic* corresponds to the case denoted by $\left\{\left(\hat{\mathbf{B}}_n\right)^{-1}, \ddot{\nabla}\mathbf{d}_i(\boldsymbol{\theta})\right\}$ where $\left(\hat{\mathbf{B}}_n\right)^{-1}$ is estimated by the OPG covariance matrix estimator (Lancaster 1984; also see Chesher 1983) and $\nabla\mathbf{d}_i(\boldsymbol{\theta})$ is calculated using the Lancaster-Chesher formula $\ddot{\nabla}\mathbf{d}_i(\boldsymbol{\theta})$. To the best of our knowledge, the use of the remaining four GIMT statistics $\left(\text{i.e., } \left\{\left(\hat{\mathbf{A}}_n\right)^{-1}, \ddot{\nabla}\mathbf{d}_i(\boldsymbol{\theta})\right\}, \left\{\hat{\mathbf{C}}_n, \ddot{\nabla}\mathbf{d}_i(\boldsymbol{\theta})\right\}, \left\{\left(\hat{\mathbf{B}}_n\right)^{-1}, \nabla\mathbf{d}_i(\boldsymbol{\theta})\right\}, \left\{\hat{\mathbf{C}}_n, \nabla\mathbf{d}_i(\boldsymbol{\theta})\right\}\right)$ associated with a single specific GIMT Hypothesis function for estimating the GIMT covariance matrix have not been discussed in the literature. However, in preliminary studies not reported here (Henley et al. 2001, 2004) we have found that these new statistics exhibit promising size and power properties.

It can be shown that for all six distinct GIMT statistics, the asymptotic distribution of $\hat{\mathcal{W}}_n$ is chi-square with r degrees of freedom when $H_0 : \mathbf{s}(\mathbf{A}^*, \mathbf{B}^*) = \mathbf{0}_r$ holds, under appropriate further regularity conditions and with a few minor modifications to the analysis presented by White (1982; see Proof of Theorem 4.1). Further, it can be shown that $\hat{\mathcal{W}}_n \rightarrow \infty$ almost surely when $H_0 : \mathbf{s}(\mathbf{A}^*, \mathbf{B}^*) = \mathbf{0}_r$ is false. Thus,

$\hat{\mathcal{W}}_n$ (or similarly the adjusted version $\hat{\mathcal{W}}_n$) can be used as a test statistic for the purpose of detecting the presence of model misspecification.

2.4 Classical IMT Family

White (1982) describes a family of IMTs that can be represented by a GIMT Hypothesis Function $\mathbf{s}$ of the form $\mathbf{s}(\mathbf{A}, \mathbf{B}) = \mathbf{S}\ \mathbf{vech}(\mathbf{A} - \mathbf{B})$, where the *selection matrix* $\mathbf{S} \in \mathcal{R}^{r \times k(k+1)/2}$ is some user-specified constant rectangular matrix of row rank r. The *Classical Full IMT* that has been widely discussed in the literature corresponds to the case where the selection matrix is a $k(k+1)/2$-dimensional identity matrix. White (1982) proposed the Classical Full IMT null hypothesis $H_0 : \mathbf{A}^* = \mathbf{B}^*$ that can be represented by a nondirectional GIMT hypothesis function. White (1982) also proposed a family of IMTs that could be represented as a set of directional GIMT hypothesis functions of the form: $\mathbf{s}(\mathbf{A}, \mathbf{B}) \equiv \mathbf{S}\ \mathbf{vech}(\mathbf{A} - \mathbf{B})$ where $\mathbf{S} \in \mathcal{R}^{r \times k(k+1)/2}$ has row rank r. Thus, the GIMT hypothesis function introduced in this chapter is a nonlinear generalization of the original Information Matrix Test hypothesis function described by White (1982), which is limited to the representation of linear combinations of the elements of the $\mathbf{A}$ and $\mathbf{B}$ matrices. Note that White's (1982) IMT theory may be viewed as special case of the GIMT theory presented in this chapter.

2.4.1 Classical Full IMT

The Classical Full IMT as described in White (1982, 1994) corresponds to the case where the *Classical Full IMT Hypothesis Function* $\mathbf{s} : \Upsilon^{k \times k} \times \Upsilon^{k \times k} \to \mathcal{R}^r$ is defined such that for every $\mathbf{A}, \mathbf{B} \in \Upsilon^{k \times k}$:

$$\mathbf{s}(\mathbf{A}, \mathbf{B}) = \mathbf{vech}(\mathbf{A}) - \mathbf{vech}(\mathbf{B})$$

yielding the null hypothesis $H_0 : \mathbf{vech}(\mathbf{A}^*) = \mathbf{vech}(\mathbf{B}^*)$. The Classical Full IMT is a nondirectional GIMT, but suffers from the disadvantage of an excessive number of degrees of freedom, $k(k+1)/2$. Thus, the associated excessive variance may yield erratic test performance for typical values of k.

2.4.2 Adjusted Classical GIMT

In simulation studies, we found that the covariance matrix of the GIMT hypothesis function estimator for White's (1982) Classical IMT tended to be singular and so we always used the "adjusted version" of the Classical Full IMT (see the discussion in Sect. 2.3), called the Adjusted Classical GIMT. We emphasize that although the performance of the Adjusted Classical GIMT has not been systematically investigated in previous empirical studies, it is actually a particular member of the family

of directional IMTs explicitly discussed in White's (1982) original paper. We also comment that the performance of the adjusted version of the Classical Full IMT depends upon the researcher's choice of the row dimension g of the adjusted hypothesis projection matrix. Theoretically, the appropriate choice of g is straightforward, but in practice, numerical definitions of the presence of excessive multicollinearity are required. To examine the presence of excessive multicollinearity we compute the ratio of the largest to the smallest eigenvalues as well as the magnitude of the largest and smallest eigenvalues of the GIMT statistic covariance matrix estimator. The performance of the Adjusted Classical GIMT in our simulation studies (and other simulation studies not reported here) tended to vary depending upon how stringently we defined a GIMT statistic covariance matrix estimator as singular or non-singular (Henley et al. 2001, 2004). Our results suggest that care in this regard is a previously unappreciated crucial element to obtaining good IMT statistic performance.

2.5 Eigenspectrum GIMT Family

The essential idea of the classical IMT family (White 1982) was to directly compare linear combinations of the elements of and $\mathbf{A}^*$ and $\mathbf{B}^*$. In this section, we propose a new approach that compares the eigenvalues of $(\mathbf{A}^*)^{-1}$ and $(\mathbf{B}^*)^{-1}$ to determine if the Fisher Information Matrix Equality holds for a probability model.

Assume $\mathbf{A}^*$ is real symmetric positive definite and that all eigenvalues of $\mathbf{A}^*$ are distinct. Let $\lambda_{j,\mathbf{A}^*}$ denote the jth eigenvalue associated with the jth unique orthonormal eigenvector $\mathbf{e}_{j,\mathbf{A}^*}$ of $\mathbf{A}^*$. Then there exists a neighborhood of $\mathbf{A}^*$, $\mathcal{N}_{\mathbf{A}^*} \subseteq \mathcal{R}^{k\times k}$, such that: $\mathbf{A}\boldsymbol{\varepsilon}_{j,\mathbf{A}^*}(\mathbf{A}) = \Lambda_{j,\mathbf{A}^*}(\mathbf{A})\,\boldsymbol{\varepsilon}_{j,\mathbf{A}^*}(\mathbf{A})$ for all $\mathbf{A} \in \mathcal{N}_{\mathbf{A}^*}$ where $\Lambda_{j,\mathbf{A}^*} : \mathcal{N}_{\mathbf{A}^*} \to \mathcal{R}$ is an infinitely differentiable function such that $\Lambda_{j,\mathbf{A}^*}(\mathbf{A}^*) = \lambda_{j,\mathbf{A}^*}$, and $\boldsymbol{\varepsilon}_{j,\mathbf{A}^*} : \mathcal{N}_{\mathbf{A}^*} \to \mathcal{R}^k$ is an infinitely differentiable function such that $\boldsymbol{\varepsilon}_{j,\mathbf{A}^*}(\mathbf{A}^*) = \mathbf{e}_{j,\mathbf{A}^*}$ (Magnus (1985) Theorem 1; also see Magnus and Neudecker (1999) p. 180). Furthermore, $\frac{d\Lambda_{j,\mathbf{A}^*}}{d\mathbf{A}}(\mathbf{A}^*) = \mathbf{e}_{j,\mathbf{A}^*}\left(\mathbf{e}_{j,\mathbf{A}^*}\right)^T$. Let $\Lambda_{\mathbf{A}^*} : \mathcal{N}_{\mathbf{A}^*} \to \mathcal{R}^{\boldsymbol{k}}$ be defined such that for all $\mathcal{N}_{\mathbf{A}^*} \subseteq \mathcal{R}^{k\times k} : \boldsymbol{\Lambda}_{\mathbf{A}^*} \equiv \left[\Lambda_{1,\mathbf{A}^*}, \ldots, \Lambda_{k,\mathbf{A}^*}\right]$. Similarly, when $\mathbf{B}^*$ is real symmetric positive definite with distinct eigenvalues, there exists a neighborhood of $\mathbf{B}^*, \mathcal{N}_{\mathbf{B}^*} \subseteq \mathcal{R}^{k\times k}$, such that: $\mathbf{B}\boldsymbol{\varepsilon}_{j,\mathbf{B}^*}(\mathbf{B}) = \Lambda_{j,\mathbf{B}^*}(\mathbf{B})\,\boldsymbol{\varepsilon}_{j,\mathbf{B}^*}(\mathbf{B})$ for all $\mathbf{B} \in \mathcal{N}_{\mathbf{B}^*}$.

