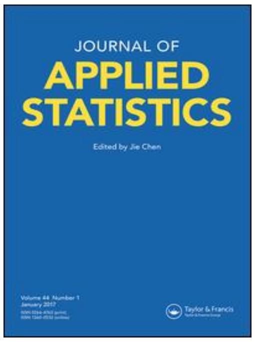

Taylor & Francis
Taylor & Francis Group

Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: https://www.tandfonline.com/loi/cjas20

Reconceptualizing the *p*-value from a likelihood ratio test: a probabilistic pairwise comparison of models based on Kullback-Leibler discrepancy measures

Benjamin Riedle, Andrew A. Neath & Joseph E. Cavanaugh

To cite this article: Benjamin Riedle, Andrew A. Neath & Joseph E. Cavanaugh (2020): Reconceptualizing the *p*-value from a likelihood ratio test: a probabilistic pairwise comparison of models based on Kullback-Leibler discrepancy measures, Journal of Applied Statistics, DOI: 10.1080/02664763.2020.1754360

To link to this article: https://doi.org/10.1080/02664763.2020.1754360

Published online: 23 Apr 2020.

JOURNAL OF APPLIED STATISTICS
https://doi.org/10.1080/02664763.2020.1754360

Reconceptualizing the *p*-value from a likelihood ratio test: a probabilistic pairwise comparison of models based on Kullback-Leibler discrepancy measures

Benjamin Riedle[a], Andrew A. Neath[b] and Joseph E. Cavanaugh[c]

[a]Eli Lilly and Company, Indianapolis, IN, USA; [b]Department of Mathematics and Statistics, Southern Illinois University, Edwardsville, IL, USA; [c]Department of Biostatistics, University of Iowa, Iowa City, IA, USA

ABSTRACT

Discrepancy measures are often employed in problems involving the selection and assessment of statistical models. A discrepancy gauges the separation between a fitted candidate model and the underlying generating model. In this work, we consider pairwise comparisons of fitted models based on a probabilistic evaluation of the ordering of the constituent discrepancies. An estimator of the probability is derived using the bootstrap. In the framework of hypothesis testing, nested models are often compared on the basis of the *p*-value. Specifically, the simpler null model is favored unless the *p*-value is sufficiently small, in which case the null model is rejected and the more general alternative model is retained. Using suitably defined discrepancy measures, we mathematically show that, in general settings, the likelihood ratio test *p*-value is approximated by the bootstrapped discrepancy comparison probability (BDCP). We argue that the connection between the *p*-value and the BDCP leads to potentially new insights regarding the utility and limitations of the *p*-value. The BDCP framework also facilitates discrepancy-based inferences in settings beyond the limited confines of nested model hypothesis testing.

ARTICLE HISTORY
Received 2 December 2019
Accepted 5 April 2020

KEYWORDS
Bootstrap; discrepancy functions; hypothesis testing; model evaluation; model selection

1. Introduction

A statistical model embodies a set of assumptions regarding how a set of data was generated. Ideally, a model will provide a good approximation to the true data-generating mechanism. Using the observed data, a hypothesis regarding some aspect of the model can be evaluated via a hypothesis test. In hypothesis testing, two hypotheses, the null and alternative, are proposed. The null typically corresponds to an assumption of no effect or no difference. The model corresponding to the alternative is assumed to be adequately specified.

In standard hypothesis testing, if the alternative model is not adequately specified, then the validity of the results of a hypothesis test, including the *p*-value, is brought into question. In many statistical modeling applications, the notion of *any* model being correct is difficult to defend, thus reminding one of the famous George Box [8] quote: ‘All models

CONTACT Joseph E. Cavanaugh ✉ joe-cavanaugh@uiowa.edu

are wrong; some are useful.' Unfortunately, hypothesis testing is often performed in settings where the alternative model is unlikely to provide an adequate characterization of the underlying phenomenon. Johnson [15] provides ecological examples in which hypothesis testing was performed when the alternative model is almost certainly underspecified.

In recent years, misunderstanding and misapplication of the *p*-value has led to skepticism in the general efficacy of hypothesis testing in scientific research [23]. This skepticism has led to some extreme decisions. For instance, in 2015, the editors of *Basic and Applied Social Psychology* decided to ban all *p*-values [32]. The mounting controversy led the American Statistical Association (ASA) to take the unprecedented step of issuing a policy statement on *p*-values, hoping to reduce confusion about their proper interpretation and use [33].

There is a growing sense that practitioners of statistical methods must rethink the standard *p*-value interpretation. For instance, Bland [6] and Boos and Stefanski [7] argue the need for a more refined interpretation of *p*-value distribution theory. Johnson [16] and McShane *et al.* [20] propose modifications to the interpretation of the *p*-value as an evidence measure in light of concerns over replication.

In this paper, we introduce a novel interpretation of the likelihood ratio (LR) test *p*-value. To develop this alternative interpretation, our work introduces the discrepancy comparison probability (DCP), which is a pairwise model comparison probability based on discrepancy measures. A discrepancy gauges the separation between a fitted candidate model and the underlying generating model. Discrepancies can be used to delineate between models, with the notion that a smaller discrepancy signifies a model that more closely adheres to the truth. When using discrepancies to select an appropriate model, it is typically unnecessary to assume one of the candidate models is the truth. Rather, most model selection techniques employing discrepancies seek to determine which model most closely corresponds to the truth, without requiring any candidate model to be precisely true. However, because a discrepancy depends on the unknown data generating process, calculating the discrepancy is impossible. Instead, under suitable conditions, model selection criteria may be derived as asymptotically unbiased estimators of expected discrepancies. Rather than relying on asymptotic theory to estimate discrepancies, this paper utilizes bootstrap resampling techniques to characterize the distribution of the discrepancy.

The DCP has broader utility and requires fewer assumptions than hypothesis testing and the *p*-value. Despite the broader applicability of the DCP, we attempt to establish a connection between the *p*-value and the DCP when hypothesis testing assumptions are met. Because the constituent discrepancies of the DCP cannot be calculated, an estimator of the DCP is derived using the bootstrap. We show that in certain large-sample settings, the LR *p*-value and the bootstrapped discrepancy comparison probability (BDCP) are approximately equal. To understand the importance of this connection, consider the standard interpretation of the *p*-value: 'the probability that if the null hypothesis is true, a test statistic will have a value as extreme or more extreme than the value we actually observe' [14]. Thus, the *p*-value is a probability conditioned on the null being true. On the other hand, the validity of discrepancy functions, and thus the BDCP, does not depend on whether the model is true. Therefore, by drawing a connection between the BDCP and the LR *p*-value, we show that rather than assuming the null is true, as the standard *p*-value interpretation requires, the LR *p*-value can at times simply be interpreted as a bootstrap-based probability that the null model is better than the alternative.

This paper establishes a connection between the BDCP and the LR p-value. This relationship between the BDCP and the p-value, however, can be demonstrated for most asymptotic tests based on a suitable discrepancy formulation. In work not shown, the connection has been established for the Wald and score tests (A detailed development can be found in [26]). Importantly, once the connection between the BDCP and the p-value is justified, a practitioner can use the connection to interpret the p-value in an alternative manner, without having to actually evaluate the BDCP. We discuss the benefits that can be gleaned from the connection between the BDCP and the p-value, including alternative interpretations of the p-value, which lead to potentially new insights regarding its behavior. While the BDCP can be connected to the p-value, the BDCP can also be employed in settings that will not lead to an approximation of the p-value. The utility of the DCP framework in settings that do not necessarily lead to a natural connection to the p-value will also be considered.

In Section 2, we formally introduce discrepancy functions and how to estimate them using the bootstrap. We then introduce the discrepancy comparison probability (DCP) as a pairwise comparison of the discrepancies of two models. In Section 3, we mathematically show that the LR p-value can be approximated by the BDCP using a suitably chosen discrepancy. The mathematical results of Section 3 are supported with a simulation study in Section 4. In Section 5, we examine the benefits that can be gleaned from connecting the p-value to a discrepancy-based comparison of models, and then explain that the DCP framework can be applied in a broader range of settings than those in which standard hypothesis testing is valid. In Section 6, we apply our methodology to a biomedical application. We end with some concluding remarks in Section 7.

2. Background

In this section, we provide an introduction to discrepancy functions and discuss how the bootstrap can be applied to estimate the distribution of the overall discrepancy.

2.1. Discrepancy functions

Model selection problems often employ discrepancy functions to aid in the choice between competing models. Suppose we have a vector of independent observations $\boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, which is generated from an unknown, true distribution $g(\boldsymbol{y})$ that is not necessarily parametric. Further, suppose a parametric candidate model $f(\boldsymbol{y} \mid \boldsymbol{\theta})$ is put forth to approximate the observed data $\boldsymbol{y}$. Specifically, assume the candidate model belongs to a parametric class of densities

$$\mathscr{F} = \left\{ f(\boldsymbol{y} \mid \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta} \right\},$$

where $\boldsymbol{\Theta}$ is the parameter space for $\boldsymbol{\theta}$. A discrepancy function $d(g,f)$ provides a measure of the disparity between the true density $g(\boldsymbol{y})$ and a parametric model $f(\boldsymbol{y} \mid \boldsymbol{\theta})$ that satisfies

$$d(g,f) \geq d(g,g).$$

A discrepancy function need not be a formal distance metric. However, a discrepancy should still behave in a manner similar to a distance. Namely, as the dissimilarity between

$g(\boldsymbol{y})$ and $f(\boldsymbol{y} \mid \boldsymbol{\theta})$ increases, the discrepancy $d(g,f)$ should increase accordingly. For notational simplicity, we assume candidate parametric models can be characterized by their parameter vector $\boldsymbol{\theta}$, and will thus denote $d(g,f)$ by $d(g,\boldsymbol{\theta})$.

Let $\ell(\boldsymbol{\theta} \mid \boldsymbol{y}) = \log f(\boldsymbol{y} \mid \boldsymbol{\theta})$ be the natural logarithm of the likelihood function for the candidate model. Accordingly, let $\ell_i(\boldsymbol{\theta} \mid y_i)$ represent the contribution of the ith observation to the log-likelihood. The Kullback-Leibler (KL) discrepancy between the true model $g(\boldsymbol{y})$ and candidate model $f(\boldsymbol{y} \mid \boldsymbol{\theta})$ is defined as

$$d_{KL}(g,\boldsymbol{\theta}) = E_g\left\{-2\ell(\boldsymbol{\theta} \mid \boldsymbol{z})\right\},$$

where E_g denotes expectation with respect to the true distribution g, and $\boldsymbol{z} = (z_1, \ldots, z_n)^{\mathrm{T}}$ is a sample of independent observations drawn from the true distribution g, generated independently of $\boldsymbol{y}$. The KL discrepancy assesses how well the candidate model predicts future data arising from the true distribution. In the subsequent development, $\boldsymbol{y}$ will serve as a fitting sample and $\boldsymbol{z}$ as a validation sample.

For the purpose at hand, the KL discrepancy may be viewed as an operationally equivalent variant of the ubiquitous Kullback-Leibler information [17]. The KL information between the true model $g(\boldsymbol{y})$ and candidate model $f(\boldsymbol{y} \mid \boldsymbol{\theta})$ may be defined as

$$I_{KL}(g,\boldsymbol{\theta}) = E_g\left\{\log \frac{g(\boldsymbol{z})}{f(\boldsymbol{z} \mid \boldsymbol{\theta})}\right\}.$$

As a consequence of Jensen's inequality, $I_{KL}(g,\boldsymbol{\theta}) \geq 0$ with equality if and only if $g(\boldsymbol{y})$ and $f(\boldsymbol{y} \mid \boldsymbol{\theta})$ are the same density. Although $I_{KL}(g,\boldsymbol{\theta})$ is not a formal metric, the KL information reflects the separation between the true model and the candidate model. Note that we can write

$$2I_{KL}(g,\boldsymbol{\theta}) = d_{KL}(g,\boldsymbol{\theta}) - E_g\left\{-2\log g(\boldsymbol{z})\right\}.$$

Since $E_g\{-2\log g(\boldsymbol{z})\}$ is a constant that does not depend on the structure of the candidate model, for the purpose of discriminating among various candidate models, the KL discrepancy $d_{KL}(g,\boldsymbol{\theta})$ serves as a valid substitute for the KL information $I_{KL}(g,\boldsymbol{\theta})$.

