A predictive approach to measuring the strength of statistical evidence for single and multiple comparisons

May 12, 2011

This is a preprint of an article accepted for publication in *Canadian Journal of Statistics*

David R. Bickel

Ottawa Institute of Systems Biology

Department of Biochemistry, Microbiology, and Immunology

University of Ottawa

451 Smyth Road

Ottawa, Ontario, K1H 8M5

Abstract

The normalized maximum likelihood (NML) is a recent penalized likelihood that has properties that justify defining the amount of discrimination information (DI) in the data supporting an alternative hypothesis over a null hypothesis as the logarithm of an NML ratio, namely, the alternative hypothesis NML divided by the null hypothesis NML. The resulting DI, like the Bayes factor but unlike the p-value, measures the strength of evidence for an alternative hypothesis over a null hypothesis such that the

probability of misleading evidence vanishes asymptotically under weak regularity conditions and such that evidence can support a simple null hypothesis. Instead of requiring a prior distribution, the DI satisfies a worst-case minimax prediction criterion. Replacing a (possibly pseudo-) likelihood function with its weighted counterpart extends the scope of the DI to models for which the unweighted NML is undefined. The likelihood weights leverage side information, either in data associated with comparisons other than the comparison at hand or in the parameter value of a simple null hypothesis. Two case studies, one involving multiple populations and the other involving multiple biological features, indicate that the DI is robust to the type of side information used when that information is assigned the weight of a single observation. Such robustness suggests that very little adjustment for multiple comparisons is warranted if the sample size is at least moderate.

Keywords: indirect evidence; information criteria; information for discrimination; minimum description length; model selection; multiple comparison procedure; multiple testing; normalized maximum likelihood; penalized likelihood; reduced likelihood; side information; weighted likelihood; worst-case inference

MSC 2010 subject classification codes: 62F03; 62F12; 62F15; 62A01; 94A24; 94A45

1 Introduction

1.1 Quantifying statistical evidence

Many areas of science involve investigations of whether some effect is present and thus call for statistical methods that assess the evidence pertaining to whether a null hypothesis or an alternative hypothesis is closer to the system studied. For example, many experimental biologists are more interested in whether gene expression levels differ between control and treatment groups than in the effect size itself.

Because not all samples are representative of their populations, the amount of evidence against the null hypothesis is misleadingly high for some samples. Although the probability of observing such an unrepresentative sample should decrease as the size of the sample increases, that would not be the case if proximity of a p-value to 0 is interpreted as the strength of evidence against the null hypothesis. (It is thus interpreted by many knowledgeable statisticians of the Fisherian school (e.g., Efron and Gous, 2001) as opposed to the Neyman-Pearson school.) Indeed, the distribution of the p-value associated with a simple (point) null hypothesis remains uniform at all sample sizes if the null hypothesis holds, making the p-value impossible to interpret as a level of evidence apart from considering the sample size, as Royall (1997), Blume and Peipert (2003), and others have argued. The lacking property has been formalized by calling a measure of evidence *interpretable* only if, according to that measure, the frequentist probability of observing a sample that has misleading evidence converges to 0 as the sample size increases (Bickel, 2010b).

Another adverse consequence of treating the p-value as a measure of evidence is its inability to indicate evidence in favor of a simple null hypothesis. In general, the amount of information in the data that favors a simple null hypothesis cannot be quantified by the p-value since it can only indicate whether there is evidence against it.

Due to such failure of the p-value to qualify as an interpretable measure of evidence, Edwards (1992), Royall (1997), Blume and Peipert (2003), Bickel (2010b), and many others have

followed Fisher (1973, pp. 71-76, 96, 142) in interpreting the likelihood ratio $f_{\phi_1}(x)/f_{\phi_0}(x)$ as the strength of evidence favoring the hypothesis that the parameter ϕ is of value ϕ_1 rather than ϕ_0. Here, $f_\phi(x)$ is the probability density or mass function indexed by parameter value $\phi \in \{\phi_1, \phi_0\}$ and evaluated at x, the observed data vector. The interpretation is often called the *law of likelihood.* If the likelihood ratio exceeds some preset threshold such as 2^3 or 2^5 (Royall, 1997), there is sufficiently strong evidence for the alternative hypothesis over the null hypothesis, much as a p-value below a significance level is interpreted to a sufficient level of evidence against the null hypothesis. In Section 2.1, thresholds in addition to 2^3 and 2^5 will be suggested to attain the level of resolution that Jeffreys (1948) introduced for the Bayes factor, the most popular generalization of the likelihood ratio.