Let $\boldsymbol{\psi} : (0,\infty)^k \times (0,\infty)^k \to \mathcal{R}^r$ be continuously differentiable in both of its arguments. An *Eigenspectrum IMT Family* is a collection of GIMT selection functions where each selection function $\mathbf{s} : \mathcal{N}_{\mathbf{A}^*} \times \mathcal{N}_{\mathbf{B}^*} \to \mathcal{R}^r$ has the property that: $\mathbf{s}(\mathbf{A},\mathbf{B}) = \boldsymbol{\psi}\left(\boldsymbol{\Lambda}_{\mathbf{A}^*}(\mathbf{A}), \boldsymbol{\Lambda}_{\mathbf{B}^*}(\mathbf{B})\right)$ for all $\mathbf{A} \in \mathcal{N}_{\mathbf{A}^*}$ and for all $\mathbf{B} \in \mathcal{N}_{\mathbf{B}^*}$.

2.5.1 Log Eigenspectrum GIMT

Let $\mathbf{log}\,\boldsymbol{\Lambda}_{\mathbf{A}^*}(\mathbf{A}) \equiv \left[\log\Lambda_{1,\mathbf{A}^*}(\mathbf{A}), \ldots, \log\Lambda_{q,\mathbf{A}^*}(\mathbf{A})\right]^T$. The *Log Eigenspectrum GIMT Hypothesis Function* is defined such that for all $\mathbf{A}, \mathbf{B} \in \Upsilon^{k\times k}$:

$$\mathbf{s}(\mathbf{A}, \mathbf{B}) = \left[\log\left(\frac{\Lambda_{1,\mathbf{A}^*}\left(\mathbf{A}^{-1}\right)}{\Lambda_{1,\mathbf{B}^*}\left(\mathbf{B}^{-1}\right)} \right), \ldots, \log\left(\frac{\Lambda_{k,\mathbf{A}^*}\left(\mathbf{A}^{-1}\right)}{\Lambda_{k,\mathbf{B}^*}\left(\mathbf{B}^{-1}\right)} \right) \right]$$
$$= \mathbf{log}\, \mathbf{\Lambda}_{\mathbf{A}^*}\left(\mathbf{A}^{-1}\right) - \mathbf{log}\, \mathbf{\Lambda}_{\mathbf{B}^*}\left(\mathbf{B}^{-1}\right).$$

Thus, the null hypothesis of the log eigenspectrum GIMT is given by:

$$H_0 : \mathbf{s}\left(\mathbf{A}^*, \mathbf{B}^*\right) = \mathbf{log}\, \mathbf{\Lambda}_{\mathbf{A}^*}\left(\left(\mathbf{A}^*\right)^{-1}\right) - \mathbf{log}\, \mathbf{\Lambda}_{\mathbf{B}^*}\left(\left(\mathbf{B}^*\right)^{-1}\right) = \mathbf{0}_k.$$

The Log Eigenspectrum GIMT is a directional GIMT because cases exist where $\mathbf{A}^* \neq \mathbf{B}^*$, yet the eigenspectra of $\mathbf{A}^*$ and $\mathbf{B}^*$ are identical. For example,

$$\mathbf{A}^* \equiv (1) \begin{bmatrix} 0.7025 \\ -0.7117 \end{bmatrix} \begin{bmatrix} 0.7025 & -0.7117 \end{bmatrix}$$
$$+ (2) \begin{bmatrix} -0.7117 \\ -0.7025 \end{bmatrix} \begin{bmatrix} -0.7117 & -0.7025 \end{bmatrix} = \begin{bmatrix} 1.5065 & 0.5 \\ 0.5 & 1.4935 \end{bmatrix}$$

and

$$\mathbf{B}^* \equiv (1) \begin{bmatrix} -0.8206 \\ 0.5715 \end{bmatrix} \begin{bmatrix} -0.8206 & 0.5715 \end{bmatrix}$$
$$+ (2) \begin{bmatrix} 0.5715 \\ 0.8206 \end{bmatrix} \begin{bmatrix} 0.5715 & 0.8206 \end{bmatrix} = \begin{bmatrix} 1.3266 & 0.4690 \\ 0.4690 & 1.6734 \end{bmatrix}$$

both have the same eigenvalues (1 and 2), yet $\mathbf{A}^* \neq \mathbf{B}^*$. On the other hand, such situations are rarely expected to occur in practice, so the Log Eigenspectrum GIMT essentially exhibits the behavioral properties of a nondirectional GIMT.

Note that the number of degrees of freedom for the Log Eigenspectrum GIMT is equal to the number of free parameters k, which is a substantial reduction from the $k(k+1)/2$ degrees of freedom of the Classical Full IMT statistic. Thus, it is expected that the variance of the Log Eigenspectrum GIMT statistic will be less than that of the Classical Full IMT statistic for even moderately small k.

2.5.2 Log Determinant GIMT

The *Log Determinant GIMT Hypothesis Function* is defined such that for every $\mathbf{A}, \mathbf{B} \in \Upsilon^{k \times k}$:

$$\mathbf{s}(\mathbf{A}, \mathbf{B}) = \log\det\left(\mathbf{A}^{-1}\mathbf{B}\right).$$

Thus, the null hypothesis of the Log Determinant GIMT is given by:

$$H_0 : \mathbf{s}\left(\mathbf{A}^*, \mathbf{B}^*\right) = \log\det\left(\left(\mathbf{A}^*\right)^{-1}\mathbf{B}^*\right)$$
$$= \log\det\left(\left(\mathbf{A}^*\right)^{-1}\right) - \log\det\left(\left(\mathbf{B}^*\right)^{-1}\right) = 0.$$

The determinant of $(\mathbf{A}^*)^{-1}$ (i.e., the product of the eigenvalues of $(\mathbf{A}^*)^{-1}$) can be interpreted as a measure of the magnitude of the Hessian covariance matrix $(\mathbf{A}^*)^{-1}$ and is sometimes referred to as the "generalized variance" (Cramér 1946, Sect. 22.7; Serfling 1980, p. 139). Thus, the Log Determinant GIMT hypothesis function compares the generalized variance of the Hessian covariance matrix $(\mathbf{A}^*)^{-1}$ to the generalized variance of the OPG covariance matrix $(\mathbf{B}^*)^{-1}$. The Log Determinant GIMT is expected to have good statistical power for two reasons: (1) it is a *one degree of freedom GIMT* regardless of the complexity of the model or the complexity of the data, and (2) it is equally sensitive to changes in the largest eigenvalues as well as changes in the smallest eigenvalues.

2.5.3 Log Trace GIMT

The Log Trace GIMT is a one-degree of freedom test that compares the magnitude of the Hessian covariance matrix $(\mathbf{A}^*)^{-1}$ to the magnitude of the OPG covariance matrix $(\mathbf{B}^*)^{-1}$ by constructing the Log Trace GIMT hypothesis function. The *Log Trace GIMT hypothesis function* is defined such that for every $\mathbf{A}, \mathbf{B} \in \Upsilon^{k\times k}$:

$$\mathbf{s}(\mathbf{A}, \mathbf{B}) = \log tr\left(\mathbf{A}^{-1}\right) - \log tr\left(\mathbf{B}^{-1}\right).$$

The null hypothesis of the Log Trace GIMT is given by:

$$H_0 : \mathbf{s}\left(\mathbf{A}^*, \mathbf{B}^*\right) = \log tr\left(\left(\mathbf{A}^*\right)^{-1}\right) - \log tr\left(\left(\mathbf{B}^*\right)^{-1}\right) = 0.$$

Note that the Log Trace GIMT hypothesis function may be interpreted as comparing the log sum of the on-diagonal variances of the Hessian covariance matrix $(\mathbf{A}^*)^{-1}$ to that of the OPG covariance matrix $(\mathbf{B}^*)^{-1}$ or equivalently, comparing the log sum of the eigenvalues of $(\mathbf{A}^*)^{-1}$ with that of $(\mathbf{B}^*)^{-1}$.

The Log Trace GIMT compares the Hessian and OPG covariance matrix structures based upon the larger eigenvalues while tending to ignore the smaller eigenvalues. This is equivalent to comparing the sums of the largest on-diagonal variance elements of both covariance matrices. Thus, the Log Trace GIMT is more sensitive to changes in the larger eigenvalues of the covariance matrices and less sensitive to changes in the smaller eigenvalues (i.e., focuses upon the major principal components of the Hessian and OPG covariance matrices). It is thus expected to be a less sensitive GIMT than the Log Determinant GIMT (i.e., it may have reduced statistical power). Depending upon the situation, this latter property of the Log Trace GIMT may be more or less desirable.

2.5.4 Log Generalized Variance GIMT

The *Log Generalized Variance GIMT Hypothesis Function* is defined such that for every $\mathbf{A}, \mathbf{B} \in \Upsilon^{k \times k}$:

$$\mathbf{s}(\mathbf{A}, \mathbf{B}) = \begin{bmatrix} \log\det\left(\mathbf{A}^{-1}\right) - \log\det\left(\mathbf{B}^{-1}\right) \\ \log tr\left(\mathbf{A}^{-1}\right) - \log tr\left(\mathbf{B}^{-1}\right) \end{bmatrix}.$$

The null hypothesis of the Log Generalized Variance GIMT is given by:

$$H_0 : \mathbf{s}\left(\mathbf{A}^*, \mathbf{B}^*\right) = \begin{bmatrix} \log\det\left(\left(\mathbf{A}^*\right)^{-1}\right) - \log\det\left(\left(\mathbf{B}^*\right)^{-1}\right) \\ \log tr\left(\left(\mathbf{A}^*\right)^{-1}\right) - \log tr\left(\left(\mathbf{B}^*\right)^{-1}\right) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

The Log Generalized Variance GIMT is a two degree of freedom GIMT and combines the Log Determinant GIMT, which focuses on both major and minor principal components of the Hessian and OPG covariance matrices, with the Log Trace GIMT, which focuses only upon the major principal components of the Hessian and OPG covariance matrices.