Quantifying how well the *fitted* candidate model approximates the true distribution is often of interest in a model selection problem. Let $\hat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \ell(\boldsymbol{\theta} \mid \boldsymbol{y})$ denote the maximum likelihood estimator of $\boldsymbol{\theta}$. Similarly, let $f(\boldsymbol{y} \mid \hat{\boldsymbol{\theta}})$ denote the corresponding fitted model, and let $\ell(\hat{\boldsymbol{\theta}} \mid \boldsymbol{y}) = \log f(\boldsymbol{y} \mid \hat{\boldsymbol{\theta}})$. The discrepancy between the true model g and the fitted candidate model $f(\boldsymbol{y} \mid \hat{\boldsymbol{\theta}})$ is referred to as the *overall* discrepancy, and is denoted $d(g,\hat{\boldsymbol{\theta}})$. The overall KL discrepancy for the fitted candidate model $f(\boldsymbol{y} \mid \hat{\boldsymbol{\theta}})$ is thereby

$$d_{KL}(g,\hat{\boldsymbol{\theta}}) = E_g\left\{-2\ell(\boldsymbol{\theta} \mid \boldsymbol{z})\right\}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

Note that the overall discrepancy $d(g,\hat{\boldsymbol{\theta}})$ is a random variable, as it is a function of the estimated parameter vector $\hat{\boldsymbol{\theta}}$, and thus depends on $\boldsymbol{y}$. Therefore, it is useful to think of the *distribution* of $d(g,\hat{\boldsymbol{\theta}})$. Model selection criteria have been developed that seek to estimate the distribution of $d(g,\hat{\boldsymbol{\theta}})$, or some characteristic of its distribution. For instance, under certain regularity conditions, the Akaike information criterion (AIC; [1,2]) serves as an asymptotically unbiased estimator of the *expected value* of the overall KL discrepancy. For an overview of model selection criteria that focus on the expected value of overall discrepancies, see McQuarrie and Tsai [19] and Burnham and Anderson [9].

2.2. Using the bootstrap to estimate discrepancy functions

Instead of relying on asymptotics to estimate the expected value of the overall discrepancy $d(g, \hat{\boldsymbol{\theta}})$, this paper uses bootstrap resampling techniques to approximate the distribution of $d(g, \hat{\boldsymbol{\theta}})$. Efron [10,11] was the first to develop the idea of using the bootstrap in the model selection context. Our method employs the non-parametric bootstrap, since this version of the bootstrap is not impacted by model misspecification.

The bootstrap samples are of size n, drawn with replacement from $\boldsymbol{y}$. Note that for most applications, for $i = 1, \ldots, n$, each observation y_i will have corresponding covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iJ})^{\mathrm{T}}$, where for $j = 1, \ldots, J$, x_{ij} denotes the ith observation on the jth covariate. For each observation y_i included in a bootstrap sample, its corresponding covariate vector $\boldsymbol{x}_i$ is also included, and thus the selection of y_i implies the selection of $(y_i, \boldsymbol{x}_i^{\mathrm{T}})^{\mathrm{T}}$. Following a common convention of modeling notation, we will often let the covariates $\boldsymbol{x}_i$ and the outcome y_i be represented simply by y_i.

We can use the bootstrap to estimate the distribution of the overall discrepancy $d(g, \hat{\boldsymbol{\theta}})$, by applying what Efron and Tibshirani [12] refer to as the 'plug-in principle.' The plug-in principle dictates that each element of the overall discrepancy is replaced by its bootstrap analogue. For instance, applying the plug-in principle to the overall KL discrepancy can be summarized by the following replacements:

$$g \to \hat{g}, \quad \boldsymbol{y} \to \boldsymbol{y}^*, \quad E_g \to E_{\hat{g}}, \quad \hat{\boldsymbol{\theta}} \to \hat{\boldsymbol{\theta}}^*.$$

Here, $\hat{g}$ is the empirical distribution; $\boldsymbol{y}^*$ is a bootstrap sample drawn from $\hat{g}$, and $\hat{\boldsymbol{\theta}}^*$ is the MLE of $\boldsymbol{\theta}$ derived under the bootstrap sample $\boldsymbol{y}^*$. Because the observations $y_1, \ldots, y_n$ are independent, the bootstrap analogue to the overall KL discrepancy is then given by

$$\begin{aligned} d_{KL}\left(\hat{g}, \hat{\boldsymbol{\theta}}^*\right) &= E_{\hat{g}}\left\{-2\ell(\boldsymbol{\theta} \mid \boldsymbol{z})\right\}\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^*} \\ &= \sum_{i=1}^{n}\left\{-2\ell_i(\hat{\boldsymbol{\theta}}^* \mid y_i)\right\} \\ &= -2\ell(\hat{\boldsymbol{\theta}}^* \mid \boldsymbol{y}). \end{aligned}$$

Rather than just thinking in terms of the overall KL discrepancy, consider a generic discrepancy d. To derive a bootstrap-based estimator of the distribution of the overall discrepancy, we first draw $b = 1, \ldots, B$ bootstrap samples from $\boldsymbol{y}$. For $b = 1, \ldots, B$, let the MLE of $\boldsymbol{\theta}$ based on the b^{th} bootstrap sample be denoted $\hat{\boldsymbol{\theta}}^*(b)$. Then, for $b = 1, \ldots, B$, calculate the bootstrap analogue to the overall discrepancy $d(\hat{g}, \hat{\boldsymbol{\theta}}^*(b))$. The set

$$\left\{d\left(\hat{g}, \hat{\boldsymbol{\theta}}^*(b)\right) : b = 1, \ldots, B\right\}$$

serves as a bootstrap-based approximation to the distribution of the overall discrepancy.

2.3. The discrepancy comparison probability (DCP)

Suppose that two nested models are put forth to approximate observed data $\boldsymbol{y}$. Delineating between these two models is often done using hypothesis testing, where we choose in favor of the null model unless the p-value is sufficiently small, in which case we reject the null and decide in favor of the alternative.

However, deciding between these two competing models could also be done using discrepancy functions. Let the MLE of $\boldsymbol{\theta}$ for the null and alternative models be denoted $\hat{\boldsymbol{\theta}}_0$ and $\hat{\boldsymbol{\theta}}$, respectively, with corresponding overall discrepancies $d(g, \hat{\boldsymbol{\theta}}_0)$ and $d(g, \hat{\boldsymbol{\theta}})$. There exist many ways in which a model can be evaluated in the discrepancy function framework. For instance, suppose one is interested in which of the two models has a smaller expected overall KL discrepancy. Then, choosing the model with the smaller AIC would be an appropriate way of proceeding. However, in this paper, we do not seek to delineate between two models using their expected overall discrepancies, but instead evaluate the models using the probability

$$P = \Pr\left[d(g, \hat{\boldsymbol{\theta}}_0) < d(g, \hat{\boldsymbol{\theta}})\right],$$

which we refer to as the *discrepancy comparison probability* (DCP). The DCP is the probability that the fitted null model will be more congruous with the true model than the fitted alternative, as measured by the discrepancy function d. To help better understand the DCP, suppose $P = 0.80$. Then, the fitted null model will have a smaller overall discrepancy than the fitted alternative in 80% of samples of size n drawn from the generating distribution. The null model may be *better* without conforming precisely to the truth; if the bias of the null model is negligible compared to the additional estimation error of the alternative model, then the null will be preferred. Importantly, the DCP is a pairwise comparison of the two competing models because it compares the null and alternative overall discrepancies *derived under the same samples*. Therefore, the DCP is actually a measure on the joint distribution of $d(g, \hat{\boldsymbol{\theta}}_0)$ and $d(g, \hat{\boldsymbol{\theta}})$.

Of course, because the true distribution g is unknown, we cannot calculate either $d(g, \hat{\boldsymbol{\theta}}_0)$ or $d(g, \hat{\boldsymbol{\theta}})$. Instead, as outlined in the previous subsection, we employ bootstrap resampling to approximate their joint distribution. For the null and alternative models, let the MLE of $\boldsymbol{\theta}$ derived using the bootstrap sample be denoted as $\hat{\boldsymbol{\theta}}_0^*$ and $\hat{\boldsymbol{\theta}}^*$, respectively. Also, let the null and alternative model bootstrap-based estimator of the overall discrepancy be denoted $d(\hat{g}, \hat{\boldsymbol{\theta}}_0^*)$ and $d(\hat{g}, \hat{\boldsymbol{\theta}}^*)$, respectively. Finally, for $b = 1, \dots, B$, let the null and alternative model MLE of $\boldsymbol{\theta}$ based on the b^{th} bootstrap sample be denoted $\hat{\boldsymbol{\theta}}_0^*(b)$ and $\hat{\boldsymbol{\theta}}^*(b)$, respectively. Then, for $b = 1, \dots, B$, we apply the plug-in principle to derive the following empirical approximation of the joint distribution of $d(g, \hat{\boldsymbol{\theta}}_0)$ and $d(g, \hat{\boldsymbol{\theta}})$:

$$\left\{\left(d\left(\hat{g}, \hat{\boldsymbol{\theta}}_0^*(b)\right), d\left(\hat{g}, \hat{\boldsymbol{\theta}}^*(b)\right)\right) : b = 1, \dots, B\right\}.$$

Because the DCP P is of particular interest in this paper, we use the bootstrap to derive an estimator. Let $\Pr^*$ denote probability with respect to the joint bootstrap distribution of $d(\hat{g}, \hat{\boldsymbol{\theta}}_0^*)$ and $d(\hat{g}, \hat{\boldsymbol{\theta}}^*)$. Following the plug-in principle, the *bootstrap-based discrepancy comparison probability* P^*, or the BDCP, is then

$$P^* = \Pr^*\left[d(\hat{g}, \hat{\boldsymbol{\theta}}_0^*) < d(\hat{g}, \hat{\boldsymbol{\theta}}^*)\right].$$

The BDCP P^* is the probability the bootstrap-based estimator of the overall discrepancy is smaller under the null than under the alternative. Let $\mathbb{1}(\cdot)$ denote the indicator function.

We can approximate P^* by drawing $b = 1, \ldots, B$ bootstrap samples, and calculating

$$\hat{P}^* = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1} \left\{ d\left(\hat{g}, \hat{\boldsymbol{\theta}}_0^*(b)\right) < d\left(\hat{g}, \hat{\boldsymbol{\theta}}^*(b)\right) \right\},$$

which is simply the proportion of the B bootstrap samples in which the bootstrap-based overall discrepancy estimator is smaller under the null model than under the alternative. Conceptually, the BDCP mimics the DCP in that the DCP is a probability based on repeated samples drawn from the generating distribution, whereas the BDCP is a probability based on drawing repeated bootstrap samples from the sample $\boldsymbol{y}$. The BDCP is then the proportion of the samples in which the fitted null model is more congruous with the 'truth' $\boldsymbol{y}$ than the fitted alternative.

3. Connection with LR *p*-value

Hypothesis testing, and by extension the p-value, is criticized for the illogical premise of testing the correctness of a hypothesis in settings in which no hypothesis is likely to be exactly correct. The discrepancy function approach to model selection does not suffer from this same criticism. Despite the differences in these approaches to model selection, in large-sample settings with nested models, under an adequately specified alternative model, we will theoretically establish that the LR p-value is approximately equal to the BDCP using specifically chosen discrepancies. Employing the BDCP under the KL discrepancy, this result is first established in the 'full null' setting in which the null hypothesis pre-specifies all parameter values. Using a variant of the KL discrepancy, we then establish the result in the 'partial null' setting, where only a portion of the parameter values are pre-specified by the null hypothesis. We derive these results under the assumptions of the LR test.

3.1. The full null setting

In the full null setting the null hypothesis pre-specifies all parameter values. Thus, the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ is tested against the general alternative $H_A : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. If we let $dim(\boldsymbol{\theta}) = p_A$, with the LR test statistic denoted by

$$L = 2\left(\ell(\hat{\boldsymbol{\theta}} \mid \boldsymbol{y}) - \ell(\boldsymbol{\theta}_0 \mid \boldsymbol{y})\right), \tag{1}$$

then the LR test p-value for these hypotheses is

$$p_{LR} = \Pr[\chi^2_{p_A} > L], \tag{2}$$

where $\chi^2_{p_A}$ is a central chi-square random variable with p_A degrees of freedom [34].