Some generalization is needed since the scope of the law of likelihood as stated is limited to the comparison of statistical hypotheses that are *simple* in the sense that each corresponds to a single probability function, either f_{ϕ_1} or f_{ϕ_0}. Thus, the law of likelihood does not apply to an alternative hypothesis that is *composite* in the sense that it corresponds to a parametric family of probability distributions. With the Bayes factor, Jeffreys (1948) extended the law to comparisons of composite hypotheses by integrating the likelihood over the nuisance parameters with respect to a prior distribution. Since the choice of that prior has a strong impact on the numeric value of the Bayes factor (Kass and Raftery, 1995; Aitkin, 2010, §2.8.2) and since the type of objectivity provided by automatic procedures is often desirable (Efron, 1986), attempts have been made to reduce or eliminate the need for subjective input. Arbitrary aspects of even the most successful proposals have prevented their widespread adoption (Aitkin, 2010, §2.9.5); one such aspect will be mentioned in Example 4.

To achieve both the objectivity of the p-value and the interpretability of the Bayes factor, this paper introduces a measure of evidence with roots in information theory. Its key element is a distribution that uses a minimax criterion to summarize the set of distributions that constitute each composite hypothesis. Such a distribution is *predictive*: it does not depend on x but rather is used to determine how well the hypothesis that it represents could have

predicted x. More precisely, if $\bar{f}_1$ and $\bar{f}_0$ are the predictive probability functions summarizing $\{f_\phi : \phi \in \Phi_1\}$ and $\{f_\phi : \phi \in \Phi_0\}$, respectively, then $\bar{f}_1(x)/\bar{f}_0(x)$ will be considered the strength of evidence in x that favors the hypothesis that $\phi \in \Phi_1$ over the hypothesis that $\phi \in \Phi_0$.

The closest analogue in wide use is the Bayes factor, which differs from $\bar{f}_1(x)/\bar{f}_0(x)$ only in that it generates each predictive function by averaging probability functions over a prior rather than generating the predictive probability function according to the minimax criterion. If π_1 and π_0 are the prior probability density functions with supports Φ_1 and Φ_0, respectively, then the Bayes factor measuring evidence for $\phi \in \Phi_1$ over $\phi \in \Phi_0$ would be $\int f_\phi(x)\,\pi_1(\phi)\,d\phi / \int f_\phi(x)\,\pi_0(\phi)\,d\phi$, a ratio of prior predictive densities. This paper will replace the prior predictive density $\int f_\phi(x)\,\pi_i(\phi)\,d\phi$ with the minimax predictive density $\bar{f}_i(x)$ for $i = 0, 1$, thereby eliminating dependence on π_1 and π_0.

In one of its simplest forms, that probability density is $\bar{f}_i(x) = C_i^{-1} \max_{\phi\in\Phi_i} f_\phi(\langle x, y_0\rangle)$, where $\langle x, y_0\rangle$ is the the vector formed by adding the pseudo-observation y_0 to the observed data vector x; C_i is the normalization constant ensuring that $\bar{f}_i$ is a probability density function on the sample space in which x lies. The pseudo-observation encodes incidental information, perhaps derived from data that do not directly bear on the value of ϕ or from the scientific knowledge that led investigators to consider the two hypotheses. The normalization constant prevents overfitting by automatically penalizing the maximum likelihood. In contrast with the Akaike information criterion, the Bayesian information criterion, and other penalized likelihoods derived from asymptotics, $\bar{f}_i(x)$ is defined for finite samples on the basis of a minimax criterion. These concepts will be more precisely stated in Section 2. In short,

$$\frac{\bar{f}_1(x)}{\bar{f}_0(x)} = \frac{C_1^{-1} \max_{\phi\in\Phi_1} f_\phi(\langle x, y_0\rangle)}{C_0^{-1} \max_{\phi\in\Phi_0} f_\phi(\langle x, y_0\rangle)} \tag{1}$$

is the proposed measure of the strength of statistical evidence for the hypothesis that $\phi \in \Phi_1$ over the hypothesis that $\phi \in \Phi_0$.

1.2 Organization of the paper

Section 1.1 provided the main reasons for seeking a new measure of statistical evidence. In order to derive equation (1), Section 2.1 will reinterpret a minimax predictive density borrowed from information theory. Those sections set the stage for presenting in Section 2.2 an exposition of the main advances made in this paper over the previous minimax predicted density: (i) that density applies to more models commonly encountered in statistical practice than does the existing minimax predictive density; (ii) the proposed minimax predictive density incorporates incidental information; (iii) the proposed reduction of data to lower-dimensional summary statistics can speed up the computation of the minimax predictive density.