2.5.5 Log GAIC GIMT

Takeuchi (1976; for relevant reviews see Konishi and Kitagawa 1996; Bozdogan 2000) showed that the GAIC defined by the formula:

$$GAIC \equiv -2\log\prod_{i=1}^{n} f\left(\mathbf{X}_i; \hat{\boldsymbol{\theta}}_n\right) + 2TRACE\left(\hat{\mathbf{A}}_n^{-1}\hat{\mathbf{B}}_n\right)$$

is an unbiased estimator of the expected value of $-2\log\prod_{i=1}^{n} f\left(\mathbf{X}_i; \hat{\boldsymbol{\theta}}_n\right)$ in the presence of model misspecification. When the model is correctly specified, then almost surely: $\hat{\mathbf{A}}_n^{-1}\hat{\mathbf{B}}_n \rightarrow \mathbf{I}_k$ where $\mathbf{I}_k$ is the k-dimensional identity matrix. Furthermore, since $2TRACE\left(\hat{\mathbf{A}}_n^{-1}\hat{\mathbf{B}}_n\right) \rightarrow 2k$, GAIC reduces to Akaike's (1973) Akaike Information Criterion (AIC) defined as:

$$AIC \equiv -2\log\prod_{i=1}^{n} f\left(\mathbf{X}_i; \hat{\boldsymbol{\theta}}_n\right) + 2k.$$

Let $\left(\boldsymbol{\Lambda}_{\mathbf{A}^*}(\mathbf{A})\right)^{-1} \equiv \left[\left(\Lambda_{1,\mathbf{A}^*}(\mathbf{A})\right)^{-1}, \ldots, \left(\Lambda_{k,\mathbf{A}^*}(\mathbf{A})\right)^{-1}\right]$ and let $\odot$ denote the Hadamard product (i.e., element-wise vector multiplication) operator. If a simultaneous diagonalization of $\mathbf{A}^*$ and $\mathbf{B}^*$ exists, $TRACE\left[\left(\mathbf{A}^*\right)^{-1}\mathbf{B}^*\right] = (\mathbf{1}_k)^T \left[\left(\boldsymbol{\Lambda}_{\mathbf{A}^*}(\mathbf{A}^*)\right)^{-1} \odot \boldsymbol{\Lambda}_{\mathbf{B}^*}(\mathbf{B}^*)\right]$. This observation suggests a new GIMT called the Log

GAIC IMT. The *Log GAIC GIMT Hypothesis Function* is defined such that for every $\mathbf{A}, \mathbf{B} \in \Upsilon^{k \times k}$:

$$\mathbf{s}(\mathbf{A}, \mathbf{B}) = \log\left(\frac{1}{k}\sum_{j=1}^{k}\left(\frac{\tilde{\lambda}_{j,\mathbf{B}^*}(\mathbf{B})}{\tilde{\lambda}_{j,\mathbf{A}^*}(\mathbf{A})}\right)\right)$$
$$= \log\left(\frac{1}{k} TRACE\left[\left(\tilde{\boldsymbol{\lambda}}_{\mathbf{A}^*}(\mathbf{A})\right)^{-1} \odot \tilde{\boldsymbol{\lambda}}_{\mathbf{B}^*}(\mathbf{B})\right]\right).$$

Thus, the null hypothesis of the Log GAIC IMT is given by:

$$H_0 : \mathbf{s}\left(\mathbf{A}^*, \mathbf{B}^*\right) = \log\left(\frac{1}{k} TRACE\left[\left(\tilde{\boldsymbol{\lambda}}_{\mathbf{A}^*}\left(\mathbf{A}^*\right)\right)^{-1} \odot \tilde{\boldsymbol{\lambda}}_{\mathbf{B}^*}\left(\mathbf{B}^*\right)\right]\right) = 0.$$

The Log GAIC GIMT is also a one-degree of freedom IMT, and is more similar to the Log Determinant GIMT than to the Log Trace GIMT because the Log GAIC GIMT is sensitive to all differences in the eigenspectra of $(\mathbf{A}^*)^{-1}$ and $(\mathbf{B}^*)^{-1}$. However, the Log GAIC GIMT differs from the Log Determinant GIMT because these changes are combined additively instead of multiplicatively.

3 Simulation Studies

In this section we describe and report findings from simulation studies designed to investigate the level and power properties of the five new Eigenspectrum GIMTs and the Adjusted Classical GIMT. Our studies here investigate the reliability of the large sample approximations for estimating Type I error probabilities (level) and evaluate the performance of the new Eigenspectrum GIMTs relative to the new Adjusted Classical GIMT. They also demonstrate the applicability of the new Eigenspectrum GIMTs to detect and assess model misspecification using a realistic epidemiological data analysis problem.

3.1 Epidemiological Data Sample

Our simulation studies were conducted using a random sample ($n = 16{,}189$) of deidentified patient discharges from the Department of Veterans Affairs (VA) Patient Treatment File between October 1, 1995 and September 30, 1996. The "deidentified Extraction Sample" of 16,189 patients included a single binary response variable (ALC) indicating the presence or absence of a primary or secondary discharge diagnosis of either: (i) alcohol dependence (IDC9#303), or (ii) alcohol abuse (ICD9#305.0), based on diagnostic codes from the International Classification of Diseases 9th

Edition (ICD9) (DHHS 1980). The simulation data contains only adults, with the ICD9 alcohol disorders occurring in approximately 20.3 % (3,283) of all patients, where in the sample 4 % are female, 25.1 % are divorced, and 4.2 % are minorities.

3.2 Logistic Regression Models

In this chapter, we investigate the performance of our new GIMTs with respect to binary logistic regression (logit) models (Christensen 1997; Hosmer and Lemeshow 2000; Harrell 2001; Agresti 2002; Collett 2003; Hilbe 2009) in which the probability that a binary response random variable R takes on the values of zero or one is functionally dependent upon $d-1$ predictor variable values denoted by the $d-1$-dimensional vector $\mathbf{u} \in \mathcal{R}^{d-1}$. Define a logistic regression model using

$$\log\left[\frac{p\,(R=1|\mathbf{u};\boldsymbol{\beta})}{p\,(R=0|\mathbf{u};\boldsymbol{\beta})}\right] = \boldsymbol{\beta}^T \begin{bmatrix}\mathbf{u}\\ 1\end{bmatrix}$$

where the last element of the k-dimensional parameter vector $\boldsymbol{\beta}$ corresponds to the intercept parameter. In order to relate this logistic regression model to the discussion in Sect. 2, let $R \equiv x_1$ and $\mathbf{u} \equiv [x_2, \ldots, x_d]$ so that $\mathbf{x} \equiv [R, \mathbf{u}] \in \mathcal{R}^d$ and let $\boldsymbol{\theta} \equiv \boldsymbol{\beta} \in \Theta \subseteq \mathcal{R}^k$ where $d = k$. Using this notation, we define

$$f\,(\mathbf{x};\boldsymbol{\theta}) \equiv [x_1\,p\,(R=1|\mathbf{u};\boldsymbol{\beta}) + (1-x_1)\,p\,(R=0|\mathbf{u};\boldsymbol{\beta})]\,p\,(x_2,\ldots,x_d)$$

where the joint predictor density $p\,(x_2, \ldots, x_d)$ is not functionally dependent upon $\boldsymbol{\beta} \in \mathcal{R}^d$. Because of this latter property, the GIMT formulas are not functionally dependent on $p\,(x_2, \ldots, x_d)$. Thus in the *i.i.d.* case the log-likelihood for a logistic regression model with sample size n is

$$L\,(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{R_i \ln[p\,(R_i=1|\mathbf{u}_i;\boldsymbol{\beta})] + (1-R_i)\ln[1-p\,(R_i=1|\mathbf{u}_i;\boldsymbol{\beta})]\}$$

where

$$p\,(R=1|\mathbf{u};\boldsymbol{\beta}) = \left(1+\exp\left[-\left(\mathbf{u}^T\boldsymbol{\beta}\right)\right]\right)^{-1}.$$

3.2.1 Logistic Regression Model with Binary Predictors

We first fitted a logistic regression model to the $n = 16{,}189$ deidentified Extraction Sample using maximum likelihood estimation to predict the presence or absence of "alcohol-disorder" (ALC) from the binary predictors "female" (FEMALE), "married" (MARRIED), recoded categorical predictor ethnicity containing "black" (BLACK) and "white" (WHITE), and the recoded predictor "age" (AGE).

The ethnicity variable was recoded into a three category design variable (white, black, other) using reference cell coding (Hosmer and Lemeshow 2000) where "other" is the reference variable. Also, the numerical AGE predictor was trichotomized into a three category design variable by applying optimally estimated cut values $\gamma_1 = 55.4$ and $\gamma_2 = 68.2$ (Henley et al. 2000; Kashner et al. 2002, 2003, 2007, 2010) where the first binary design variable AGE_1 (age ≤ 55.4) is the reference variable.

In addition to reporting our model fit results using a negative log-likelihood score, we report fitness results in terms of a GAIC, also known as the Takeuchi Information Criterion (TIC) (Takeuchi 1976; Konishi and Kitagawa 1996; Bozdogan 2000). GAIC is a misspecification robust extension of the Akaike Information criterion (AIC) (Akaike 1973; Burnham and Anderson 2002, pp. 65, 362–372). The resulting fitted logistic regression model had a negative log-likelihood of 6,718.2 (GAIC/$2n$ = 0.415420, $p = 0.0000$) with estimated parameter values

$$\hat{\beta}_0 = -0.7397, \quad \hat{\beta}_1 = -1.3099, \quad \hat{\beta}_2 = -2.2946, \quad \hat{\beta}_3 = -1.4249,$$
$$\hat{\beta}_4 = -0.9784, \quad \hat{\beta}_5 = 1.0000, \quad \hat{\beta}_6 = 0.6822$$

respectively for the intercept, AGE_2 ($55.4 < \text{age} \leq 68.2$), AGE_3 ($68.2 < \text{age} \leq 85$), FEMALE, MARRIED, BLACK, and WHITE predictors. Wald tests computed using robust standard errors (e.g., Wald 1943; White 1982; Golden 1996) showed each estimated parameter value was significantly different from zero ($p < 0.001$). All six GIMTs applied to this model failed to reject the null hypothesis (Adjusted Classical, $p = 0.6113$; Log Eigenspectrum, $p = 0.3618$; Log Determinant, $p = 0.6138$; Log Trace, $p = 0.4063$; Log Generalized Variance, $p = 0.6890$; Log GAIC, $p = 0.6004$) indicating no evidence of model misspecification. Thus, simulated data samples generated from this fitted model were expected to be more representative of real world data.