For the asymptotic χ^2 distribution of the LR test statistic to hold, both the true parameter vector and the parameter vector under the null hypothesis must lie on the interior of the parameter space. If some element(s) of the null hypothesis parameter vector lie on the boundary of the parameter space, then the null distribution of LR test statistic is more complicated and may be difficult to characterize. In such settings, the null distribution is generally a mixture of a degenerate distribution at zero and a χ^2 distribution; see, for instance, Self and Liang [28]. Additional relevant references include [3,13,31].

The overall KL discrepancy of the model corresponding to the alternative hypothesis is

$$d_{KL}(g, \hat{\boldsymbol{\theta}}) = E_g \{-2\ell(\boldsymbol{\theta} \mid \boldsymbol{z})\} \mid_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}},$$

whose bootstrap-based estimator is

$$d_{KL}(\hat{g}, \hat{\boldsymbol{\theta}}^*) = -2\ell(\hat{\boldsymbol{\theta}}^* \mid \boldsymbol{y}). \tag{3}$$

Similarly, the null overall KL discrepancy is

$$d_{KL}(g, \boldsymbol{\theta}_0) = E_g \{-2\ell(\boldsymbol{\theta}_0 \mid \boldsymbol{z})\}.$$

Unlike the alternative model whose parameter vector is maximized over its parameter space, the null parameter vector is fixed at $\boldsymbol{\theta}_0$ in the full null setting. Therefore, the bootstrap-based 'estimator' of $\boldsymbol{\theta}$ is $\boldsymbol{\theta}_0$ for all bootstrap samples. The bootstrap-based estimator of the null overall KL discrepancy is then

$$d_{KL}(\hat{g}, \boldsymbol{\theta}_0) = -2\ell(\boldsymbol{\theta}_0 \mid \boldsymbol{y}). \tag{4}$$

Note that conditioned upon the observed data $\boldsymbol{y}$, the bootstrap-based estimator of the null overall KL discrepancy $d_{KL}(\hat{g}, \boldsymbol{\theta}_0)$ is actually fixed, and thus does not vary from one bootstrap sample to the next. Applying the null and alternative bootstrap-based KL discrepancy estimators in (3) and (4), respectively, yields the following BDCP:

$$P^*_{KL} = \Pr{}^* \left[-2\ell(\boldsymbol{\theta}_0 \mid \boldsymbol{y}) < -2\ell(\hat{\boldsymbol{\theta}}^* \mid \boldsymbol{y}) \right]. \tag{5}$$

To present the proof connecting the likelihood ratio test p-value with the BDCP, we introduce the following distributional constructs. Let the observed Fisher information matrix be denoted by

$$I(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{-\partial^2 \ell(\boldsymbol{\theta} \mid \boldsymbol{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} = -\sum_{i=1}^{n} \left(\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} \ell_i(\boldsymbol{\theta} \mid y_i) \right),$$

and let the expected Fisher information be denoted

$$\mathscr{I}(\boldsymbol{\theta}) = E\left[I(\boldsymbol{\theta} \mid \boldsymbol{y}) \right].$$

Proposition 3.1: *Assuming that the large-sample properties of the MLEs hold, that the alternative model is adequately specified, and that the true parameter vector and the parameter vector under the null hypothesis lie in the interior of the parameter space, then for testing a full null hypothesis of* $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ *versus the alternative of* $H_A : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$,

$$p_{LR} \approx P^*_{KL}.$$

Proof: We begin by stating a well-known result from maximum likelihood estimation that will be applied later in the proof. Recall that for large n, under certain regularity conditions with the alternative model being adequately specified,

$$\hat{\boldsymbol{\theta}} \,\dot{\sim}\, N_{p_A}\left(\boldsymbol{\theta}, \mathscr{I}^{-1}(\boldsymbol{\theta})\right).$$

It follows that

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^{\mathrm{T}} \mathscr{I}(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \,\dot{\sim}\, \chi^2_{p_A}. \tag{6}$$

Recall from (5) that the equation for the BDCP is

$$P^*_{KL} = \Pr{}^*\left[-2\ell(\boldsymbol{\theta}_0 \,|\, \boldsymbol{y}) < -2\ell(\hat{\boldsymbol{\theta}}^* \,|\, \boldsymbol{y})\right].$$

Because Pr* denotes probability with respect to the joint distribution of $d_{KL}(\hat{g}, \hat{\boldsymbol{\theta}}^*)$ and $d_{KL}(\hat{g}, \boldsymbol{\theta}_0)$, we can conceptualize the observed data $\boldsymbol{y}$ as being fixed under Pr*. The bootstrap sample, and thus $\hat{\boldsymbol{\theta}}^*$, are random under Pr*. We rearrange P^*_{KL} so as to introduce the likelihood ratio statistic L from (1):

$$\begin{aligned} P^*_{KL} &= \Pr{}^*\left[-2\ell(\boldsymbol{\theta}_0 \,|\, \boldsymbol{y}) < -2\ell\left(\hat{\boldsymbol{\theta}}^* \,|\, \boldsymbol{y}\right)\right] \\ &= \Pr{}^*\left[2\left(\ell(\hat{\boldsymbol{\theta}} \,|\, \boldsymbol{y}) - \ell(\boldsymbol{\theta}_0 \,|\, \boldsymbol{y})\right) < 2\left(\ell(\hat{\boldsymbol{\theta}} \,|\, \boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}^* \,|\, \boldsymbol{y})\right)\right] \\ &= \Pr{}^*\left[2\left(\ell(\hat{\boldsymbol{\theta}} \,|\, \boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}^* \,|\, \boldsymbol{y})\right) > L\right]. \end{aligned} \tag{7}$$

Under fixed observed data $\boldsymbol{y}$, the likelihood ratio statistic L is fixed. In order to show that the LR p-value in (2) is approximated by P^*_{KL} in (7), we need to show that under Pr*,

$$2\left(\ell(\hat{\boldsymbol{\theta}} \,|\, \boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}^* \,|\, \boldsymbol{y})\right) \,\dot{\sim}\, \chi^2_{p_A}.$$

Consider taking a second-order Taylor series expansion of $\ell(\hat{\boldsymbol{\theta}}^* \,|\, \boldsymbol{y})$ about $\hat{\boldsymbol{\theta}}$, which yields

$$\ell(\hat{\boldsymbol{\theta}}^* \,|\, \boldsymbol{y}) \approx \ell(\hat{\boldsymbol{\theta}} \,|\, \boldsymbol{y}) - \tfrac{1}{2}(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}})^{\mathrm{T}} I(\hat{\boldsymbol{\theta}} \,|\, \boldsymbol{y})(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}).$$

For large n, the observed information is approximated by the expected information, and thus we can write

$$2\left(\ell(\hat{\boldsymbol{\theta}} \,|\, \boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}^* \,|\, \boldsymbol{y})\right) \approx (\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}})^{\mathrm{T}} \mathscr{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}). \tag{8}$$

Assuming the data at hand adequately characterizes the sampling distribution of $\hat{\boldsymbol{\theta}}$ via the bootstrap distribution of $\hat{\boldsymbol{\theta}}^*$, then applying the large-sample result (6) to the bootstrapping context yields

$$(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}})^{\mathrm{T}} \mathscr{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}) \dot{\sim} \chi^2_{p_A}. \tag{9}$$

Applying (8) and (9), we have

$$2\left(\ell(\hat{\boldsymbol{\theta}} \mid \boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}^* \mid \boldsymbol{y})\right) \dot{\sim} \chi^2_{p_A}.$$

Referring back to (7), we can establish our desired result:

$$\begin{aligned} P^*_{KL} &= \Pr{}^*\left[2\left(\ell(\hat{\boldsymbol{\theta}} \mid \boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}^* \mid \boldsymbol{y})\right) > L\right] \\ &\approx \Pr\left[\chi^2_{p_A} > L\right] \\ &= p_{LR}. \end{aligned} \tag{10}$$

We have thus shown that, assuming an adequately specified alternative model, $P^*_{KL} \approx p_{LR}$ when the null hypothesis pre-specifies all parameter values. This completes the proof. ■

Result (10) establishes the LR p-value as a bootstrap-based estimator of the probability that the null is 'better' than the alternative, as measured by the bootstrapped overall KL discrepancy. Clearly, the p-value is not the probability that the null is *true*, but it is not a requirement for the null model to be true for it to better than the alternative. If $\boldsymbol{\theta}_0$ provides a reasonable characterization of $\boldsymbol{\theta}$, then due to the sampling variability incurred under the alternative model, the null model may be more accurate than the estimated alternative model.

Because the null hypothesis pre-specifies all parameter values in the full null setting, the bootstrap-based parameter vector estimator for the null model is fixed under Pr*, leading to the null bootstrap-based KL discrepancy estimator $d(\hat{g}, \boldsymbol{\theta}_0)$ also being fixed. In the following subsection, we address the partial null setting, in which the null hypothesis does not pre-specify all parameter values.

3.2. The partial null setting

In the partial null setting, we again wish to show that the LR p-value is approximated by the BDCP. Suppose the null hypothesis pre-specifies k of the p_A parameter values. Let the parameter vector $\boldsymbol{\theta}$ be partitioned into the vector of parameters that the null pre-specifies, denoted by $\boldsymbol{\theta}_{(1)}$, and the parameters that are not pre-specified, denoted $\boldsymbol{\theta}_{(2)}$. We refer to $\boldsymbol{\theta}_{(1)}$ as the parameters of interest and to $\boldsymbol{\theta}_{(2)}$ as the nuisance parameters. The null hypothesis of the form $H_0 : \boldsymbol{\theta}_{(1)} = \boldsymbol{\theta}_{0(1)}$ is tested against the alternative $H_A : \boldsymbol{\theta}_{(1)} \neq \boldsymbol{\theta}_{0(1)}$. Let $\hat{\boldsymbol{\theta}}_{0(2)} = \operatorname{argmax}_{\boldsymbol{\theta}_{(2)}} \ell(\boldsymbol{\theta}_{0(1)}, \boldsymbol{\theta}_{(2)} \mid \boldsymbol{y})$ be the MLEs of the nuisance parameters derived under the null hypothesis. Then, let the MLEs derived under the null hypothesis and the sample $\boldsymbol{y}$ be denoted $\hat{\boldsymbol{\theta}}_0 = (\boldsymbol{\theta}^{\mathrm{T}}_{0(1)}, \hat{\boldsymbol{\theta}}^{\mathrm{T}}_{0(2)})^{\mathrm{T}}$. Let the unrestricted MLEs of $\boldsymbol{\theta}$ derived from the sample $\boldsymbol{y}$ also be partitioned so that $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}^{\mathrm{T}}_{(1)}, \hat{\boldsymbol{\theta}}^{\mathrm{T}}_{(2)})^{\mathrm{T}}$. Similarly, let the unrestricted MLEs of $\boldsymbol{\theta}$ derived from the bootstrap sample $\boldsymbol{y}^*$ also be partitioned so that $\hat{\boldsymbol{\theta}}^* = (\hat{\boldsymbol{\theta}}^{*T}_{(1)}, \hat{\boldsymbol{\theta}}^{*T}_{(2)})^{\mathrm{T}}$.

Letting the LR test statistic be denoted

$$L = 2\left(\ell(\hat{\boldsymbol{\theta}} \mid \boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}_0 \mid \boldsymbol{y})\right),$$

the LR test p-value in the partial null setting is

$$p_{LR} = \Pr\left[\chi_k^2 > L\right].$$

As in the full null setting, we need the parameter vectors to lie on the interior of the parameter space for the preceding result to hold.

Unlike in the full null setting, we will be unable to connect the LR p-value to the BDCP under the conventional KL discrepancy. That the BDCP under the conventional KL discrepancy does not approximate the p-value in this setting does not preclude one from using the KL discrepancy; the BDCP under the KL discrepancy is still a valid tool for deciding between two competing models, it simply does not have the connection with the LR p-value which we seek.