Section 3 will present the proposed framework with greater mathematical precision and generality. Thus, whereas previous sections of this paper gradually extend a special case, this section will follow the complementary procedure of deriving special cases from the general framework. The redundancy is intended to facilitate understanding and intuition. Section 3 also includes a computationally efficient approximation and some results relevant to the probability of observing misleading statistical evidence.

One sense in which the extended framework is more general than the sketch in Section 2 is that it allows alternative pseudo-log-likelihood weights, including data-dependent weights. The problem of setting weights is addressed in Section 4.

The proposed minimax measure of quantifying statistical evidence methodology is illustrated in Section 5 with both a multiple-population data set and a multiple-feature data set. Finally, Section 6 concludes the paper with desirable properties of the new measure that go beyond those identified above.

The concept of a minimax predictive distribution needed for the measure of evidence is presented in this paper largely without the terminology of its origin in universal coding (Shtarkov, 1987). Discussion of the coding metaphor will be confined to Appendix A.

2 Heuristics

2.1 Worst-sample minimaxity

Consider the observed data vector $x \in \mathcal{X}^n$. Let $\mathcal{E}(\Omega)$ denote the set of all probability density functions on any sample space Ω, and let $\mathcal{F} = \{f_\phi : \phi \in \Phi\} \subset \mathcal{E}(\mathcal{X}^n)$ denote a parametric family of density functions on $\mathcal{X}^n$ for parameter space Φ. Herein, the probability densities are Radon-Nikodym derivatives, reducing to probability masses if $\mathcal{X}$ is countable. In the notation of equation (1), $\Phi \in \{\Phi_0, \Phi_1\}$. The maximum likelihood estimate of ϕ, denoted by $\hat{\phi}(x)$, is assumed to be unique.

The *logarithmic prediction error* of a probability density function $\bar{f} \in \mathcal{E}(\mathcal{X}^n)$ is $-\log \bar{f}(x)$, the negative logarithm of its probability density evaluated at x. The *regret* of a probability density function is the difference between its logarithmic prediction error and that of the best-predicting density function in $\mathcal{F}$. That is,

$$\text{regret}\left(\bar{f}, x; \Phi\right) \quad = \quad -\log \bar{f}(x) - \inf_{\phi \in \Phi} \left(-\log f_\phi(x)\right) = \sup_{\phi \in \Phi} \log \frac{f_\phi(x)}{\bar{f}(x)} \tag{2}$$

for any $x \in \mathcal{X}^n$. The $\mathcal{E}(\mathcal{X}^n)$-*minimax predictive density function relative to* $\mathcal{F}$,

$$\bar{f} = \arg \inf_{\bar{f} \in \mathcal{E}(\mathcal{X}^n)} \sup_{u \in \mathcal{X}^n} \text{regret}\left(\bar{f}, u; \Phi\right), \tag{3}$$

while by definition in $\mathcal{E}(\mathcal{X}^n)$, is not necessarily in $\mathcal{F}$. Rather, $\bar{f}$ is a probability density function that summarizes the entire family $\mathcal{F}$ with a single distribution, much as does a prior predictive density function (§1.1). Instead of averaging the members of $\mathcal{F}$ with respect to a prior distribution, the present definition employs $\mathcal{F}$ in equation (3) for each $u \in \mathcal{X}^n$ through the maximization of the likelihood over $\phi \in \Phi$, as seen by substituting u for x in equation (2).

Originally motivated in the information theory literature by a need to minimize code-length (Shtarkov, 1987), equation (3) defines the type of minimaxity employed. (Appendix

A briefly describes the codelength interpretation.) The predictive density function $\bar{f}$ solves the minimax problem involving all $u \in \mathcal{X}^n$, and thus for the observed sample $x \in \mathcal{X}^n$, rather than the more usual minimax problem involving an expectation value over all samples, as in the standard decision theory of frequentism. The following result (Shtarkov, 1987; Rissanen, 2007; Grünwald, 2007), to be proved in Section 3.2.1 for a more general optimization problem, sheds light on the nature of the minimaxity considered.