3.2.2 Alternative Logistic Regression Model with Numerical and Binary Predictors

We also fitted a different (alternative) logistic regression model that replaced the trichotimized age predictor with the numerical predictor for "age" (AGE*) and added a "divorced" (DIVORCED*) binary variable so each model had seven free parameters. The model was otherwise identical to the first one. The resulting fitted logistic regression model had a negative log-likelihood of 6,743 (GAIC/$2n = 0.416965$, $p = 0.0000$) with estimated parameter values

$$\hat{\beta}_0 = 1.8448, \quad \hat{\beta}_1 = -0.0646, \quad \hat{\beta}_2 = -1.6057, \quad \hat{\beta}_3 = -0.7972,$$
$$\hat{\beta}_4 = 0.3353, \quad \hat{\beta}_5 = 1.0082, \quad \hat{\beta}_6 = 0.7065$$

respectively for the intercept, AGE*, FEMALE, MARRIED, DIVORCED*, BLACK, and WHITE predictors. Wald tests computed using robust standard errors

(e.g., Wald 1943; White 1982; Golden 1996) again showed each estimated parameter value was significantly different from zero ($p < 0.001$). All six GIMTs applied to the alternative logit model rejected the null hypothesis (Adjusted Classical, $p = 0.0000$; Log Eigenspectrum, $p = 0.0000$; Log Determinant, $p = 0.0028$; Log Trace, $p = 0.0282$; Log Generalized Variance, $p = 0.0112$; Log GAIC, $p = 0.0026$) indicating the presence of model misspecification.

In practice, researchers may inadvertently use a misspecified model that nevertheless provides a good fit, as measured by log-likelihood or GAIC, to the observed data. We selected an alternative logistic regression model, which provided a fit (GAIC/$2n = 0.416965$, $p = 0.0000$) to the observed data that is comparable to the fit (GAIC/$2n = 0.415420$, $p = 0.0000$) of the original logit model described in Sect. 3.2.1. This difference in model fit was not statistically significant ($p = 0.1960$) using the Discrepancy Risk Model Selection Test (DRMST) (Vuong 1989; Golden 2000, 2003; Henley et al. 2000, 2003, 2008) for comparing nonnested and possibly misspecified models.

3.3 Simulation Study

3.3.1 GIMT Level and Power Estimation Procedures

The procedure for estimating the observed level of a GIMT is shown in Fig. 1. Four simulated data samples of n^* records ($n^* = 1{,}619$, $n^* = 4{,}047$, $n^* = 8{,}095$, and $n^* = 16{,}189$) were generated by sampling with replacement from the original representative sample (see Politis et al. 1999; Davison et al. 2003). This process was repeated m times for each of four sample sizes. The conditional probability for the binary ALC outcome variable was then computed and assigned the value one or zero, based on the minimum probability of decision error rule, for each record using predictor values and the estimated coefficients of the seven-parameter logistic regression model with binary predictors. Thus, all simulated data samples had predictor values with synthetic ALC outcome values that had been generated from the specified logistic regression model estimated on the original representative sample ($n = 16{,}189$). To calculate level estimation results, we then fit the logistic regression model to each of the m simulated data samples for the four sample sizes and computed 10,000 significance levels in the range of zero to one for all the GIMTs. The percentage of times that a GIMT incorrectly rejects the null hypothesis of correct specification as the "observed incorrect rejection rate" or "observed level" was calculated.

The procedure for estimating the observed power of a GIMT is shown in (Fig. 2). In this experiment we created an alternative logistic regression model by changing two of the six binary predictor variables in the logistic regression model from the level estimation procedure (Fig. 1). As previously described, the numerical AGE and binary DIVORCED predictor variables in the original representative data sample replaced the binary design variables AGE_2 and AGE_3. This predictor variable change introduced a relatively subtle, but realistic misspecification into the alterna-

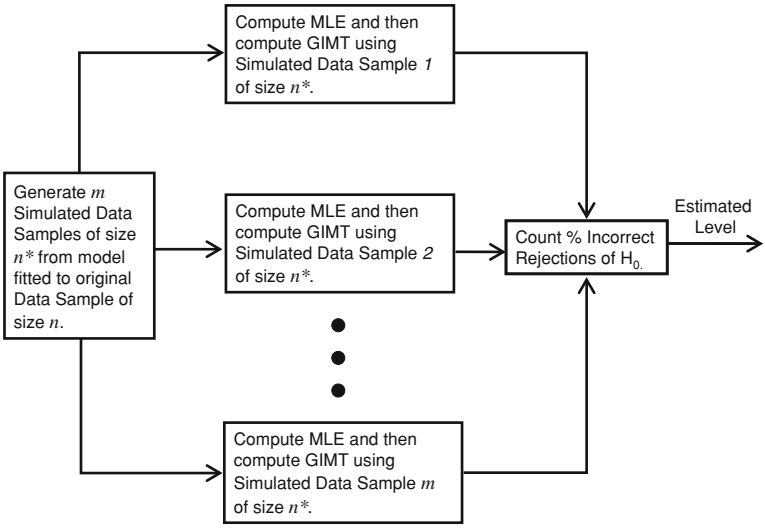

Fig. 1 Simulation procedure for estimation of level

tive model because the known (i.e., simulated) data generating process stems from the original logistic regression model containing only binary predictors. Further, the use of observationally equivalent original and alternative logit models (see discussion in Sect. 3.2.2) for the simulation design minimizes the confounding issue of model fit (GAIC) with specification, thus enabling the effects of model specification (goodness-of-fit) on GIMT performance to be more effectively studied. To calculate power estimation results, we then fit the alternative logistic regression model to each of the simulated data samples from the level analysis for the four sample sizes and computed 10,000 significance levels in the range of zero to one for all the GIMTs. The percentage of times that a GIMT correctly rejects the null hypothesis of correct specification as the "observed correct rejection rate" or "observed power" was calculated.

In our simulation studies, an MLE was defined as a set of parameter values such that the sup norm of the gradient of the negative log-likelihood evaluated at the MLE was less than 1e−8. Further, we avoided fitting models to degenerate simulated data by omitting samples with condition numbers greater than 4.5e+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 $m = 100{,}000$ simulated data samples of size n^* was reached. The sample sizes n^* for the simulated data represented 10 %, 25 %, 50 %, and 100 % of the original 16,189 record data set. In all simulations, we utilized the Hessian-GIMT statistic as defined in Sect. 2.3.

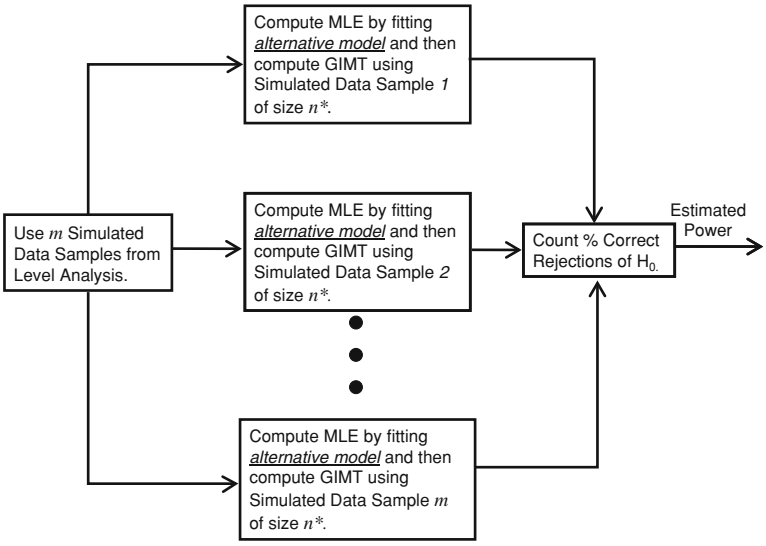

Fig. 2 Simulation procedure for estimation of power

3.3.2 Simulation Study Results

In this section we present level-discrepancy and level-power simulation results for the proposed GIMTs.

Level-Discrepancy Analyses

We first examined the performance for the six GIMTs using a *P*-value plot analysis (Davidson and MacKinnon 1998). This method plots the empirical level (observed rejection rate of the null hypothesis, i.e., Type I error) of a GIMT against its nominal level (specified rejection rate of the null hypothesis). To enable *P*-value plot comparisons, we also define a summary deviation measure for the *level-discrepancy* as the root mean square error (RMSE) between empirical and nominal levels over the specified range of interest (e.g., [0, 0.1] or [0, 1.0]). Thus, an ideal estimation of the Type I error rate corresponds to a level-discrepancy of zero (i.e., RMSE $= 0$). In our studies, the level-discrepancy for each GIMT was estimated on simulated data for each sample size.