To draw a connection between the p-value and the BDCP, we instead use a modified version of the KL discrepancy, which we refer to as the *parameter of interest Kullback-Leibler* (PIKL) *discrepancy*. The PIKL discrepancy constitutes only a small modification of the conventional KL discrepancy and is a sensible tool for evaluating models that contain nuisance parameters. Of particular importance to this paper, the BDCP under the PIKL discrepancy approximates the LR p-value in the partial null setting, as we will soon establish. To understand the PIKL discrepancy, we first introduce the notion of the pseudo-true parameter $\bar{\boldsymbol{\theta}}$. The pseudo-true parameter is defined as the parameter value that minimizes the KL discrepancy. Write

$$\bar{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta}} d_{KL}(g, \boldsymbol{\theta}).$$

If the model is adequately specified (i.e. $g(\boldsymbol{y}) \in \mathscr{F}$), then $\bar{\boldsymbol{\theta}}$ is the true value of $\boldsymbol{\theta}$. However, $\bar{\boldsymbol{\theta}}$ is well-defined, regardless of whether the model is adequately specified. Let the pseudo-true nuisance parameter vector for the null model be denoted $\bar{\boldsymbol{\theta}}_{0(2)} = \operatorname{argmin}_{\boldsymbol{\theta}_{(2)}} d_{KL}(g, (\boldsymbol{\theta}_{0(1)}, \boldsymbol{\theta}_{(2)}))$. For the alternative model, let the pseudo-true parameter vector be denoted by

$$\left(\bar{\boldsymbol{\theta}}_{(1)}^{\mathrm{T}}, \bar{\boldsymbol{\theta}}_{(2)}^{\mathrm{T}}\right)^{\mathrm{T}} = \operatorname{argmin}_{(\boldsymbol{\theta}_{(1)}, \boldsymbol{\theta}_{(2)})} d_{KL}\left(g, \left(\boldsymbol{\theta}_{(1)}, \boldsymbol{\theta}_{(2)}\right)\right).$$

Also, let the MLEs of the parameters of interest, derived under the restriction that $\boldsymbol{\theta}_{(2)} = \bar{\boldsymbol{\theta}}_{(2)}$, be denoted $\hat{\boldsymbol{\theta}}_{C(1)} = \operatorname{argmax}_{\boldsymbol{\theta}_{(1)}} \ell((\boldsymbol{\theta}_{(1)}, \bar{\boldsymbol{\theta}}_{(2)}) \mid \boldsymbol{y})$. We will refer to $\hat{\boldsymbol{\theta}}_{C(1)}$ as the conditional MLEs, to emphasize their dependence on the condition $\boldsymbol{\theta}_{(2)} = \bar{\boldsymbol{\theta}}_{(2)}$.

Under the alternative model, the overall PIKL discrepancy evaluates the KL discrepancy at the pseudo-true values of the nuisance parameters and at the conditional MLEs of the parameters of interest. Specifically, for the alternative model, the measure is given by

$$d_{PIKL}\left(g, (\hat{\boldsymbol{\theta}}_{C(1)}, \bar{\boldsymbol{\theta}}_{(2)})\right) = E_g\left\{-2\ell(\boldsymbol{\theta}_{(1)}, \bar{\boldsymbol{\theta}}_{(2)} \mid \boldsymbol{z})\right\} |_{\boldsymbol{\theta}_{(1)} = \hat{\boldsymbol{\theta}}_{C(1)}},$$

whereas under the null model, the measure is

$$d_{PIKL}\left(g, (\boldsymbol{\theta}_{0(1)}, \bar{\boldsymbol{\theta}}_{0(2)})\right) = E_g\left\{-2\ell(\boldsymbol{\theta}_{0(1)}, \bar{\boldsymbol{\theta}}_{0(2)} \mid \boldsymbol{z})\right\}.$$

Note that the null model overall PIKL discrepancy $d_{PIKL}(g, (\boldsymbol{\theta}_{0(1)}, \bar{\boldsymbol{\theta}}_{0(2)}))$ is fixed, as is $d_{KL}(g, \boldsymbol{\theta}_0)$ in the full null setting.

To use the bootstrap to estimate the null and alternative discrepancies, we must apply the plug-in principle to the pseudo-true parameter vector. The pseudo-true parameter vector $\bar{\boldsymbol{\theta}}$ minimizes the KL discrepancy under the true distribution g, so the bootstrap-based version of $\bar{\boldsymbol{\theta}}$ should minimize the empirical KL discrepancy $-2\ell(\boldsymbol{\theta}\,|\,\boldsymbol{y})$. Therefore, we use $\hat{\boldsymbol{\theta}}$ as the plug-in for $\bar{\boldsymbol{\theta}}$. Accordingly, the bootstrap-based pseudo-true nuisance parameters for the null and alternative models are $\hat{\boldsymbol{\theta}}_{0(2)}$ and $\hat{\boldsymbol{\theta}}_{(2)}$, respectively. Let $\hat{\boldsymbol{\theta}}^*_{C(1)} = \operatorname{argmax}_{\boldsymbol{\theta}_{(1)}} \ell((\boldsymbol{\theta}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)})\,|\,\boldsymbol{y}^*)$. The null model bootstrap-based PIKL discrepancy estimator is thereby

$$d_{PIKL}\left(\hat{g}, (\boldsymbol{\theta}_{0(1)}, \hat{\boldsymbol{\theta}}_{0(2)})\right) = -2\ell(\boldsymbol{\theta}_{0(1)}, \hat{\boldsymbol{\theta}}_{0(2)}\,|\,\boldsymbol{y}),$$

and the alternative model bootstrap-based estimator is

$$d_{PIKL}\left(\hat{g}, (\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)})\right) = -2\ell(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}\,|\,\boldsymbol{y}).$$

The BDCP under the PIKL discrepancy is then

$$\begin{aligned} P^*_{PIKL} &= \Pr{}^*\left[d_{PIKL}\left(\hat{g}, (\boldsymbol{\theta}_{0(1)}, \hat{\boldsymbol{\theta}}_{0(2)})\right) < d_{PIKL}\left(\hat{g}, (\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)})\right)\right] \\ &= \Pr{}^*\left[-2\ell(\boldsymbol{\theta}_{0(1)}, \hat{\boldsymbol{\theta}}_{0(2)}\,|\,\boldsymbol{y}) < -2\ell(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}\,|\,\boldsymbol{y})\right]. \end{aligned}$$

We now introduce some notation that is needed in the proof establishing that the BDCP under the PIKL discrepancy approximates the LR p-value. Let the score vector based on the entire parameter vector be

$$U(\boldsymbol{\theta}\,|\,\boldsymbol{y}) = \frac{\partial \ell(\boldsymbol{\theta}\,|\,\boldsymbol{y})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}\,|\,y_i).$$

Let the elements of the score vector corresponding to the derivatives taken with respect to the parameters of interest be denoted

$$U_{(1)}(\boldsymbol{\theta}\,|\,\boldsymbol{y}) = \frac{\partial \ell(\boldsymbol{\theta}_{(1)}, \boldsymbol{\theta}_{(2)}\,|\,\boldsymbol{y})}{\partial \boldsymbol{\theta}_{(1)}} = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}_{(1)}} \ell_i(\boldsymbol{\theta}_{(1)}, \boldsymbol{\theta}_{(2)}\,|\,y_i).$$

We also need to partition the observed and expected informations into components consisting of the parameters of interest and nuisance parameters. For the observed information, write

$$I(\boldsymbol{\theta}\,|\,\boldsymbol{y}) = \begin{pmatrix} I_{11}(\boldsymbol{\theta}\,|\,\boldsymbol{y}) & I_{12}(\boldsymbol{\theta}\,|\,\boldsymbol{y}) \\ I_{21}(\boldsymbol{\theta}\,|\,\boldsymbol{y}) & I_{22}(\boldsymbol{\theta}\,|\,\boldsymbol{y}) \end{pmatrix}.$$

Let the expected information be similarly partitioned:

$$\mathscr{I}(\boldsymbol{\theta}) = \begin{pmatrix} \mathscr{I}_{11}(\boldsymbol{\theta}) & \mathscr{I}_{12}(\boldsymbol{\theta}) \\ \mathscr{I}_{21}(\boldsymbol{\theta}) & \mathscr{I}_{22}(\boldsymbol{\theta}) \end{pmatrix}.$$

Proposition 3.2: *Assuming that the large-sample properties of the MLEs hold, that the alternative model is adequately specified, and that the true parameter vector and the parameter* *vector under the null hypothesis lie in the interior of the parameter space, then for testing a partial null hypothesis of* $H_0 : \boldsymbol{\theta}_{(1)} = \boldsymbol{\theta}_{0(1)}$ *versus the alternative of* $H_A : \boldsymbol{\theta}_{(1)} \neq \boldsymbol{\theta}_{0(1)}$,

$$p_{LR} \approx P^*_{PIKL}.$$

Proof: We begin by stating well-known results pertaining to score statistics that will be applied later in the proof. First, recall that for large n, under certain regularity conditions with the alternative model being adequately specified,

$$\boldsymbol{U}(\boldsymbol{\theta} \mid \boldsymbol{y}) \dot{\sim} N_{p_A}\left(\mathbf{0}, \mathscr{I}(\boldsymbol{\theta})\right).$$

Thus, the score vector for the parameters of interest also follows an approximate normal distribution:

$$\boldsymbol{U}_{(1)}(\boldsymbol{\theta} \mid \boldsymbol{y}) \dot{\sim} N_k\left(\mathbf{0}, \mathscr{I}_{11}(\boldsymbol{\theta})\right).$$

The preceding result leads to

$$\boldsymbol{U}^{\mathrm{T}}_{(1)}(\boldsymbol{\theta} \mid \boldsymbol{y})\mathscr{I}_{11}^{-1}(\boldsymbol{\theta})\boldsymbol{U}_{(1)}(\boldsymbol{\theta} \mid \boldsymbol{y}) \dot{\sim} \chi^2_k. \qquad (11)$$

We now start by adding $2\ell(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)} \mid \boldsymbol{y})$ to each side of the inequality that defines P^*_{PIKL}, yielding

$$\begin{aligned} P^*_{PIKL} &= \Pr{}^*[-2\ell(\boldsymbol{\theta}_{0(1)}, \hat{\boldsymbol{\theta}}_{0(2)} \mid \boldsymbol{y}) < -2\ell(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)} \mid \boldsymbol{y})] \\ &= \Pr{}^*[2\left(\ell(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)} \mid \boldsymbol{y}) - \ell(\boldsymbol{\theta}_{0(1)}, \hat{\boldsymbol{\theta}}_{0(2)} \mid \boldsymbol{y})\right) \\ &< 2\left(\ell(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)} \mid \boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)} \mid \boldsymbol{y})\right)] \\ &= \Pr{}^*\left[2\left(\ell(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)} \mid \boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)} \mid \boldsymbol{y})\right) > L\right]. \end{aligned} \qquad (12)$$

Thus, in order to complete the proof, we need to show that under Pr*, the term on the left-hand side of the inequality in (12) follows an approximate χ^2_k distribution:

$$2\left(\ell(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)} \mid \boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)} \mid \boldsymbol{y})\right) \dot{\sim} \chi^2_k.$$

To establish this result, consider taking a second-order Taylor series expansion of $\ell(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)} \mid \boldsymbol{y})$ around $(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)})$. Write

$$\ell(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)} \mid \boldsymbol{y}) \approx \ell(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)} \mid \boldsymbol{y}) - \tfrac{1}{2}\left(\hat{\boldsymbol{\theta}}^*_{C(1)} - \hat{\boldsymbol{\theta}}_{(1)}\right)^{\mathrm{T}} I_{11}(\hat{\boldsymbol{\theta}} \mid \boldsymbol{y})\left(\hat{\boldsymbol{\theta}}^*_{C(1)} - \hat{\boldsymbol{\theta}}_{(1)}\right). \qquad (13)$$

Replacing the observed information with the expected information, approximation (13) implies that

$$2\left[\ell(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)} \mid \boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)} \mid \boldsymbol{y})\right] \approx \left(\hat{\boldsymbol{\theta}}^*_{C(1)} - \hat{\boldsymbol{\theta}}_{(1)}\right)^{\mathrm{T}} \mathscr{I}_{11}(\hat{\boldsymbol{\theta}})\left(\hat{\boldsymbol{\theta}}^*_{C(1)} - \hat{\boldsymbol{\theta}}_{(1)}\right). \qquad (14)$$