Theorem 1. *If* $\int_{\mathcal{X}^n} f_{\hat{\phi}(u)}(u)\, du < \infty$, *then* $\bar{f}$, *the* $\mathcal{E}(\mathcal{X}^n)$*-minimax predictive density function relative to* $\mathcal{F}$, *is the probability density function that satisfies*

$$\bar{f}(x) = \bar{f}(x; \Phi) = \frac{f_{\hat{\phi}(x)}(x)}{\int_{\mathcal{X}^n} f_{\hat{\phi}(u)}(u)\, du} \tag{4}$$

for all $x \in \mathcal{X}^n$.

Proof. This proof by contradiction is based on the direct proof given by Grünwald (2007, §6.2.1). Assume, contrary to the claim, that the density function $\bar{f}$ that satisfies equation (4) is not the optimal predictive density function. Since, for any $v \in \mathcal{X}^n$, the ratio $\bar{f}(v) / f_{\hat{\phi}(v)}(v)$ does not depend on v, it follows that, for any $\breve{f} \in \mathcal{E}(\mathcal{X}^n) \setminus \{\bar{f}\}$, there is a $v \in \mathcal{X}^n$ such that $\breve{f}(v) / f_{\hat{\phi}(v)}(v) < \bar{f}(v) / f_{\hat{\phi}(v)}(v)$. Therefore, given any $\breve{f} \in \mathcal{E}(\mathcal{X}^n) \setminus \{\bar{f}\}$, there is a $v \in \mathcal{X}^n$ such that $\text{regret}\left(\bar{f}, v; \Phi\right) < \text{regret}\left(\breve{f}, v; \Phi\right)$, which contradicts the assumption. □

Note that u, the dummy variable of integration over $\mathcal{X}^n$, appears twice in the integrand. For the observed $x \in \mathcal{X}^n$, the quantity $\bar{f}(x) = \bar{f}(x; \Phi)$ is called the *normalized maximum likelihood* (NML) with respect to Φ.

For many commonly used statistical models, the condition of the theorem does not hold since $\int_{\mathcal{X}^n} f_{\hat{\phi}(u)}(u)\, du = \infty$ under them. The main contribution of this paper is a variation of the NML that is defined for such models.

The density function $\bar{f}$ is worst-sample minimax in the sense that $\text{regret}\left(\bar{f}, x; \Phi\right)$, the regret due to the observed sample, cannot exceed $\sup_{u \in \mathcal{X}^n} \text{regret}\left(\bar{f}, u; \Phi\right)$, the regret due to the worst-case sample. Such observed-sample minimaxity suggests selecting the model or

Information (bits)	$(0, 1)$	$[1, 2)$	$[2, 3)$	$[3, 5)$	$[5, 7)$	$[7, \infty)$
Evidence grade	Negligible	Weak	Moderate	Strong	Very strong	Overwhelming

Table 1: Heuristic grades of evidence for an alternative hypothesis over a null hypothesis corresponding to intervals of the information for discrimination. The absolute value of a negative amount of information gives the grade of evidence favoring the null hypothesis. The discrimination information may also be viewed as an approximate change in log odds (§5.2.3).

hypothesis corresponding to the family $\mathcal{F}$ that minimizes the regret. Following the terminology of Kullback (1968) and Bickel (2010a), $-\log \bar{f}\,(x; \{\phi_0\}) - \big(-\log \bar{f}\,(x; \Phi)\big)$ would be the *information in x for discrimination* in favor of the alternative hypothesis that $\phi \neq \phi_0$ over the null hypothesis that $\phi = \phi_0$. Such information is an interpretable measure of evidence (§1.1) under general conditions and can quantify the strength of any evidence in favor of the null hypothesis as well as that of any evidence against it. More important, the information for discrimination uniquely quantifies the difference in how well each model or hypothesis predicts relative to ideal predictors.

Since the base of the logarithm is inconsequential, it may be chosen for convenience of interpretation. The binary logarithm ($\log_2$), yielding the number of *bits* of information, enables not only immediate exponentiation back to the ratio domain but also the use of grades of evidence that are both broad enough and refined enough for applications across scientific disciplines (Table 1). Except for the distinction between negligible and weak evidence, the grades closely mirror those Jeffreys (1948) originally proposed for the Bayes factor; cf. Bickel (2010b). Accordingly, the $[3, 5)$ grade of Table 1 is what Royall (1997, §1.12) considers "fairly strong evidence" for one simple hypothesis over another, and the $[5, 7)$ and $[7, \infty)$ grades together constitute his "quite strong evidence." Guidance on using Table 1 in applications will appear in Section 5.2.3.