The Adjusted Classical GIMT is a member of the family of Classical IMTs that includes White (1982) Full IMT. Figure 3 depicts the *P*-value plots with level-discrepancies for the Adjusted Classical GIMT on 100,000 simulated data samples for n ranging from 1,619 to 16,189 for level ranges on [0, 0.10]. These results show that the level-discrepancy deviation decreases from 0.0261 to 0.0091 RMSE as sample size increases, thus approaching an ideal estimation Type I error rate at larger sample sizes. Further, the exhibited Type I error rate convergence for the Adjusted Classical GIMT indicated level-discrepancy performance that was much better than the performance of the Classical Full IMT (not shown). We attribute this to the par-

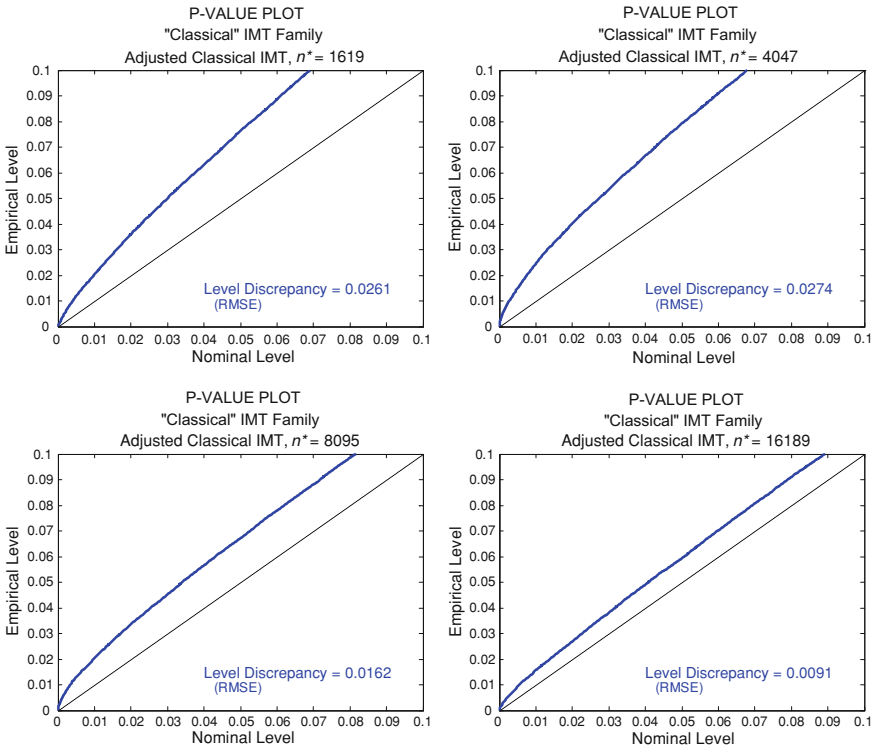

Fig. 3 *P*-value plots for the White's (1982) Adjusted Classical GIMT show empirical level [0, 0.1] versus nominal level [0, 0.1] by sample size. The displayed level-discrepancy is defined as the root mean square error (RMSE) between the empirical and nominal levels. Thus, an ideal estimation of the Type I error rate corresponds to a discrepancy between the empirical (simulated) and nominal levels of zero (i.e., RMSE = 0). The data points on the graphs are computed for 100,000 simulated data samples for $n^* = 1{,}619$, $n^* = 4{,}047$, $n^* = 8{,}095$ and $n^* = 16{,}189$

ticular care with which singularity or near-singularity of the test statistic covariance matrix is handled.

Next, we present the simulation results for the new Log Eigenspectrum GIMT. Figure 4 depicts the *P*-value plots with level-discrepancies for the Log Eigenspectrum GIMT on 100,000 simulated data samples for n ranging from 1,619 to 16,189, which again shows RMSE decreasing as sample size increases. Notably, the level-discrepancy (RMSE = 0.0030) for the Log Eigenspectrum GIMT at $n = 16{,}189$ is less than the level-discrepancy (RMSE = 0.0091) for the Adjusted Classical GIMT (Fig. 3).

The simulation results for the new Log GAIC GIMT, which is a directional GIMT, are also presented for comparison. Figure 5 shows the *P*-value plots with level-discrepancies for the Log GAIC GIMT on 100,000 simulated data samples for n ranging from 1,689 to 16,189. Again, the empirical and nominal levels of interest

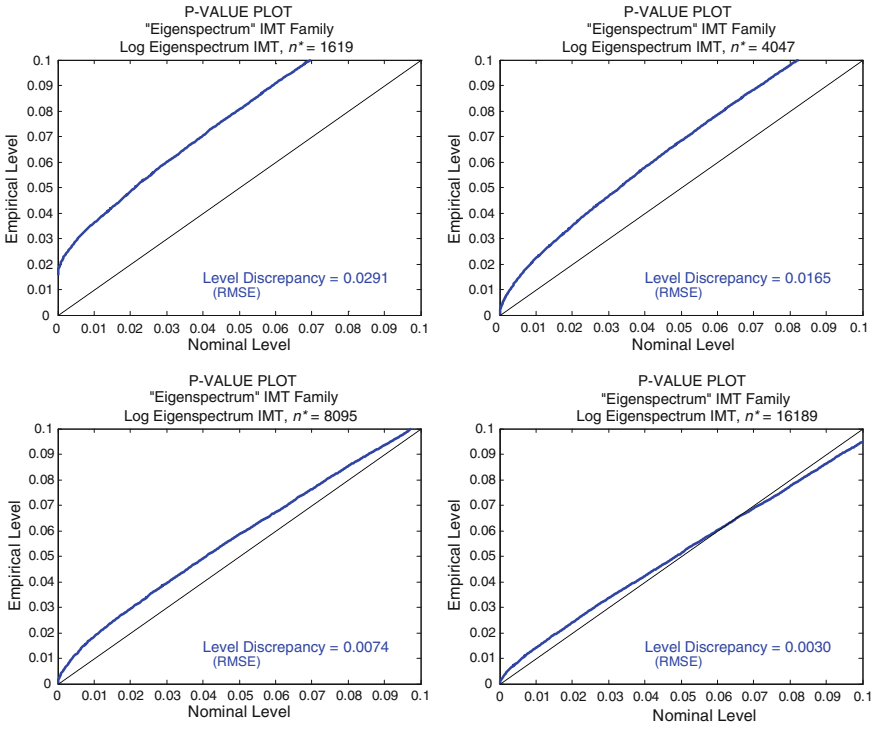

Fig. 4 *P*-value plots for the Log Eigenspectrum GIMT show empirical level [0, 0.1] versus nominal level [0, 0.1] by sample size. The level-discrepancy is defined as the deviation measured by root mean square error (RMSE) between the empirical and nominal levels. Thus, an ideal estimation of the Type I error rate corresponds to a discrepancy between the empirical (simulated) and nominal levels of zero (i.e., RMSE = 0). The data points on the graphs are computed for 100,000 simulated data samples for $n^* = 1{,}619$, $n^* = 4{,}047$, $n^* = 8{,}095$ and $n^* = 16{,}189$. The level-discrepancy (RMSE = 0.0030) at $n = 16{,}189$ for the Log Eigenspectrum GIMT with seven degrees of freedom is less than the level-discrepancy (RMSE = 0.0091) reported for the Adjusted Classical GIMT (Fig. 3), which has up to 28 degrees of freedom

range over [0, 0.10]. These simulation results show the level-discrepancy for the Log GAIC GIMT is converging to zero as sample size increases. The level-discrepancy (RMSE = 0.0045) at $n = 16{,}189$ for the directional Log GAIC GIMT is greater than the level-discrepancy (RMSE = 0.0030) reported for the Log Eigenspectrum GIMT (Fig. 4), but less than the level-discrepancy (RMSE = 0.0091) reported for the Adjusted Classical Full GIMT (Fig. 3). A similar pattern of results was observed using the *P*-value plot analyses for the remaining three new directional Eigenspectrum GIMTs. All observed rejection rates were very close to the nominal levels.

The level-discrepancy performance of all GIMTs is depicted in Fig. 6, which displays *P*-value plot results as a function of sample size. As shown, the new Eigenspectrum GIMTs exhibit excellent performance for large sample sizes. In addition, they exhibited better performance than the Adjusted Classical GIMT with

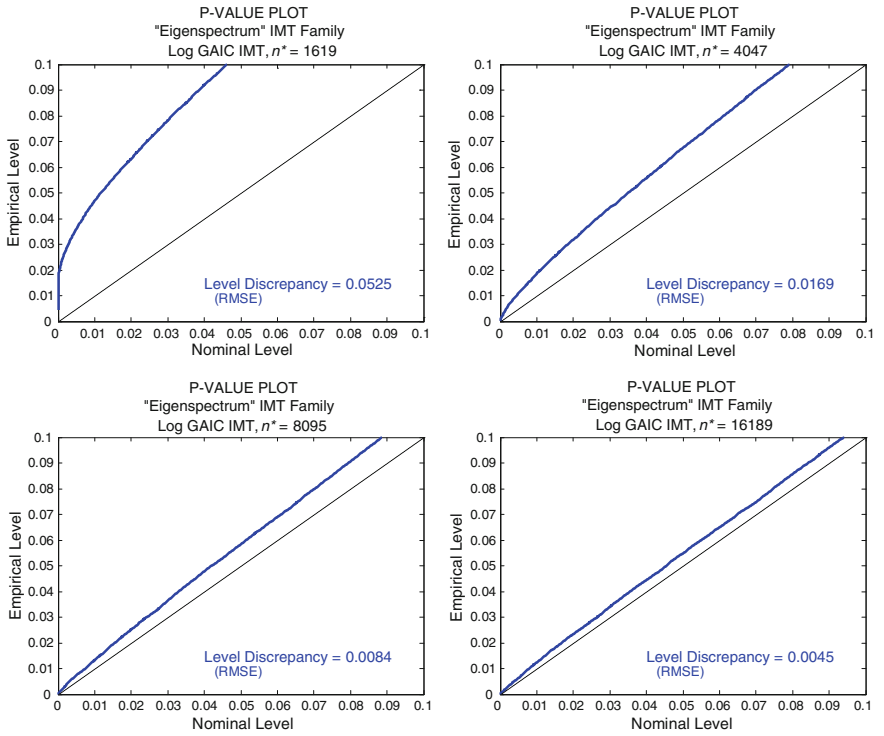

Fig. 5 *P*-value plots for the directional Log GAIC GIMT show empirical level [0, 0.1] versus nominal level [0, 0.1] by sample size. The level-discrepancy is defined as the deviation measured by root mean square error (RMSE) between the empirical and nominal levels. Thus, an ideal estimation of the Type I error rate corresponds to a discrepancy between the empirical (simulated) and nominal levels of zero (i.e., RMSE = 0). The data points on the graphs are computed for 100,000 simulated data samples for $n^* = 1{,}619$, $n^* = 4{,}047$, $n^* = 8{,}095$ and $n^* = 16{,}189$. The level-discrepancy (RMSE = 0.0045) at $n = 16{,}189$ for the directional Log GAIC GIMT with one-degree of freedom is larger than the level-discrepancies obtained for the Log Eigenspectrum GIMT (RMSE = 0.0030), though smaller than the Adjusted Classical GIMT (RMSE = 0.0091) shown respectively in Figs. 3 and 4

level-discrepancies approaching zero in all cases. The Log Eigenspectrum GIMT exhibited the best (i.e., smallest) level-discrepancy performance of all GIMTs at larger sample sizes.