Applying a first-order Taylor series expansion of $\boldsymbol{U}_{(1)}(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)} \mid \boldsymbol{y}^*)$ about $(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)})$ leads to

$$\boldsymbol{U}_{(1)}(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)} \mid \boldsymbol{y}^*) \approx \boldsymbol{U}_{(1)}(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)} \mid \boldsymbol{y}^*) - I_{11}(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)} \mid \boldsymbol{y}^*)\left(\hat{\boldsymbol{\theta}}^*_{C(1)} - \hat{\boldsymbol{\theta}}_{(1)}\right).$$

Based on the definition of the conditional MLE $\hat{\boldsymbol{\theta}}^*_{C(1)}$, we have that $\boldsymbol{U}_{(1)}(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)} \,|\, \boldsymbol{y}^*) = \mathbf{0}$. Thus, we write

$$\boldsymbol{U}_{(1)}(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)} \,|\, \boldsymbol{y}^*) \approx I_{11}(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)} \,|\, \boldsymbol{y}^*) \left(\hat{\boldsymbol{\theta}}^*_{C(1)} - \hat{\boldsymbol{\theta}}_{(1)}\right). \tag{15}$$

For large n, the bootstrap distribution of $\boldsymbol{y}^*$ should mimic the sampling distribution of $\boldsymbol{y}$, thus leading to $I_{11}(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)} \,|\, \boldsymbol{y}^*) \approx I_{11}(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)} \,|\, \boldsymbol{y})$. Also for large n, under certain regularity conditions, we have that $I_{11}(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)} \,|\, \boldsymbol{y}) \approx \mathscr{I}_{11}(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)})$. Therefore, under these conditions, $I_{11}(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)} \,|\, \boldsymbol{y}^*) \approx \mathscr{I}_{11}(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)})$. This approximation in combination with (15) leads to

$$\hat{\boldsymbol{\theta}}^*_{C(1)} - \hat{\boldsymbol{\theta}}_{(1)} \approx \mathscr{I}^{-1}_{11}(\hat{\boldsymbol{\theta}}) \boldsymbol{U}_{(1)}(\hat{\boldsymbol{\theta}} \,|\, \boldsymbol{y}^*). \tag{16}$$

Result (16) implies that

$$\begin{aligned}
&\left(\hat{\boldsymbol{\theta}}^*_{C(1)} - \hat{\boldsymbol{\theta}}_{(1)}\right)^{\mathrm{T}} \mathscr{I}_{11}(\hat{\boldsymbol{\theta}}) \left(\hat{\boldsymbol{\theta}}^*_{C(1)} - \hat{\boldsymbol{\theta}}_{(1)}\right) \\
&\quad \approx \boldsymbol{U}^{\mathrm{T}}_{(1)}(\hat{\boldsymbol{\theta}} \,|\, \boldsymbol{y}^*) \left[\mathscr{I}^{-1}_{11}(\hat{\boldsymbol{\theta}})\right]^{\mathrm{T}} \mathscr{I}_{11}(\hat{\boldsymbol{\theta}}) \mathscr{I}^{-1}_{11}(\hat{\boldsymbol{\theta}}) \boldsymbol{U}_{(1)}(\hat{\boldsymbol{\theta}} \,|\, \boldsymbol{y}^*) \\
&\quad = \boldsymbol{U}^{\mathrm{T}}_{(1)}(\hat{\boldsymbol{\theta}} \,|\, \boldsymbol{y}^*) \mathscr{I}^{-1}_{11}(\hat{\boldsymbol{\theta}}) \boldsymbol{U}_{(1)}(\hat{\boldsymbol{\theta}} \,|\, \boldsymbol{y}^*).
\end{aligned} \tag{17}$$

Applying (11) to the bootstrapping context, we have

$$\boldsymbol{U}^{\mathrm{T}}_{(1)}(\hat{\boldsymbol{\theta}} \,|\, \boldsymbol{y}^*) \mathscr{I}^{-1}_{11}(\hat{\boldsymbol{\theta}}) \boldsymbol{U}_{(1)}(\hat{\boldsymbol{\theta}} \,|\, \boldsymbol{y}^*) \mathrel{\dot{\sim}} \chi^2_k. \tag{18}$$

Combining (17) and (18), we see that

$$\left(\hat{\boldsymbol{\theta}}^*_{C(1)} - \hat{\boldsymbol{\theta}}_{(1)}\right)^{\mathrm{T}} \mathscr{I}_{11}(\hat{\boldsymbol{\theta}}) \left(\hat{\boldsymbol{\theta}}^*_{C(1)} - \hat{\boldsymbol{\theta}}_{(1)}\right) \mathrel{\dot{\sim}} \chi^2_k.$$

In conjunction with (14), the preceding distributional result yields the desired distributional result:

$$2\left[\ell(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)} \,|\, \boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)} \,|\, \boldsymbol{y})\right] \mathrel{\dot{\sim}} \chi^2_k. \tag{19}$$

Finally, applying (19) to the inequality involving P^*_{PIKL} displayed in (12) allows us to assert

$$\begin{aligned}
P^*_{PIKL} &= \Pr{}^*\left[2\left(\ell(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)} \,|\, \boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)} \,|\, \boldsymbol{y})\right) > L\right] \\
&\approx \Pr[\chi^2_k > L] \\
&= p_{LR}.
\end{aligned}$$

■

We have thus shown that the LR p-value can be approximated by a BDCP under an appropriately chosen discrepancy, regardless of whether the null hypothesis pre-specifies all parameter values. In this subsection, we focus on the PIKL discrepancy because its BDCP provides an approximation to the LR p-value in the partial null setting. However, we again note that the BDCP under the conventional KL discrepancy is also a valid tool for choosing between two competing models in the partial null setting.

We have rigorously shown the connection between the LR p-value and the BDCP framework. However, if an appropriate discrepancy measure is chosen, the connection between the BDCP and most asymptotic tests can also be formally established. In fact, in work not shown, we have established this connection for the Wald and score test p-values. We focus on the connection between the KL discrepancy and LR p-value because of their utility and ubiquity. Evaluating the BDCP can be computationally expensive because it requires fitting the null and alternative models across numerous bootstrap samples. Fortunately, by justifying this connection between the p-value and the BDCP, one need not actually calculate the BDCP in order to interpret the p-value in the prescribed manner.

At first glance, the overall PIKL discrepancy may seem convoluted because it depends on the pseudo-true nuisance parameters, which are unknown. Thus, we cannot evaluate the overall PIKL. However, we are also unable to evaluate the conventional overall KL discrepancy, and thus the inability to evaluate the PIKL discrepancy causes no additional duress. Instead, both the PIKL and KL discrepancies are easily estimated using the bootstrap. The PIKL may also be appealing from a practical standpoint; if one is concerned with only the parameters of interest, then setting the nuisance parameters to their best possible values, as both the PIKL and its bootstrap-based estimator do, is reasonable. The PIKL discrepancy and its estimator are also akin to a plug-in likelihood in which the nuisance parameter vector is evaluated at its global MLE.

Various likelihood-based methods for contending with or eliminating nuisance parameters have been developed. One such method is the integrated likelihood, in which the joint likelihood of the parameters of interest and nuisance parameters $L(\boldsymbol{\theta}_{(1)}, \boldsymbol{\theta}_{(2)})$ is integrated with respect to the nuisance parameters $\boldsymbol{\theta}_{(2)}$. For a thorough treatment of integrated likelihood approaches, see [5]. For recommendations on which likelihood methods to use for likelihood ratio testing in the presence of censored and missing data, see [4]. While the integrated likelihood is a valuable alternative method of eliminating nuisance parameters, because of our desire to draw a connection between the BDCP and the standard LR p-value, we do not explore the integrated likelihood in this paper.

4. Simulation study

To further support the mathematical results showing the connection between the BDCP and the LR p-value, we have also performed a simulation study. The simulation study employs a factorial design composed of three factors. First, we consider both a linear and a logistic regression modeling framework. Within both frameworks, we consider full and partial null hypotheses. Finally, within each combination of modeling framework and type of null, we examine a setting in which the null is adequately specified and another setting in which the alternative is true and null is underspecified. Based on compiled results not shown, the sample size n affects the quality of the approximations more in the true alternative, false null setting than for an adequately specified null. Therefore, in the underspecified null settings, we present results for 2 sample sizes, namely $n = 100$ and $n = 1000$, and when the null is adequately specified, we present results for one sample size, $n = 500$.

In both the linear and logistic regression modeling frameworks, we draw independent samples of size n of an outcome variable y, as well as corresponding covariates x_1, x_2 and x_3. In the linear regression setting, for $i = 1, \ldots, n$, the generating model is of the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i,$$

where $\epsilon_i \backsim N(0, \sigma^2)$. Similarly, in the logistic regression setting, for $i = 1, \ldots, n$, the observed data is generated as $y_i \backsim bin(1, \pi_i)$, where

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}.$$

In both settings the distribution of the covariates is

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \backsim N_3\left(\begin{pmatrix} 2 \\ 2 \\ -2 \end{pmatrix}, \begin{pmatrix} 100 & 64 & 64 \\ 64 & 100 & 64 \\ 64 & 64 & 100 \end{pmatrix}\right).$$

In the linear regression setting, the alternative model always corresponds to the model in which the parameter vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^{\mathrm{T}}$ is unrestricted. In each linear regression simulation, the true variance is $\sigma^2 = 50$. In the full null setting, the null model sets $\boldsymbol{\beta} = \mathbf{0}$. While not typically done in practice, in the full null setting, the null model must provide a pre-specified value of σ^2, denoted by σ_0^2. The pre-specified σ_0^2 varies across simulation sets. Thus, we can write the hypotheses for the full null linear regression setting as $H_0 : \begin{pmatrix} \boldsymbol{\beta} \\ \sigma^2 \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \sigma_0^2 \end{pmatrix}$ versus $H_A : \begin{pmatrix} \boldsymbol{\beta} \\ \sigma^2 \end{pmatrix} \neq \begin{pmatrix} \mathbf{0} \\ \sigma_0^2 \end{pmatrix}$. The partial null in the linear regression framework tests $H_0 : \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ versus the general alternative. Simulation sets differ according to the values of the true parameter vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^{\mathrm{T}}$.

Like the linear regression simulations, in the logistic regression setting, the alternative model parameter vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^{\mathrm{T}}$ is unrestricted. In the logistic regression setting, the full null tests $H_0 : \boldsymbol{\beta} = \mathbf{0}$ versus the general alternative. Also similar to the linear regression setting, the partial null in the logistic regression setting corresponds to a test of $H_0 : \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ versus the general alternative. Simulation sets vary according to the values of the true parameter vector $\boldsymbol{\beta}$.

Each set of simulation results is based on drawing 50 original samples. From each original sample, we calculate the full and partial null LR test p-values. Based on $B = 10,000$ bootstrap samples, we evaluate the BDCP under the KL discrepancy in the full null setting and under the PIKL discrepancy in the partial null setting.

For each set of simulation results, we present a scatterplot with the corresponding p-value on the x-axis and the BDCP on the y-axis. A 45 degree line running through the origin is placed on each graph to aid in determining how close the BDCP is to its corresponding p-value. Each graph will contain 50 ordered pairs, one for each of the original samples. For each of the simulation results, we also present an estimate of the concordance correlation coefficient (CCC), labeled $\hat{\rho}_c$, which is a numerical measure of how close a set of ordered pairs falls to the line $y = x$ [18]. The CCC is a measure that lies between -1 and 1, inclusive, with $\rho_c = 1$ indicating exact agreement.

All simulations, calculations and scatterplots were performed and created using R [25] and RStudio [27].

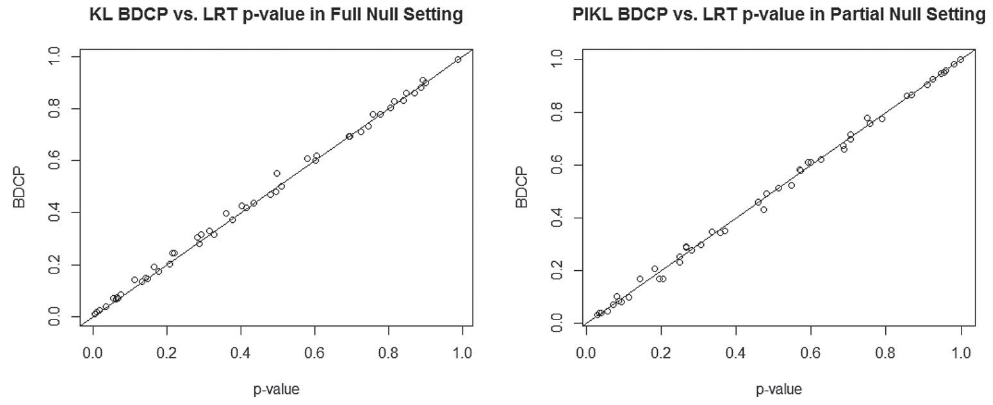

Figure 1. Scatterplots of BDCPs vs. their respective p-values in the linear regression setting with an adequately specified null. Here, $\beta_0 = 0, \sigma^2 = 50$ and $n = 500$.