2.2 Extending the worst-sample minimaxity

Despite the power of the minimaxity of the NML to uniquely quantify the strength of evidence in the form of discrimination information, three shortcomings render it impractical for use in many statistics applications. This paper generalizes the NML to overcome each limitation. The shortcomings will now be stated in order of decreasing severity.

First, the normalizing denominator of equation (4) is infinite for typical families of distributions, including the normal family. Each of the variant NMLs proposed to address the limitation introduces its own conceptual difficulties (Lanterman, 2005; Grünwald, 2007). For example, Rissanen (2007, §5.2.4), Rissanen and Roos (2007), and Grünwald (2007, §11.4.2) proposed conditional versions of the NML. Cf. related work by Takimoto and Warmuth (2000).

Second, the NML only uses information that is in x, but considering such information about the parameter in isolation from other available information can be misleading unless the sample size is sufficiently large. Additional information may be available in data from other populations, from other biological features such as genes or SNPs, or from other feature-feature comparisons. Even in the absence of such incidental data, there would be some information in the fact that the null hypothesis that $\phi = \phi_0$ is seriously considered.

Third, the integral $\int_{\mathcal{X}^n} f_{\hat{\phi}(u)}(u)\,du$ in the NML (4) tends to be difficult to compute for a moderate-to-large sample size n, for it requires integration over the entire n-dimensional sample space. The integral cannot generally be expressed in closed form, and numeric integration can be slow even for small n. Since an asymptotic expansion can be used to avoid such computation (Grünwald, 2007, chapter 7), the problem is most acute when n is too large for efficient computation and yet not large enough to warrant reliance on the asymptotics.

Heuristic outlines of the proposed solutions to these limitations are organized below into subsections. As the solution to the first limitation follows from that of the second, the latter will be explained first.

2.2.1 Limitation 2: Incidental information

The second problem can often be addressed by encoding the incidental information as a single observation $y_0 \in \mathcal{X}$, thereby increasing the effective sample size to $n+1$. (The practice of encoding such indirect evidence (Efron, 2010) as direct-evidence data has a long history; see, e.g., Edwards (1992, §3.5), Lele (2004), Greenland (2006), Bickel (2009), and especially Kass and Wasserman (1995).) The information encoded in y_0 may be incorporated by using

$$\bar{f}\left(x; y_0, \Phi\right) = \frac{f_{\hat{\phi}(\langle x, y_0 \rangle)}\left(\langle x, y_0 \rangle\right)}{\int_{\mathcal{X}^n} f_{\hat{\phi}(\langle u, y_0 \rangle)}\left(\langle u, y_0 \rangle\right) du} \tag{5}$$

instead of the NML of equation (4). Holding y_0 constant in the integration solves the minimax problem involving the worst-case sample x rather than the worst-case $\langle x, y_0 \rangle$, as will be seen in Section 3. The strength of evidence is then given by making the substitutions $\bar{f}_1\left(x\right) = \bar{f}\left(x; y_0, \Phi_1\right)$ and $\bar{f}_0\left(x\right) = \bar{f}\left(x; y_0, \Phi_0\right)$ in equation (1).

Let $x = \langle y_1, \ldots, y_n \rangle$. If the observations $y_1, \ldots, y_n$ are realizations of independent random variables, then $f_\phi\left(\langle x, y_0 \rangle\right) = f_\phi\left(y_0\right) \prod_{i=1}^{n} f_\phi\left(y_i\right)$, and the log-likelihood is $\log f_\phi\left(y_0\right) + \sum_{i=1}^{n} \log f_\phi\left(y_i\right) = \log f_\phi\left(y_0\right) + \log f_\phi\left(x\right)$. As the left-hand side makes clear, the log-likelihood is a weighted sum in which y_0 has $\left(100\%\right) / \left(n+1\right)$ of the total weight. That suggests replacing $f_\phi\left(\langle x, y_0 \rangle\right)$ in equation (5) with the weighted pseudo-likelihood $L\left(\theta; t\right)$ defined by $\log L\left(\theta; t\right) = w_0 \log g_\theta\left(t_0\right) + w \log g_\theta\left(t\right)$, where t_0, suppressed on the left-hand side, is the transformation of x_0 by the map that transforms x to t, $w_0 = \left(100\%\right) / \left(n+1\right)$, and $w = \left(100\%\right) n / \left(n+1\right)$. The NML based on $L\left(\theta; t\right)$ is often more computationally efficient than that based on $f_\phi\left(\langle x, y_0 \rangle\right)$ for the reasons to be given in Section 2.2.3.