The observed rejection rates (estimated Type I errors) for each of the six new GIMTs are reported in Table 1 for the nominal significance levels of 0.001, 0.005, 0.01, 0.025, 0.05, and 0.10 for the full sample size of $n = 16{,}189$. The simulated standard errors of the estimated Type I error rates are shown in parentheses. Note that these standard errors will converge to zero as $m \to \infty$ for a fixed sample size $n = 16{,}189$. Our findings show that the estimated Type I error rates for all six new GIMTs are, in general, very close to their specified error rates. The Log Eigenspectrum GIMT exhibited the smallest level-discrepancy of all GIMTs at the

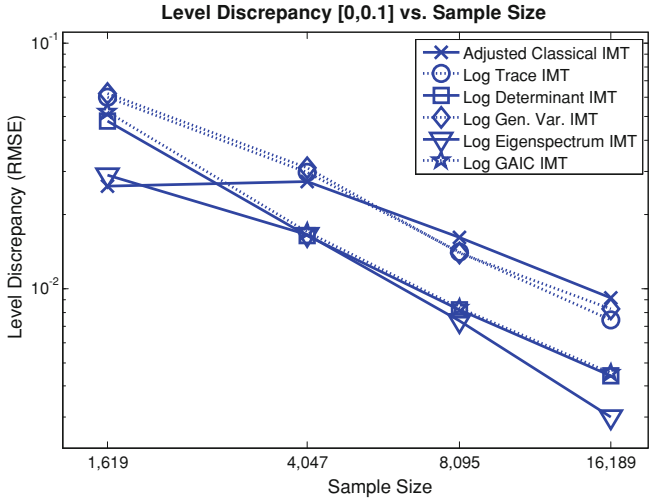

Fig. 6 Level-discrepancy performance by sample size for the six GIMTs in the simulation study. Each data point corresponds to 100,000 simulated data samples. The Adjusted Classical GIMT and all the Eigenspectrum GIMTs exhibit level-discrepancy convergence towards zero as sample size increases. The Log Eigenspectrum GIMT exhibited the smallest level-discrepancy of all GIMTs at the larger sample sizes

larger sample sizes. We also performed additional simulation studies (Henley et al. 2001, 2004), and found that the performance of the six new GIMTs was always better than White's (1982) Classical Full IMT.

Level-Power Analyses

Next, we perform a level-power analysis to examine all six GIMTs by generating a level-power curve (Davidson and MacKinnon 1998) for each GIMT. A level-power curve plots the power (i.e., 1-Type II error) of a statistical test as a function of the level (rejection rate or Type I error). Accordingly, we interpret a statistical test as a binary classifier that divides the decision space into two regions: *reject* or *fail to reject* (Wickens 2002; Pepe 2004, p. 152).

An important performance measure for the evaluation of binary classifiers is the Area Under the Response Operating Characteristic Curve (AUROC; also known as AUC) (Hanley and McNeil 1982; Bradley 1997; Wickens 2002; Pepe 2004; Fawcett 2006). In the context of a level-power analysis, this corresponds to the area under the level-power curve. A level-power AUROC equal to one corresponds to perfect classification (i.e. test) performance. Figure 7 shows the level-power curves for the Log Eigenspectrum GIMT for $m = 100{,}000$ simulated data samples with sample sizes of $n^* = 1{,}619$, $n^* = 4{,}047$, $n^* = 8{,}095$, and $n^* = 16{,}189$. The Log Eigenspectrum GIMT exhibited ideal level-power performance (AUROC = 1.00) at the two larger samples sizes (not shown).

Level-power curves for all sample sizes ($n^* = 1{,}619$, $n^* = 4{,}047$, $n^* = 8{,}095$, and $n^* = 16{,}189$) were also generated for the other GIMTs using 100,000 simulated

Table 1 Empirical level (observed Type I error rates) obtained in the simulation studies for pre-specified (nominal) significance levels: 0.001, 0.005, 0.01, 0.025, 0.05, and 0.10

Generalized IM test	Nominal level					
	$p = 0.001$	$p = 0.005$	$p = 0.01$	$p = 0.025$	$p = 0.05$	$p = 0.10$
Adjusted classical[a,b] (df $\leq$ 28)	0.0029 (0.0002)	0.0093 (0.0003)	0.0156 (0.0004)	0.0326 (0.0006)	0.0593 (0.0007)	0.1112 (0.0010)
Log eigenspectrum[b] (7 df)	0.0029 (0.0002)	0.0084 (0.0003)	0.0140 (0.0004)	0.0289 (0.0005)	0.0511 (0.0007)	0.0947 (0.0009)
Log determinant (1 df)	0.0013 (0.0001)	0.0061 (0.0002)	0.0120 (0.0003)	0.0282 (0.0005)	0.0548 (0.0007)	0.1063 (0.0010)
Log trace (1 df)	0.0017 (0.0001)	0.0072 (0.0003)	0.0134 (0.0004)	0.0306 (0.0005)	0.0576 (0.0007)	0.1101 (0.0010)
Log generalized variance (2 df)	0.0017 (0.0001)	0.0074 (0.0003)	0.0139 (0.0004)	0.0310 (0.0005)	0.0586 (0.0007)	0.1110 (0.0010)
Log GAIC (1 df)	0.0016 (0.0001)	0.0063 (0.0003)	0.0122 (0.0003)	0.0284 (0.0005)	0.0547 (0.0007)	0.1063 (0.0010)

Results are for the six GIMTs where the sample size $n^* = 16,189$ and the number of simulated data samples $m = 100,000$. Bootstrapped standard errors, reflecting simulation sampling error, are shown in parentheses. In general, empirical levels agreed with the prespecified nominal significance levels

[a]Adjusted to remove multicollinearity from the Classical Full IMT selection statistic covariance matrix

[b]Degrees of freedom (df) is a function of the number of free parameters

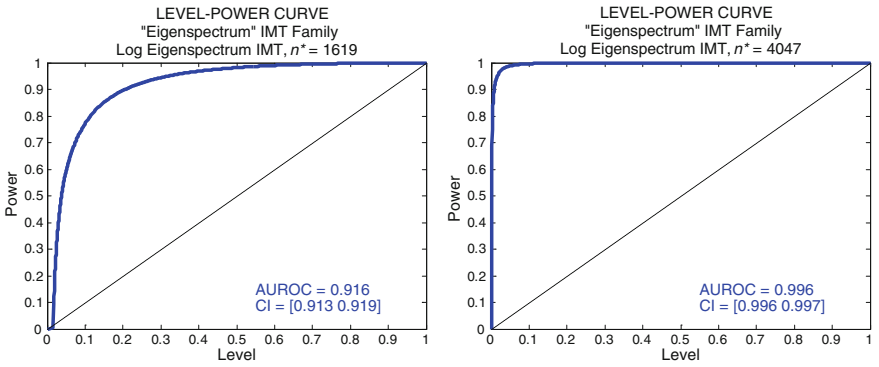

Fig. 7 Level-power curves for Log Eigenspectrum GIMT exhibit convergence to ideal GIMT decision performance as sample size increases using simulated epidemiological data. Each data point on the graphs represents 100,000 simulated data samples under the null and alternative hypotheses for the sample sizes $n^* = 1{,}619$ and $n^* = 4{,}047$ respectively. This two graph sequence depicts convergence to an ideal level-power curve (i.e., AUROC = 1.00). The level-power performance for the larger sample sizes $n^* = 8{,}095$ and $n^* = 16{,}189$ (not shown) achieved an ideal AUROC = 1.00

data samples per data point under the null and alternative hypotheses. Figure 8 depicts the level-power performance of all GIMTs as a function of sample size. As shown, the new Log Eigenspectrum GIMT and the Adjusted Classical GIMT have good power for both small and large sample sizes, although all of the GIMTs exhibit useful power for large sample sizes. A possible explanation for the increased power of the Log Eigenspectrum and the Adjusted Classical GIMTs is that these GIMTs test more comprehensive composite null hypotheses that result in increased opportunities to detect the presence of model misspecification.

4 Summary and Conclusions

In this chapter, we have introduced a general approach to the development of Generalized Information Matrix Tests that are intended to detect the presence of model misspecification. Such situations occur when the Hessian inverse covariance matrix $\mathbf{A}^*$ and the OPG inverse covariance matrix $\mathbf{B}^*$ are different. In particular, we introduced the new Generalized Information Matrix Test (GIMT) that tests $H_0 : \mathbf{s}(\mathbf{A}^*, \mathbf{B}^*) = \mathbf{0}_r$ and provided a Wald test version of the GIMT based on the asymptotic distribution of $n^{1/2}\hat{\mathbf{s}}_n \equiv n^{1/2}\mathbf{s}\left(\hat{\mathbf{A}}_n, \hat{\mathbf{B}}_n\right)$, along the lines of (White 1982, Theorem 4.2). For a given GIMT Selection Hypothesis Function, we also provided six distinct formulas for computing each GIMT test statistic and introduced the new concept of an "adjusted" GIMT statistic for dealing with issues of multicollinearity and demonstrated its utility by applying it to White's (1982) Classical Full IMT.