Table 1. $\hat{\rho}_c$ comparing $\hat{P}^*$ to its respective p-value in the linear regression setting.

Null specification	n	β_0	σ_0^2	Null Type	Disc.	$\hat{\rho}_c$
Adequate	500	0	50	full null	KL	0.99878
Adequate	500	0		partial null	PIKL	0.99884
Underspecified	100	0.175	60	full null	KL	0.95098
Underspecified	100	0.175		partial null	PIKL	0.99235
Underspecified	1000	0.05	55	full null	KL	0.99590
Underspecified	1000	0.05		partial null	PIKL	0.99910

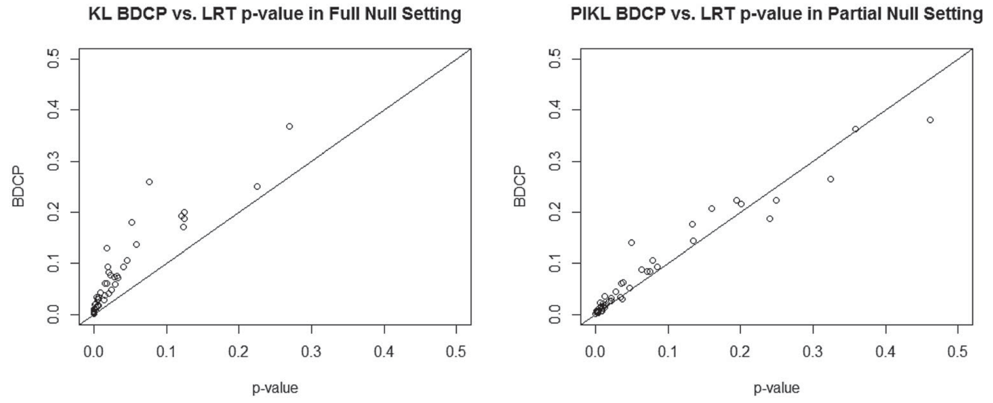

Figure 2. Scatterplots of BDCPs vs. their respective p-values in the linear regression setting with an underspecified null. Here, $\beta_0 = 0.175, \sigma^2 = 60$ and $n = 100$.

4.1. Linear regression

Figure 1 presents the scatterplots comparing the BDCPs to their respective p-values in the adequately specified null setting. The first two entries in the last column of Table 1 present the CCC values comparing the BDCPs to their respective p-values.

In the underspecified null setting, we wish to avoid a scenario in which most p-values are very close to zero. To achieve this goal, as we increase the sample size, the absolute value of

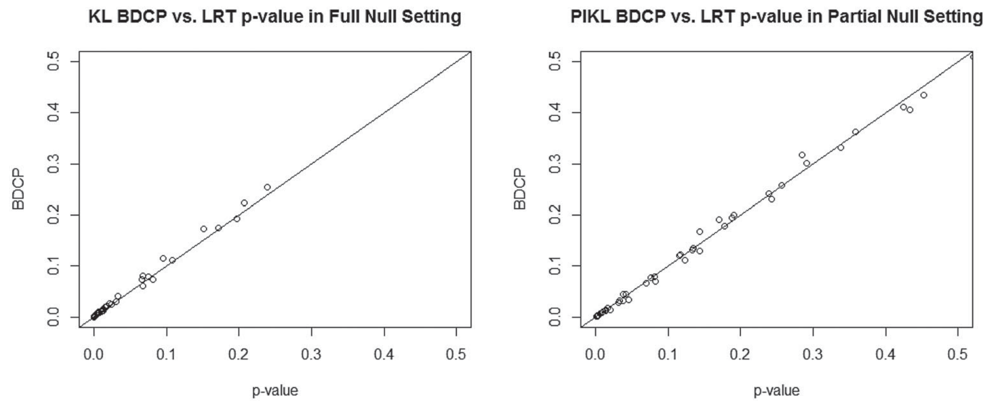

Figure 3. Scatterplots of BDCPs vs. their respective *p*-values in the linear regression setting with an underspecified null. $\beta_0 = 0.05, \sigma^2 = 55$ and $n = 1000$.

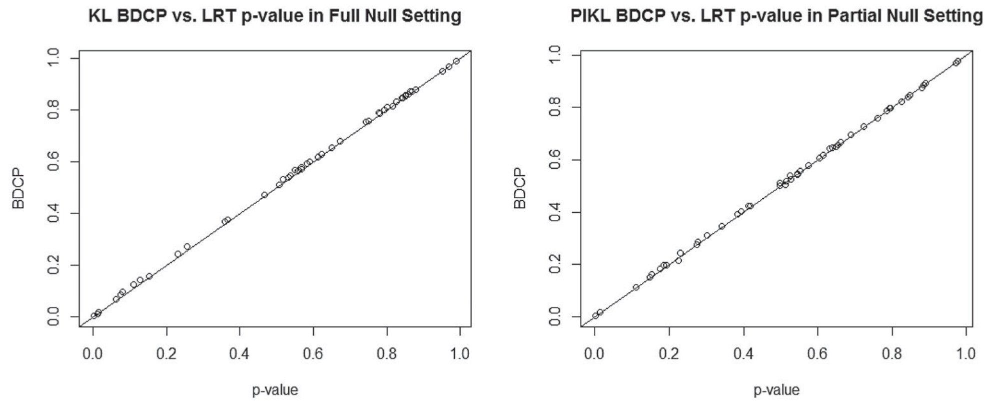

Figure 4. Scatterplots of BDCPs vs. their respective *p*-values in the logistic regression setting with an adequately specified null. Here, $\beta_0 = 0$, and $n = 500$.

the elements of the true parameter vector $\boldsymbol{\beta}$ must get smaller. With this in mind, for each of the three sample sizes, we vary $\boldsymbol{\beta}$ as well as the pre-specified null variance 'estimator' σ_0^2. For each sample size, samples are drawn from the generating distribution in which we set $\beta_0 = -\beta_1 = \beta_2 = -\beta_3$, and $\sigma^2 = 50$. For $n = 100$, we set $\beta_0 = 0.175$ and set the null model variance estimator to $\sigma_0^2 = 60$; and for $n = 1000$, set $\beta_0 = 0.05$ and $\sigma_0^2 = 55$.

Figure 2 presents scatterplots for the $n = 100$ setting; the $n = 1000$ setting is presented in Figure 3. Table 1 presents the CCC values for these two settings.

4.2. Logistic regression

The logistic regression results are presented in a fashion similar to the linear regression setting. In the adequately specified null setting $\boldsymbol{\beta} = \mathbf{0}$. In the underspecified null setting, we again set $\beta_0 = -\beta_1 = \beta_2 = -\beta_3$. When $n = 100$, we set $\beta_0 = 0.07$; and for $n = 1000$, $\beta_0 = 0.01$. The scatterplots for the adequately specified null, where we set $n = 500$, are presented in Figure 4. Figure 5 presents the scatterplots for the underspecified null with

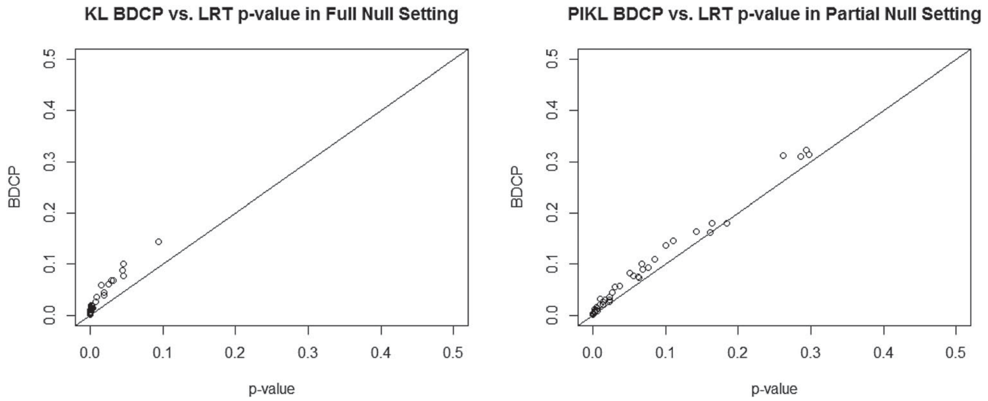

Figure 5. Scatterplots of BDCPs vs. their respective p-values in the logistic regression setting with an underspecified null. Here, $\beta_0 = 0.07$, and $n = 100$.

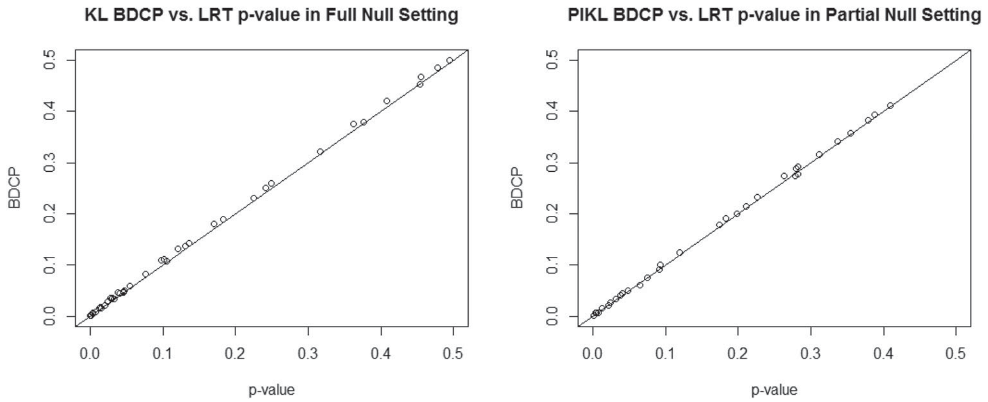

Figure 6. Scatterplots of BDCPs vs. their respective p-values in the logistic regression setting with an underspecified null. Here, $\beta_0 = 0.01$, and $n = 1000$.

Table 2. $\hat{\rho}_c$ comparing $\hat{P}^*$ to its respective p-value in the logistic regression setting.

Null Specification	n	β_0	Null Type	Disc.	$\hat{\rho}_c$
Adequate	500	0	full null	KL	0.99986
Adequate	500	0	partial null	PIKL	0.99980
Underspecified	100	0.07	full null	KL	0.83085
Underspecified	100	0.07	partial null	PIKL	0.99352
Underspecified	1000	0.01	full null	KL	0.99990
Underspecified	1000	0.01	partial null	PIKL	0.99993

$n = 100$; and Figure 6 presents results for $n = 1000$. Table 2 gives the CCC values for each of these combinations.

4.3. Interpretation of results

For both the linear and logistic regression modeling frameworks, Figures 1 and 4 support that when the null model is adequately specified, the p-values are closely approximated by their respective BDCPs. This finding is especially strong in the logistic regression setting, where most points on the scatterplot fall very close to the 45 degree line.

In the $n = 100$ underspecified null setting, as illustrated by Figures 2 and 5, the BDCPs exhibit a considerable amount of positive bias for their respective p-values in both the linear and logistic regression settings. The bias is more pronounced in the full null setting.

When we increase the sample size from $n = 100$ to $n = 1000$, the bias disappears. Figure 3 for the linear regression setting and Figure 6 for the logistic regression setting show that nearly all ordered pairs fall very close to the 45 degree line.

We find that the simulation results strongly support the mathematical findings that connect the BDCPs to their respective p-values. For adequately specified null hypotheses, the approximations hold quite well for moderate sample sizes. For underspecified null hypotheses, the approximations improve as the sample size increases. This, however, is to be expected because the proofs connecting the BDCPs to their p-values rely on large sample theory.