In the simplest case, t_0 is a guessed value or a value determined by the null hypothesis $(\theta = \theta_0)$. For example, $\theta_0 = 0$ would yield $t_0 = 0$ if θ is a normal mean and t is the Student t-statistic.

A more complex case arises in the setting of multiple comparisons. For concreteness, suppose each comparison has a one-to-one correspondence with a gene or other biological

feature. Let $t_1, \ldots, t_N$ denote the statistics observed for N features. Each feature is considered to be the feature of *focus* when evaluating evidence in favor of its alternative hypothesis over its null hypotheses, with the other $N-1$ features considered to be *incidental* to that evaluation. For an example appropriate to most gene expression studies, a gene is in focus when one is computing its discrimination information in favor of the hypothesis of differential expression over the hypothesis of equivalent expression, and the other genes are incidental to that computation; details will appear in Example 2 of Section 3.1.1.

Supposing the first feature is currently in focus, $t = t_1$ in the above notation. A way to directly apply the above method of utilizing incidental information is to let t_0 be an average of the incidental statistics $t_2, \ldots, t_N$ such that each has equal weight. As a result, while t_1, the focus statistic, still has $w_1 = (100\%)\, n\, (n+1)^{-1}$ of the pseudo-log-likelihood weight, each of the $N-1$ incidental statistics is effectively left with only $w_i = (100\%)\, (N-1)^{-1}\, (n+1)^{-1}$ of the pseudo-log-likelihood weight ($i = 2, \ldots, N$).

The same distribution of weights is employed in the following method of incorporating incidental information. Let $\theta_1, \ldots, \theta_N$ denote the parameters of interest for the N features. With the first feature still in focus, the pseudo-log-likelihood $\sum_{i=1}^{N} w_i \log g_{\theta_1}(t_i)$ generates the NML by integrating over t_1 while holding $t_2, \ldots, t_N$ fixed. Thus, instead of directly averaging the incidental statistics, this method averages their pseudo-log-likelihoods. Only the latter average is invariant to whether the joint density $g_{\theta_1}(t_1) \cdots g_{\theta_1}(t_N)$ of independent incidental statistics is used in place of their marginal densities $g_{\theta_1}(t_1), \ldots, g_{\theta_1}(t_N)$.

At least for some applications, the above choice of pseudo-log-likelihood weights can probably the improved. Adaptive (data-dependent) weights will be considered in Section 3.2.1.

2.2.2 Limitation 1: Infinite normalization constant

As a by-product for commonly used distribution families, the solution to the second problem (§2.2.1) simultaneously solves the problem of an infinite value of $\int_{\mathcal{X}^n} f_{\hat{\phi}(u)}(u)\, du$ or

$\int_{\mathcal{T}} g_{\hat{\theta}(z)}(z)\,dz$. The finiteness of the analogous integral of the weighted likelihood is due to the fact that the integration is performed over the focus statistic with the incidental statistics held constant. The integrand then becomes negligible given sufficient distance of the variable of integration from the incidental statistics.

2.2.3 Limitation 3: Computational efficiency

To overcome the last of the three identified problems with NML, it is generalized in Section 3 by replacing the n-dimensional data vector x with a ν-dimensional statistic $t \in \mathcal{T}$ that is a function of the data; $\nu \leq n$. The distribution of the statistic must depend on θ, a parameter of interest in Θ, but not on λ, a scalar or vector nuisance parameter. Then $f_\phi = f_{\theta,\gamma}$, the density function of the original data, is replaced with g_θ, the density function of the statistic. The replacement is useful to the extent that the data relevant to the hypothesis about the interest parameter is largely confined to the statistic.

As a result, the ν-dimensional integral $\int_{\mathcal{T}} g_{\hat{\theta}(z)}(z)\,dz$ may be computed in place of the n-dimensional integral $\int_{\mathcal{X}^n} f_{\hat{\phi}(u)}(u)\,du$, where $\hat{\theta}(t) = \arg\sup_{\theta\in\Theta} g_\theta(t)$, the estimate defined by maximizing the pseudo-likelihood $g_\theta(t)$ over Θ, which is generally of smaller dimension than Φ. Thus, the speed of calculating the normalizing integral increases to the extent that $\nu \ll n$ and, when the likelihood functions must be maximized numerically, to the extent that $\dim\Theta \ll \dim\Phi$. The examples of this paper will use scalar statistics ($\nu = 1$) and scalar interest parameters ($\dim\Theta = 1$).