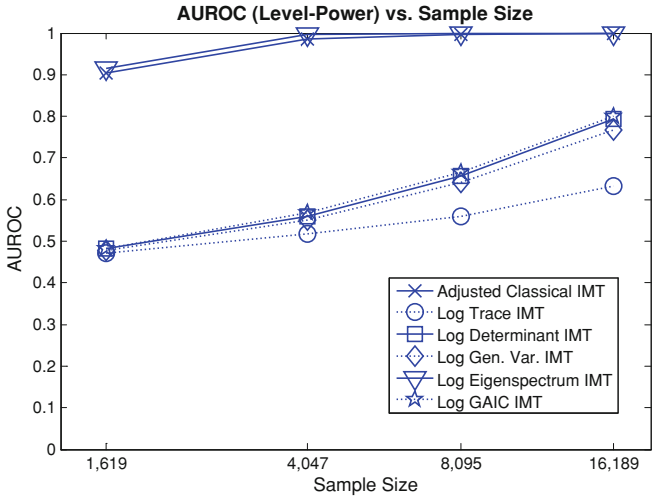

Fig. 8 AUROC (level-power) performance as a function of sample size for the six GIMTs in the simulation study. Each data point corresponds to 100,000 simulated data samples under the null and 100,000 simulated data samples under the alternative hypothesis. The Adjusted Classical and Log Eigenspectrum GIMTs converged at a faster rate to ideal level-power (i.e., AUROC = 1.00) as sample size increases, indicating more efficient level-power performance when compared to the other GIMTs

Further, we introduced the idea of constructing GIMTs by comparing nonlinear functions of the eigenspectra of the Hessian and OPG covariance matrices. Next, we developed five new GIMTs based upon the Eigenspectrum GIMT Family. These are the Log Eigenspectrum GIMT, Log Determinant GIMT, Log Trace GIMT, Generalized Variance GIMT, and Log GAIC GIMT. Analytic formulas for these five new Eigenspectrum GIMTs were derived and implemented in computer software.

We studied the performance of these five new Eigenspectrum GIMTs and an adjusted version of White's (1982) Classical Full IMT (i.e., Adjusted Classical GIMT) in a series of simulation experiments using a realistic 16,189 record data set typical of data encountered in epidemiological studies. By comparing a correctly specified model and a misspecified model with approximately equivalent fits to the observed data, our simulation studies focus specifically on the effects of model misspecification. Using *P*-value plots and level-power plots, we found that the Adjusted Classical GIMT and the five new Eigenspectrum GIMTs exhibited reliable performance, in the sense that their asymptotic behavior was correctly captured by the large sample statistical theory under the null. In particular, the empirically observed Type I error rates for all six new GIMTs were very close to their nominal error rates. Additionally, they also exhibited useful power. This is in stark contrast to the familiar poor performance of the unadjusted form of White (1982) Classical Full IMT (e.g., Davidson and MacKinnon 1992; Stomberg and White 2000; Aparicio and Villanua 2001).

For the larger sample sizes, the level-discrepancy performance (i.e., Type I error performance) of the high degree of freedom GIMT (i.e., Log Eigenspectrum) was better than those of all the low degree of freedom GIMTs (i.e., Log Determinant, Log Trace, Log Generalized Variance, Log GAIC), which in turn exceeded the performance of the high degree of freedom Adjusted Classical GIMT. However, the power performance (i.e., Type II error performance) of the Adjusted Classical and Log Eigenspectrum GIMTs was always superior to that of the low degree of freedom GIMTs over all sample sizes. We conjecture that the reduced variance of the low degree of freedom GIMTs decreased the efficiency of the large sample approximation when compared to the Log Eigenspectrum GIMT. We further conjecture that because the Eigenspectrum GIMTs have fewer degrees of freedom they were more robust to sampling error when compared to the Adjusted Classical GIMT, which adjusts its degrees of freedom to control for multicollinearity. The greater power of the larger degree of freedom GIMTs is most likely explained by noting that these GIMTs are simultaneously testing multiple hypotheses, thus providing additional opportunities to detect model misspecification.

We used our Adjusted Classical GIMT instead of White's (1982) Classical Full IMT because in additional simulation studies not reported here, the asymptotic covariance matrix for the Classical Full IMT was frequently observed to be singular and exhibited much worse performance in our investigations. However, in all cases, the level-discrepancy and the level-power performance of the new Adjusted Classical GIMT and the new Eigenspectrum GIMTs were superior to those of the Classical Full IMT. Moreover, the reliable performance of the Adjusted Classical GIMT as compared to the Classifical Full IMT is notable, and we emphasize that this GIMT is a special case of the original IMT theory proposed by White (1982).

In conclusion, the generalized IMT theory (Henley et al. 2001, 2004, 2008) presented here provides a novel framework for developing a wide range of model specification tests for a broad range of probability models. In particular, the new Eigenspectrum Family GIMTs have degrees of freedom less than or equal to k, in contrast to the Classical Full IMT (White 1982), which has $k(k+1)/2$ degrees of freedom for a k-parameter model. Further, our five new Eigenspectrum GIMTs and new Adjusted Classical GIMT for logistic regression models all have appealing level and power properties, as seen in a series of simulation experiments involving a realistic epidemiologic modeling problem. These six new GIMTs are therefore expected to provide useful new tools for detecting model misspecification across a broad class of probability models (Hastie and Tibshirani 1986; McCullagh and Nelder 1989; Wei 1998; Harrell 2001; Hastie et al. 2009), thus decreasing the chance that a misspecified model is inadvertently used to make inferences in practice. The reduction of incorrect statistical inferences, in turn, has fundamentally important consequences for making critical decisions in many areas, including the social, behavioral, and physical sciences, as well as engineering, financial, medical, and public health research (Kashner et al. 2002, 2003, 2007, 2010).

Acknowledgments This research was made possible by grants from the National Cancer Institute (NCI) (R44CA139607, PI: S.S. Henley) and the National Institute on Alcohol Abuse and Alcoholism

(NIAAA) (R43AA014302, PI: S.S. Henley; R43/44AA013351, PI: S.S. Henley; R44AA011607, PI: S.S. Henley) under the Small Business Innovation Research (SBIR) program. The authors wish to gratefully acknowledge this support. This chapter reflects the authors' views and not necessarily the opinions or views of the NCI or the NIAAA. The authors would also like to thank the anonymous referee for helpful comments and suggestions.

References

Agresti, A.: Categorical data analysis. New York: Wiley-Interscience, 2002.

Akaike, H.: "Information theory and an extension of the maximum likelihood principle", 1973.

Alonso, A., S. Litière, and G. Molenberghs: "A family of tests to detect misspecifications in the random-effects structure of generalized linear mixed models", Computational Statistics and Data Analysis, 52(2008), 4474–4486.

Aparicio, T., and I. Villanua: "The asymptotically efficient version of the information matrix test in binary choice models. A study of size and power", Journal of Applied Statistics, 28(2001), 167–182.

Archer, K. J., and S. Lemeshow: "Goodness-of-fit test for a logistic regression model fitted using survey sample data", The Stata Journal, 6(2006), 97–105.

Arminger, G., and M. E. Sobel: "Pseudo-maximum likelihood estimation of mean and covariance structures with missing data", Journal of the American Statistical Association, 85(1990), 195–203.

Begg, M. D., and S. Lagakos: "On the consequences of model misspecification in logistic regression", Environmental Health Perspectives, 87(1990), 69–75.

Bera, A. K., and S. Lee: "Information Matrix Test, Parameter Heterogeneity and ARCH: A Synthesis", The Review of Economic Studies, 60(1993), 229–240.

Bertolini, G., R. D'Amico, D. Nardi, A. Tinazzi, and G. Apolone: "One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model", Journal of Epidemiology and Biostatistics, 5(2000), 251–3.

Box, E. P., G. M. Jenkins, and G. C. Reinsel: Time Series Analysis: Forecasting and Control. New York: John Wiley & Sons, 2008.

Bozdogan, H.: "Akaike's Information Criterion and Recent Developments in Information Complexity", Journal of Mathematical Psychology, 44(2000), 62–91.

Bradley, A. P.: "The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms", Pattern Recognition, 30(1997), 1145–1159.

Burnham, K. P., and D. R. Anderson: Model selection and multimodel inference : a practical information-theoretic approach. New York: Springer, 2002.

Chesher, A.: "The information matrix test: Simplified calculation via a score test interpretation", Economics Letters, 13(1983), 45–48.

Chesher, A., and R. Spady: "Asymptotic Expansions of the Information Matrix Test Statistic", Econometrica, 59(1991), 787–815.

Christensen, R.: Log-Linear Models and Logistic Regression. Springer Texts in, Statistics, 1997.

Collett, D.: Modelling Binary Data. Chapman & Hall/CRC, 2003.

Copas, J.B.: "Unweighted sum of squares test for proportions", Applied Statistics, 38(1989), 71–80.

Cox, D.R.: "Role of models in statistical analysis", Statistical Science, 5(1990), 169–174.

Cramér, H.: Mathematical Methods of Statistics. Princeton: Princeton University Press, 1946.

Davidson, R., and J. G. MacKinnon: "A New Form of the Information Matrix Test", Econometrica, 60(1992), 145–157.

Davidson, R., and J. G. MacKinnon: "Graphical Methods for Investigating the Size and Power of Hypothesis Tests", The Manchester School, 66(1998), 1–26.