5. Benefits of DCP / BDCP framework

In this section, we address two important contributions of this work. First, by drawing a connection between the LR p-value and the BDCP, we can provide an alternative interpretation of the p-value, from which we gain new insights regarding the behavior of the p-value. Second, while we have shown a connection between the LR p-value and the BDCP when hypothesis testing assumptions are met, the BDCP framework can be applied to a considerably broader collection of settings than those in which the p-value is valid.

5.1. Insights gained from BDCP / p-value connection

The standard interpretation of the p-value is arguably confusing and counterintuitive, especially to students or researchers who must use statistics in their work but who may not specialize in the field. By drawing a connection between the LR p-value and the BDCP when hypothesis testing assumptions are met, we allow for a perhaps more intuitively pleasing interpretation of the p-value. Instead of interpreting the LR p-value in the usual manner, we can instead interpret it as a reflection of the probability, based on the sample at hand, that the fitted null model is closer to the 'truth' than the fitted alternative, where proximity is based on a suitably chosen discrepancy. In other words, rather than assuming the null is true and calculating a quantity that reflects the probability of what was observed under this assumption, we can instead think of the LR p-value as a bootstrap-based probability that the null is, in a certain sense, 'better' than the alternative, without the null having to be strictly true. This interpretation offers a frequentist interpretation of the p-value that is better aligned with assessing the probability of a hypothesis.

Providing a non-standard interpretation of the p-value also allows us to assess its behavior in a different light. For instance, consider a setting in which a subtle effect yields a small p-value due to a very large sample size. The standard interpretation of the p-value indicates that the result is statistically significant even though it may not be of practical importance. Viewing this phenomenon in the BDCP framework may be beneficial. In choosing between competing models based on the BDCP, concepts pertinent to statistical modeling naturally arise. For instance, there is a bias-variability tradeoff that is inherent in statistical modeling;

a larger model should be less biased, but at the cost of increased variability. A small BDCP indicates that, for most bootstrap samples, the discrepancy tends to prefer the more complex fitted alternative model to the simpler fitted null model, since the subtle effect can be estimated with sufficient accuracy to justify its inclusion. Stated another way, the adverse impact of the increased variability of the alternative model is outweighed by the impact of the null model bias due to omitting a nonzero effect. This interpretation of the BDCP may provide a clearer way of understanding the bias-variability tradeoff than the standard interpretation of the p-value.

5.2. Broader utility of DCP / BDCP framework

Beyond the advantage of providing a new interpretation of the LR p-value, the BDCP also has the strength that it can be applied in a broader collection of settings than hypothesis testing. For instance, hypothesis testing requires the alternative model to be adequately specified or else the corresponding p-value may be invalid. The BDCP, on the other hand, provides a valid comparison of competing models, regardless of the veracity of the alternative model (although the BDCP may not approximate the p-value in this setting). Because the notion of either the null or alternative being true is hard to defend in many practical settings, this advantage of the BDCP greatly enhances its utility.

Hypothesis testing typically requires the null model to be nested within the larger alternative, but the BDCP under the KL and PIKL discrepancies does not require nested models in order to be valid. There are many settings in which we would like to compare nonnested models. For instance, suppose we wish to compare a model that enters an effect linearly and another that enters the effect as a categorical variable. Standard hypothesis testing cannot be used to distinguish between these models, while the BDCP under the KL or PIKL discrepancies can easily be used.

Formal hypothesis testing requires pre-planned hypotheses in order to control Type I and Type II error rates. However, in practice, hypothesis testing is often applied in instances in which the data is used to make decisions regarding the selection of model and which hypotheses to test. While standard hypothesis testing techniques will typically no longer control for Type I and Type II error rates when applied in this manner, such hypothesis tests can still provide useful information regarding the implausibility of the null hypothesis. Nevertheless, in instances in which the hypotheses are not strictly pre-planned, a model evaluation tool that is not associated with the formality of controlling error rates may be preferable, because the use of such a tool could reduce the risk of incorrect interpretations. The BDCP simply seeks to quantify which of two models is closer to the truth and is therefore unconcerned with these long-term error rates. Accordingly, the BDCP is well-suited for use in settings without pre-planned hypotheses because it may produce a model evaluation which is less likely to be misconstrued than the evaluation provided by the p-value.

The BDCP can be evaluated for any discrepancy in which the plug-in principle can be applied. For instance, suppose we were interested in assessing the sum of weighted absolute deviations between a fitted model's J-dimensional parameter vector $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_J)^{\mathrm{T}}$ and the true parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_J)^{\mathrm{T}}$. Then, for $j = 1, \ldots, J$, we could put forth a discrepancy based on the weighted sum of scaled absolute deviations

(WAD), such as

$$d_{WAD}(g, \tilde{\boldsymbol{\theta}}) = \sum_{j=1}^{J} \frac{w_j |\tilde{\theta}_j - \theta_j|}{\sqrt{\iota^{jj}(\boldsymbol{\theta})}},$$

where $\iota^{jj}(\cdot \mid \boldsymbol{y})$ represents the (j, j)th element of the inverse expected information matrix, and $w_1, \ldots, w_J$ are the user-defined weights. Let the estimated parameter vector using the bootstrap sample be denoted $\tilde{\boldsymbol{\theta}}^* = (\tilde{\theta}_1^*, \tilde{\theta}_2^*, \ldots, \tilde{\theta}_J^*)^{\mathrm{T}}$, and let the estimated parameter vector under the general model be denoted $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_J)^{\mathrm{T}}$. Then, applying the plug-in principle, the bootstrap-based estimator of the WAD discrepancy is

$$d_{WAD}(\hat{g}, \tilde{\boldsymbol{\theta}}^*) = \sum_{j=1}^{J} \frac{w_j |\tilde{\theta}_j^* - \hat{\theta}_j|}{\sqrt{i^{jj}(\hat{\boldsymbol{\theta}} \mid \boldsymbol{y})}},$$

where $i^{jj}(\cdot \mid \boldsymbol{y})$ represents the (j, j)th element of the inverse observed information matrix.

To assess whether the null or alternative model has a smaller weighted sum of scaled absolute deviations in a hypothesis testing framework would require developing distributional theory. However, distributional theory for absolute values is notoriously difficult, and thus deriving a p-value to compare models on the basis of absolute deviations would be challenging. Yet by applying the plug-in principle and drawing repeated bootstrap samples, one can easily evaluate the BDCP for this discrepancy. Importantly, the BDCP under the WAD discrepancy need not approximate a known p-value. In Section 6.3 we calculate the BDCP under the WAD discrepancy for a variety of weighting schemes.

In this paper we primarily focused on discrepancies whose BDCP approximates the LR p-value, but practitioners do not need to confine their choice of discrepancy to this small class. We believe that, regardless of the connection to the p-value, the BDCP can be a valuable piece of information in choosing between two competing statistical models.

6. ACE/ARB and survival study

In this section, we apply our methodology to a biomedical application. The application investigates the effects of certain blood pressure medications on the probability of one-year survival in a high-risk Medicare cohort. The primary purpose of this application is to show that the LR p-value is approximated by the BDCP under the PIKL discrepancy. We also provide secondary results to illustrate that the BDCP can be applied in settings that do not provide a connection with the p-value. Specifically, we first illustrate that the BDCP can employ discrepancies that do not yield an approximation to the p-value. We then use the BDCP to compare nonnested models, which standard hypothesis testing is unable to do.

6.1. Overview

The present study seeks to determine the effects of angiotensin converting enzyme inhibitors (ACEs) and angiotensin II receptor blockers (ARBs) on one-year survival for a high-risk Medicare population, all of whom have suffered an acute myocardial infarction (AMI). There exists some evidence supporting that ACEs and ARBs may be beneficial to patients who have suffered an AMI. For instance, Setoguchi *et al.* [29] found that the use of ACE/ARBs helped explain a reduction in post-AMI patient mortality from 1995 to 2004. Also, in a large clinical trial, known as The Survival and Ventricular Enlargement (SAVE) study, Pfeffer *et al.* [24] found treating patients who recently suffered an AMI with captopril, an ACE, led to a significantly decreased risk of mortality when compared to patients receiving a placebo. However, the mean age for patients in that clinical trial is 59.4 years ($sd = 10.6$), whereas the youngest a member of the present study's cohort can be is 66, and the mean age is 78.3 years ($sd = 7.9$). Thus, while the SAVE study presents strong evidence that captopril decreases the risk of mortality among a cohort of patients considerably younger than the present study's cohort, it is unable to definitively confirm that using an ACE is beneficial to a more elderly, and perhaps sicker, patient cohort.

The cohort consists of 8,682 Medicare beneficiaries, all of whom suffered an AMI (an inpatient stay with an ICD-9 diagnosis code of 410.x1) in 2007 or 2008. All patients were also discharged alive from the hospital stay in which the AMI was diagnosed and survived for at least 30 days post-discharge.

Unless a patient has a drug contraindication, medical practice dictates that patients suffering an AMI should typically be placed on either an ACE or an ARB [22,30]. Despite the medical recommendation, only 4,327 (49.8%) members of the cohort filled a prescription for an ACE or ARB in the month following their discharge. A patient was considered an ACE/ARB user if and only if he or she filled a prescription for an ACE/ARB within 30 days of discharge. Note that in this study we do not differentiate between ACEs and ARBs; we simply create an indicator of whether the patient filled a prescription for either of the drugs in the 30 days post-discharge.

To help better understand the relationship between ACE/ARB use and survival, Table 3 presents a 2 × 2 table of ACE/ARB use and one-year survival. From Table 3, we determine that the unadjusted odds ratio comparing ACE/ARB use and one-year survival is 1.686 (95% CI: (1.496, 1.900); $p < 0.0001$). Thus, this perhaps naive analysis suggests quite strongly that ACE/ARB use increases the probability of one-year survival. However, this analysis is unable to account for the fact that patients who fill a prescription for an ACE/ARB may be considerably different than patients who do not. The result may simply be indicating that patients who receive an ACE/ARB are healthier on average and are thus less likely to die. Therefore, rather than rely on unadjusted analyses, we instead employ a multivariable logistic regression model to assess the relationship between ACE/ARB use and one-year survival. The model will control for a variety of covariates, including measures of patient demographics and socioeconomic status, measures of patient severity, comorbidities, drugs taken before the AMI, procedures before and during the AMI stay, drug contraindications, etc. To determine the importance of ACE/ARB use, we test the null hypothesis $H_0 : \beta = 0$ versus the general alternative $H_A : \beta \neq 0$, where β is the parameter corresponding to ACE/ARB use. The null model includes the control variables, and the alternative model includes the same control variables and the ACE/ARB indicator.

6.2. Primary results

To further understand this application, consider that the adjusted odds ratio estimate comparing ACE/ARB use and one-year survival is 1.151 with a Wald-based 95% confidence interval of $(1.002, 1.322)$. Thus, at the 0.05 significance level, we find ACE/ARB use to

Table 3. 2 × 2 table showing the relationship between ACE/ARB use and one-year post-index discharge survival. A patient was considered an ACE/ARB user if and only if he or she filled a prescription for an ACE or ARB within 30 days of the index discharge date.

		One-Year Survival	
		yes	no
ACE/ARB Use	yes	3814	513
	no	3550	805

increase the probability of one-year survival, holding all other covariates constant. However, the adjusted result is considerably less significant ($p = 0.0440$) than the unadjusted results ($p < 0.0001$). Also, the estimated effect size is smaller in the adjusted results, with an estimated odds ratio of 1.151, whereas the unadjusted odds ratio estimate is 1.686. This illustrates the general concept that, when using observational data, the effect of a treatment may not be adequately characterized if we do not control for important covariates related to the probability of receiving treatment.

For this application, the LR p-value is 0.0440, and the BDCP under the PIKL discrepancy is 0.0480. These results suggest that we can then interpret the LR p-value as a bootstrap-based estimator of the probability that the fitted null model will have smaller overall PIKL discrepancy than the fitted alternative. The idea of ACE/ARB use having no effect is not a scientifically valid hypothesis, so rejection of the null adds little to the underlying science. Instead, we may interpret the BDCP, and by its approximate equivalence the LR p-value, as a low probability that the model that does not account for ACE/ARB use is better than the alternative model that includes this predictor, without having to assume either candidate model precisely matches the truth. We can then conclude that the information from the sample is enough to estimate the effect of ACE/ARB use on survival with sufficient accuracy.