3 Predictive theory of evidence

This is the most abstract section of the paper. Section 2.2 is more straightforward, having only the special case of the proposed methodology that will be applied to data in Section 5.

The notation of Sections 1-2 will be extended to more general settings in Section 3.1. Sections 3.2 and 3.3 provide general definitions and properties of the minimax predictive

distribution and the minimax discrimination information, respectively.

3.1 Additional notation

3.1.1 Weighted likelihood

The technique of data reduction can simplify computations by eliminating a nuisance parameter from the full parameter ϕ. Consider a measurable map $\tau_n : \mathcal{X}^n \to \mathcal{T}(n)$. Let $\theta : \Phi \to \Theta$ denote a subparameter function such that the probability density of $\tau_n(X)$ is $g_{\theta(\phi)}(\tau_n(X))$, abbreviated as $g_\theta(\tau(X))$; the dependence of the density function g_θ on n is suppressed. Thus, the reduction of the data X to a statistic $\tau(X)$ has the effect of replacing the full parameter ϕ with the interest parameter θ. Important special cases of $L(\theta; \tau(x)) = g_\theta(\tau(x))$ as a function of θ are conditional likelihood functions and marginal likelihood functions (Royall, 1997; Severini, 2000; Bickel, 2010a).

The framework is now extended to N comparisons between null and alternative hypotheses. For each $i = 1, \ldots, N$, let $\mathcal{G}_{n,i} = \{g_{n,i,\theta} : \theta \in \Theta\} \subset \mathcal{E}(\mathcal{T}(n))$ denote the parametric family of density functions on $\mathcal{T}(n)$ for parameter space Θ. The "n" and "i" subscripts will be dropped when their values are clear. For the ith of N null hypotheses or comparisons, suppose $x_i \in \mathcal{X}^{n_i}$ is a realization of the random vector X_i of n_i independent components. Then each $T_i = \tau(X_i)$ is distributed with density $g_{\theta_i} = g_{n_i,i,\theta_i}$, and each outcome $t_i = \tau(x_i)$ is an element of $\mathcal{T}_i = \mathcal{T}(n_i)$. Let $L_i(\theta; \tau(x_i)) = g_{n,i,\theta}(\tau(x_i)) = g_\theta(\tau(x_i))$, giving each comparison its own likelihood function.

Mapping $\mathcal{X}^{n_i}$ to $\mathcal{T}_i = \mathbb{R}^D$ is common in data reduction applications in which $\Theta = \mathbb{R}^D$. Assigning a common parametric family to all comparisons ($\mathcal{G}_{n,i} = \mathcal{G}_{n,1}$ for all i) is usually appropriate when each comparison corresponds to a biological feature, as in Sections 2.2.1 and 5.2.

The observation $\mathbf{x} = \langle x_1, \ldots, x_N \rangle$ generates the statistic vector $\mathbf{t} = \langle t_1, \ldots, t_N \rangle = \langle \tau(x_1), \ldots, \tau(x_N) \rangle$, an outcome of $\mathbf{T} = \langle T_1, \ldots, T_N \rangle = \langle \tau(X_1), \ldots, \tau(X_N) \rangle$. For inference about θ_i on the basis of $\mathbf{t}$, the *weighted likelihood function* $\bar{L}_i(\bullet; \mathbf{t}) : \Theta \to [0, \infty)$ is

defined by

$$\log \bar{L}_i(\theta_i; \mathbf{t}) = \sum_{j=1}^{N} w_{ij} \log L_j(\theta_i; t_j), \tag{6}$$

where the weights $w_i = \langle w_{i1}, \ldots, w_{iN} \rangle$ are real numbers that may depend on $\langle n_1, \ldots, n_N \rangle$ and that satisfy $w_{ii} \geq w_{ij}$ (Hu and Zidek, 2002). The weights normally also conform to $\sum_{j=1}^{N} w_{ij} = 1$, a requirement that will be temporarily relaxed in Section 3.2.1. See Section 2.2.1 for a simple special case of a weighted likelihood.