Davison, A. C., D. V. Hinkley, and G. A. Young: "Recent Developments in Bootstrap Methodology", Statistical Science, 18(2003), 141–157.
Davison, A. C., and C. L. Tsai: "Regression model diagnostics", International Statistical Review, 60(1992), 337–353.
Deng, X., S. Wan, and B. Zhang: "An improved goodness-of-test for logistic regression models based on case-control data by random partition", Communications in statistics: Simulations and computation, 38(2009), 233–243.
Dhaene, G., and D. Hoorelbeke: "The information matrix test with bootstrap-based covariance matrix estimation", Economics Letters, 82(2004), 341–347.
DHHS: "The International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM). DHHS Publication No. (PHS) 80–1280", Washington D.C.: Department of Health and Human Services, 1980.
Farrington, C.P.: "On assessing goodness of fit of generalized linear models to sparse data", Journal of the Royal Statistical Society, Series B, 58(1996), 349–360.
Fawcett, T.: "An introduction to ROC analysis", Pattern Recognition Letters, 27(2006), 861–874.
Fisher, R.A.: "On the mathematical foundations of theoretical statistics", Philosophical Transactions of the Royal Society of London, Series A, 222(1922), 309–368.
Gallini, J.: "Misspecifications that can result in path analysis structures", Applied Psychological Measurement, 7(1983), 125–137.
Golden, R.M.: Mathematical methods for neural network analysis and design. Cambridge, Mass.: MIT Press, 1996.
Golden, R. M.: "Statistical tests for comparing possibly misspecified and nonnested models", Journal of Mathematical Psychology, 44(2000), 153–170.
Golden, R.M.: "Discrepancy risk model selection test theory for comparing possibly misspecified or nonnested models", Psychometrika, 68(2003), 229–249.
Greene, W.: Econometric Analysis. New Jersey: Prentice-Hall, 2003.
Hall, A.: "The Information Matrix Test for the Linear Model", The Review of Economic Studies, 54(1987), 257–263.
Hamilton, J. D.: Time Series Analysis. Princeton, New Jersey: Princeton University Press, 1994.
Hanley, J. A., and B. J. McNeil: "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve", Radiology, 143(1982), 29–36.
Harrell, F. E.: Regression modeling strategies : with applications to linear models, logistic regression, and survival analysis. New York: Springer, 2001.
Hastie, T., R. Tibshirani, and J. Friedman: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in, Statistics, 2009.
Hastie, T. J., and R. J. Tibshirani: "Generalized additive models", Statistical Science, 3(1986), 297–318.
Hastie, T. J., and R. J. Tibshirani: Generalized Additive Models. Chapman & Hall/CRC, 1990.
Henley, S. S., R. M. Golden, T. M. Kashner, and H. White: "Exploiting Hidden Structures in Epidemiological Data: Phase II Project", (R44AA011607) National Institute on Alcohol Abuse and Alcoholism, 2000. http://www.sbir.gov/sbirsearch/detail/223679
Henley, S. S., R. M. Golden, T. M. Kashner, H. White, and R. D. Katz: "Improving Validity Measures for Alcohol-Related Models: Phase I Project", (R43AA013351) National Institute on Alcohol Abuse and Alcoholism, 2001. http://www.sbir.gov/sbirsearch/detail/223681
Henley, S. S., R. M. Golden, T. M. Kashner, H. White, and R. D. Katz: "Robust Classification Methods for Categorical Regression: Phase I Project", (R43AA014302) National Institute on Alcohol Abuse and Alcoholism, 2003. http://www.sbir.gov/sbirsearch/detail/223689
Henley, S. S., R. M. Golden, T. M. Kashner, H. White, and D. Paik: "Robust Classification Methods for Categorical Regression: Phase II Project", (R44CA139607) National Cancer Institute, 2008. http://www.sbir.gov/sbirsearch/detail/223709
Henley, S. S., R. M. Golden, T. M. Kashner, H. White, L. Xuan, D. Paik, and R. D. Katz: "Improving Validity Measures in Alcohol-Related Models: Phase II Project", (R44AA013351) National Institute on Alcohol Abuse and Alcoholism, 2004. http://www.sbir.gov/sbirsearch/detail/223693

Hilbe, J. M.: Logistic Regression Models. New York: Chapman and Hall, 2009.
Horowitz, J.L.: "Bootstrap-based critical values for the information matrix test", Journal of Econometrics, 61(1994), 395–411.
Horowitz, J.L.: "The bootstrap in econometrics", Statistical Science, 18(2003), 211–218.
Hosmer, D. W., T. Hosmer, S. LeCessie, and S. Lemeshow: "A comparison of goodness-of-fit tests for the logistic regression model", Statistics in Medicine, 16(1997), 965–980.
Hosmer, D. W., and S. Lemeshow: "A goodness-of-fit test for the multiple logistic regression model", Communication in Statistics, A10(1980), 1043–1069.
Hosmer, D. W., and S. Lemeshow: Applied Logistic Regression. New York: John Wiley & Sons, 2000.
Hosmer, D. W., S. Lemeshow, and J. Klar: "Goodness-of-Fit Testing for Multiple Logistic Regression Analysis when the Estimated Probabilities are Small", Biometrical Journal, 30(1988), 1–14.
Hosmer, D. W., S. Taber, and S. Lemeshow: "The importance of assessing the fit of logistic regression models: a case study", American Journal of Public Health, 81(1991), 1630–1635.
Huber, P.: "The behavior of maximum likelihood estimates under non-standard conditions", University of California Press, 1967.
Kashner, T. M., T. J. Carmody, T. Suppes, A. J. Rush, M. L. Crismon, A. L. Miller, M. Toprac, and M. Trivedi: "Catching up on health outcomes: The Texas Medication Algorithm Project", Health Services Research, 38(2003), 311–331.
Kashner, T. M., S. S. Henley, R. M. Golden, J. M. Byrne, S. A. Keitz, G. W. Cannon, B. K. Chang, G. J. Holland, D. C. Aron, E. A. Muchmore, A. Wicker, and H. White: "Studying the Effects of ACGME Duty Hours Limits on Resident Satisfaction: Results From VA Learners' Perceptions Survey", Academic Medicine, 85(2010), 1130–1139.
Kashner, T. M., S. S. Henley, R. M. Golden, A. J. Rush, and R. B. Jarrett: "Assessing the preventive effects of cognitive therapy following relief of depression: A methodological innovation", Journal of Affective Disorders, 104(2007), 251–261.
Kashner, T. M., R. Rosenheck, A. B. Campinell, A. Suris, and C. W. T. S. Team: "Impact of work therapy on health status among homeless, substance-dependent veterans - A randomized controlled trial", Archives of General Psychiatry, 59(2002), 938–944.
Konishi, S., and G. Kitagawa: "Generalized information criteria in model selection", Biometrika, 83(1996), 875–890.
Kuss, O.: "Global goodness-of-fit tests in logistic regression with sparse data", Statistics in Medicine, 21(2002), 3789–3801.
Lancaster, T.: "The Covariance Matrix of the Information Matrix Test", Econometrica, 52(1984), 1051–1054.
Lehmann, E. L.: "Model specification: The views of Fisher and Neyman, and later developments", Statistical Science, 5(1990), 160–168.
Maddala, G. S.: Limited-dependent and Qualitative Variables in Econometrics. New York: Cambridge, 1999.
Magnus, J. R.: "On differentiating eigenvalues and eigenvectors", Econometric Theory, 1(1985), 179–191.
Magnus, J. R., and H. Neudecker: Matrix Differential Calculus with Applications in Statistics and Econometrics. New York: John Wiley & Sons, 1999.
McCullagh, P.: "On the asymptotic distribution of Pearson's statistic in linear exponential family models", International Statistical Review, 53(1985), 61–67.
McCullagh, P., and J. A. Nelder: Generalized linear models. New York: Chapman and Hall, 1989.
Orme, C.: "The Calculation of the Information Matrix Test for Binary Data Models", The Manchester School, 56(1988), 370–376.
Orme, C.: "The small-sample performance of the information-matrix test", Journal of Econometrics, 46(1990), 309–331.
Osius, G., and D. Rojek: "Normal goodness-of-fit tests for multinomial models with large degrees-of-freedom", Journal of the American Statistical Association, 87(1992), 1145–1152.

Pepe, M. S.: The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford: Oxford University Press, 2004.
Politis, D. N., J. P. Romano, and M. Wolf: Subsampling. New York: Springer, 1999.
Qin, J., and B. Zhang: "A goodness-of-fit test for logistic regression models based on case-control data", Biometrika, 84(1997), 609–618.
Raudenbush, S. W., and A. S. Bryk: Hierarchical Linear Models: Applications and Data Analysis Methods. Thousand Oaks, CA: Sage Publications, Inc., 2002.
Sarkar, S. K., and H. Midi: "Importance of assessing the model adequacy of binary logistic regression", Journal of Applied Sciences, 10(2010), 479–486.
Serfling, R. J.: Approximation theorems of mathematical statistics. New York: John Wiley & Sons, 1980.
Stomberg, C., and H. White: "Bootstrapping the Information Matrix Test", University of California, San Diego Department of Economics Discussion Paper, 2000.
Stukel, T.A.: "Generalized logistic models", Journal of the American Statistical Association, 83(1988), 426–431.
Takeuchi, K.: "Distribution of information statistics and a criterion of model fitting for adequacy of models", Mathematical Sciences, 153(1976), 12–18.
Taylor, L.W.: "The Size Bias of White's Information Matrix Test", Economics Letters, 24(1987), 63–67.
Tsay, R.S.: Analysis of Financial Time Series. New York: John Wiley & Sons, 2010.
Tsiatis, A.A.: "A Note on a goodness-of-fit test for the logistic regression model", Biometrika, 67(1980), 250–251.
Verbeke, G., and E. Lesaffre: "The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data", Computational Statistics and Data Analysis, 23(1997), 541–556.
Vuong, Q.H.: "Likelihood ratio tests for model selection and non-nested hypotheses", Econometrica, 57(1989).
Wald, A.: "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large", Transactions of the American Mathematical Society, 54(1943), 426–482.
Wei, B.: Exponential Family Nonlinear Models. New York: Springer, 1998.
White, H.: "Using least squares to approximate unknown regression functions", International Economic Review, 21(1980), 149–170.
White, H.: "Consequences and detection of misspecified nonlinear regression models", Journal of the American Statistical Association, 76(1981), 419–433.
White, H.: "Maximum Likelihood Estimation of Misspecified Models", Econometrica, 50(1982), 1–25.
White, H.: "Specification Testing in Dynamic Models", Cambridge University Press, 1987.
White, H.: Estimation, inference, and specification analysis. Cambridge: Cambridge University Press, 1994.
Wickens, T.D.: Elementary Signal Detection Theory. New York: Oxford University Press, 2002.
Winkler, G.: Image Analysis, Random Fields, and Dynamic Monte Carlo Methods. New York: Springer-Verlag, 1991.
Zhang, B.: "A chi-squared goodness-of-fit test for logistic regression models based on case-control data", Biometrika, 86(1999), 531–539.
Zhang, B.: "An information matrix test for logistic regression models based on case-control data", Biometrika, 88(2001), 921–932.