6.3. Secondary results

As previously mentioned, use of the BDCP need not be limited to discrepancies that yield a connection with the p-value. To illustrate that the BDCP can be defined and estimated for arbitrary discrepancies, we estimate the BDCP under the WAD discrepancy, which is described in Section 5.2. The BDCP will again be comparing the null model, which sets the ACE/ARB parameter to zero, and the alternative model, which estimates the ACE/ARB parameter. The WAD discrepancy requires a user-specified weighting scheme, so we apply a variety of weighting schemes to compare the two models. In each scheme, the ACE/ARB parameter receives a certain weight w, with the remaining $1 - w$ weight being distributed equally among the remaining 101 model parameters. We will let the weight on the ACE/ARB parameter be $w = 1, 0.50, 0.10, 0.05, 0.02$ and 0.0098. The weight of $w = 0.0098$ constitutes equal weighting across all model parameters. If a practitioner is interested only in the ACE/ARB indicator, then assessing the null and alternative models with $w = 1$ is a reasonable choice. On the other hand, if one is interested in an overall

Table 4. The BDCP under the WAD discrepancy for a variety of weighting schemes. The weighting schemes place probability w on the ACE/ARB parameter and distribute the remaining $1-w$ equally across the remaining parameters.

w	1	0.50	0.10	0.05	0.02	0.0098
$\hat{P}^*$	0.0504	0.0568	0.0824	0.1121	0.2630	0.2721

assessment of the model, without singling out any particular parameter, then calculating the BDCP with $w = 0.0098$ is justified.

Table 4 displays the BDCPs under the WAD discrepancy with the given weighting schemes. From Table 4, first note that each BDCP is less than 0.50, indicating that the alternative model is preferred in a majority of bootstrap samples for each studied weighting scheme. However, as the weight placed on the ACE/ARB indicator decreases, the corresponding BDCP increases. In other words, the null model fares better when it is compared across all parameters than when more weight is placed on its 'estimator' of the ACE/ARB parameter. For instance, if we compare the models only on the basis of the ACE/ARB parameter estimator (i.e. using $w = 1$), then the null model is preferred in a small percentage (5.04%) of bootstrap samples. On the other hand, when each parameter receives equal weighting, then the null model has a smaller discrepancy estimate in more than a quarter (27.21%) of bootstrap samples. Thus, depending on how we compare models, considerably different model assessments are possible.

This result illustrates that the BDCP can be evaluated for discrepancies that do not necessarily have a connection with the p-value. All that is required for use of the BDCP is successful application of the plug-in principle, so users have wide latitude in choosing appropriate context-specific discrepancies. While the connection between the BDCP and the p-value is useful, practitioners may in certain instances prefer to use a discrepancy that does not provide such an approximation.

Suppose now that rather than determining the effect of ACE/ARBs, we are instead interested in determining whether age should be entered linearly or categorically. Using the same modeling framework as before, we compare two models that contain the same set of predictors, except that the 'null' model enters age linearly, and the 'alternative' model enters ages categorically, with categories of 66–70, 71–75, 76–80, 81–85 and over 85. Here, the BDCP under the KL discrepancy is approximately 0.519, indicating no strong preference between the competing models. This analysis illustrates that the BDCP can easily compare nonnested models, a comparison for which standard hypothesis testing cannot be used.[1]

7. Concluding remarks

When evaluating models on the basis of discrepancy functions, we merely wish to know which model is most congruous with the truth, without having to assume one of the candidate models is true. On the other hand, the paradigm for hypothesis testing typically assumes one of two competing models is precisely true. Despite these differences in underlying philosophy, when the assumptions of hypothesis testing are met, we have shown that a bootstrap-based discrepancy comparison probability estimator can approximate the likelihood ratio (LR) p-value.

The primary purpose of this paper is to introduce the discrepancy comparison probability (DCP) and show that, under specifically formulated discrepancies, its bootstrap-based estimator approximates the LR p-value. Because the bootstrap-based DCP (BDCP) approximates the LR p-value, our work does not alleviate many of the problems with or abuses of the p-value. Instead, this methodology allows us to conceptualize the p-value in a different way. The alternative interpretation of the p-value that our work provides is better aligned with assessing a specific type of probability on the null hypothesis, without having to assume the null hypothesis matches the truth. Notably, one can interpret the p-value in this alternate fashion, without having to evaluate the BDCP across numerous bootstrap samples. We chose to focus on the relationship between the Kullback-Leibler discrepancy and the LR test p-value. However, in work not shown, we have also established a connection between the BDCP and the Wald and score test p-values using suitably defined discrepancy measures.

While the principal goal of this paper is to connect the BDCP with the p-value, the BDCP is more broadly applicable than hypothesis testing and the p-value. For instance, to establish a connection between the BDCP and the p-value, we needed to assume that the larger model was adequately specified. However, this assumption does not have to be met in order for the BDCP to be valid. Also, unlike the p-value derived from standard hypothesis testing, the BDCP under certain discrepancies can compare nonnested models. Further, while we have only considered the BDCP under a small number of discrepancies, the methodology presented here can be implemented for any discrepancy in which the plug-in principle can be applied. Finally, while the BDCP as it is formulated in this paper can only compare two models, using bootstrap-based estimators of overall discrepancies to delineate between models need not be limited to comparisons of only two, as illustrated by the methodology presented in [21].

Note

1. The R software that was used to produce the results for the simulation study and the application can be obtained by request from the first author, and is available on github at https://github.com/briedle-lilly/bdcp.

Acknowledgments

We wish to thank the referees for their valuable feedback, which served to substantially improve the original version of this manuscript. The content of this manuscript, including incorporated figures and tables, has been derived from the PhD thesis of the first author [26]. This content has not been published elsewhere.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Joseph E. Cavanaugh http://orcid.org/0000-0002-0514-7664

References

[1] H. Akaike, *Information theory and an extension of the maximum likelihood principle*, in *2nd International Symposium on Information Theory*, B. N. Petrov and F. Csáki, eds., Akadémia Kiadó, Budapest, 1973, pp. 267–281.

[2] H. Akaike, *A new look at the statistical model identification*, IEEE Trans. Automat. Contr. 19 (1974), pp. 716–723.

[3] D.W. Andrews, *Testing when a parameter is on the boundary of the maintained hypothesis*, Econometrica 69 (2001), pp. 683–734.

[4] N. Balakrishnan and M. Stehlik, *Likelihood testing with censored and missing duration data*, J. Stat. Theory Pract. 9 (2015), pp. 2–22.

[5] J.O. Berger, B. Liseo, and R.L. Wolpert, *Integrated likelihood methods for eliminating nuisance parameters*, Stat. Sci. 14 (1999), pp. 1–28.

[6] M. Bland, *Do baseline p-values follow a uniform distribution in randomized trials?* PLoS ONE 8 (2013), p. e76010.

[7] D. Boos and L. Stefanski, *P*-value precision and reproducibility, Am. Stat. 65 (2011), pp. 213–221.

[8] G.E. Box, *Science and statistics*, J. Am. Stat. Assoc. 71 (1976), pp. 791–799.

[9] K.P. Burnham and D.R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, Springer, New York, 2003.

[10] B. Efron, *Estimating the error rate of a prediction rule: Improvement on cross-validation*, J. Am. Stat. Assoc. 78 (1983), pp. 316–331.

[11] B. Efron, *How biased is the apparent error rate of a prediction rule?* J. Am. Stat. Assoc. 81 (1986), pp. 461–470.

[12] B. Efron and R.J. Tibshirani, *An Introduction to the Bootstrap*, CRC Press, Boca Raton, FL, 1994.

[13] B. Garel, *Recent asymptotic results in testing for mixtures*, Computational Statistics and Data Analysis 51 (2007), pp. 5295–5304.

[14] R. Gould and C. Ryan, *Introductory Statistics: Exploring the World through Data*, Pearson Education, London, England, 2013.

[15] D.H. Johnson, *Statistical sirens: The allure of nonparametrics*, Ecology 76 (1995), pp. 1998–2000.

[16] V. Johnson, *Revised standards for statistical evidence*, Proc. Natl. Acad. Sci. 110 (2013), pp. 19313–19317.

[17] S. Kullback and R.A. Leibler, *On information and sufficiency*, Ann. Math. Statist. 22 (1951), pp. 79–86.

[18] I. Lawrence and K. Lin, *A concordance correlation coefficient to evaluate reproducibility*, Biometrics 45 (1989), pp. 255–268.

[19] A.D. McQuarrie and C. Tsai, *Regression and Time Series Model Selection*, World Scientific, Hackensack, NJ, 1998.

[20] B. McShane, D. Gal, A. Gelman, C. Robert, and J. Tackett, *Abandon statistical significance*, Am. Stat. 73 (2019), pp. 235–245.

[21] A.A. Neath, J.E. Cavanaugh, and B. Riedle, *A bootstrap method for assessing uncertainty in Kullback-Leibler discrepancy model selection problems*, Math. Eng. Sci. Aerosp. 3 (2012), pp. 381–391.

[22] P. T. O'Gara, F. G. Kushner, D. D. Ascheim, D. E. Casey, M. K. Chung, J. A. de Lemos, S. M. Ettinger, J. C. Fang, F. M. Fesmire, B. A. Franklin, C. B. Granger, H. M. Krumholz, J. A. Linderbaum, D. A. Morrow, L. K. Newby, J. P. Ornato, N. Ou, M. J. Radford, J. E. Tamis-Holland, C. L. Tommaso, C. M. Tracy, Y. J. Woo, and D. X. Zhao, *2013 ACCF/AHA guideline for the management of ST-elevation myocardial infarction*, J. Am. Coll. Cardiol. 61 (2013), pp. e78–e140.

[23] R. Peng, *The reproducibility crisis in science: A statistical counterattack*, Significance 12 (2015), pp. 30–32.

[24] M. A. Pfeffer, E. Braunwald, L. A. Moyé, L. Basta, E. J. Brown, T. E. Cuddy, B. R. Davis, E. M. Geltman, S. Goldman, G. C. Flaker, M. Klein, G. A. Lamas, M. Packer, J. Rouleau, J. L. Rouleau, J. Rutherford, J. H. Wertheimer, and C. M. Hawkins, *Effect of captopril on*

mortality and morbidity in patients with left ventricular dysfunction after myocardial infarction: Results of the survival and ventricular enlargement trial, New Engl. J. Med. 327 (1992), pp. 669–677.

[25] R Core Team, R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, 2017. Available at http://www.R-project.org/.

[26] B.N. Riedle, *Probabilistic Pairwise Model Comparisons Based on Discrepancy Measures and a Reconceptualization of the p-Value*, Ph.D. thesis, Department of Biostatistics, University of Iowa, 2018. Available at https://ir.uiowa.edu/etd/6257/.

[27] RStudio Team, RStudio: Integrated development environment for R. *RStudio, Inc*, 2017. Available at http://www.rstudio.com/.

[28] S.G. Self and K.Y. Liang, *Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions*, J. Am. Stat. Assoc. 82 (1987), pp. 605–610.

[29] S. Setoguchi, R.J. Glynn, J. Avorn, M.A. Mittleman, R. Levin, and W.C. Winkelmayer, *Improvements in long-term mortality after myocardial infarction and increased use of cardiovascular drugs after discharge: a 10-year trend analysis*, J. Am. Coll. Cardiol. 51 (2008), pp. 1247–1254.

[30] S. C. Smith, E. J. Benjamin, R. O. Bonow, L. T. Braun, M. A. Creager, B. A. Franklin, R. J. Gibbons, S. M. Grundy, L. F. Hiratzka, D. W. Jones, D. M. Lloyd-Jones, M. Minissian, L. Mosca, E. D. Peterson, R. L. Sacco, J. Spertus, J. H. Stein, and K. A. Taubert, *AHA/ACCF secondary prevention and risk reduction therapy for patients with coronary and other atherosclerotic vascular disease: 2011 update*, Circulation 124 (2011), pp. 2458–2473.

[31] M. Stehlik and H. Wagner, *Exact likelihood ratio testing for homogeneity of the exponential distribution*, Comm. Statist. Simul. Comput. 40 (2011), pp. 663–684.

[32] D. Trafimow and M. Marks, *Editorial*, Basic. Appl. Soc. Psych. 37 (2015), pp. 1–2.

[33] R.L. Wasserstein and N.A. Lazar, *The ASA's statement on p-values: Context, process, and purpose*, Am. Stat. 70 (2016), pp. 129–133.

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