Example 2. In most microarray studies, the expression levels of N genes are measured with the goal of determining which genes are differentially expressed between a treatment/perturbation group of m replicates and a control group of n replicates; each of these biological replicates represents one or more organisms. (Single-channel arrays do not require the pairing of replicates between groups as did the dual-channel arrays.) Following the typical assumption that intensity values are lognormally distributed, let $x_i = \langle x_{i1}, \ldots, x_{im} \rangle$ and $y_i = \langle y_{i1}, \ldots, y_{in} \rangle$ denote the logarithms of the intensities of the ith gene in the perturbation and control group, respectively. Let $X_{ij} \sim \mathrm{N}(\xi_i, \sigma_i^2)$ and $Y_{ij} \sim \mathrm{N}(\eta_i, \sigma_i^2)$ with realized values $X_{ij} = x_{ij}$ for $j = 1, \ldots, m$ and $Y_{ij} = y_{ij}$ for $j = 1, \ldots, n$. If θ_i is the absolute value of the *inverse coefficient of variation* $(\xi_i - \eta_i)/\sigma_i$, then t_i is conveniently taken as the absolute value of the two-sample, equal-variance t-statistic, which has a noncentral t distribution with noncentrality $(m^{-1} + n^{-1})^{-1/2} \theta_i$ and $m + n - 2$ degrees of freedom.

The sampling distribution of $\mathbf{T}$ is denoted by P to specify properties of the weights while accommodating model misspecification, the case that there is not a $\theta_i \in \Theta$ such that g_{θ_i} is a density admitted by the marginal distribution $P(T_i \in \bullet)$ for all $i \in \{1, \ldots, N\}$. With suitable weights and the assumption that $\hat{\theta}_i(\mathbf{T}) = \arg\sup_{\theta \in \Theta} \bar{L}_i(\theta; \mathbf{T})$ is almost surely unique for all $i \in \{1, \ldots, N\}$, the difference between $\hat{\theta}_i(\mathbf{T})$ and the conventional maximum likelihood estimator of θ_i almost surely converges to 0 as n_i diverges with N held fixed. Specifically, $w_{ii} = 1 + o_P(1)$ and $i \neq j \implies w_{ij} = o_P(1)$ ensure that $\hat{\theta}_i(\mathbf{T}) = \arg\sup_{\theta \in \Theta} L(\theta; x_i) + o_P(1)$, where the term $o_P(1)$ converges to 0 with P-probability 1 as $n_i \to \infty$ with any ratio n_j/n_k

bounded by constants: $n_j = O(n_k)$ for all $j, k \in \{1, \ldots, N\}$.

3.1.2 Prediction error

For some $\bar{g} \in \mathcal{E}(\mathcal{T}_i)$, the *generalized regret*

$$\text{regret}_i(\bar{g}, \mathbf{t}; \Theta) = -\log \bar{g}(\tau(x_i)) - \inf_{\theta \in \Theta} \left(-\log \bar{L}_i(\theta; \mathbf{t})\right) = \log \frac{\bar{L}_i\left(\hat{\theta}_i(\mathbf{t}); \mathbf{t}\right)}{\bar{g}(\tau(x_i))} \tag{7}$$

measures loss incurred by the likelihood associated with $\bar{g}$, the predictive distribution, relative to $\bar{L}_i\left(\hat{\theta}_i(\mathbf{t}); \mathbf{t}\right)$, the maximum weighted likelihood of θ_i. In other words, $\text{regret}_i(\bar{g}, \mathbf{t}; \Theta)$ is the discrepancy between the error in predicting the value of $\tau(x)$ on the basis of $\bar{g}$ and the prediction error minimized over the interest parameter. The latter error is easier to compute numerically and is potentially more relevant to hypotheses about the value of θ than is a regret minimized over the full parameter $\phi = \langle \theta, \lambda \rangle$. Thus, $\text{regret}_i(\bar{g}, \mathbf{t}; \Theta)$ replaces regret $\left(\bar{f}, x; \Phi\right)$ as the regret in the presence of the nuisance parameter or a nonzero weight other than w_{ii}.

3.2 Minimax predictive distribution

3.2.1 Exact predictive distribution

For each $t \in \mathcal{T}_i$, let $\mathbf{t}_i(t)$ denote the N-tuple of statistics that is equal to $\mathbf{t}$ in all components except the ith, which has t in place of t_i. For example, $\mathbf{t}_1(t) = \langle t, t_2, \ldots, t_N \rangle$, but $\mathbf{t}_i(t) = \langle t_1, \ldots, t_{i-1}, t, t_{i+1}, \ldots, t_N \rangle$ if $3 \leq i \leq N-2$.

The minimax predictive density function of Section 2.1 is a special case of

$$\bar{g}_i = \arg \inf_{\bar{g} \in \mathcal{E}(\mathcal{T}_i)} \sup_{t \in \mathcal{T}_i} \text{regret}_i(\bar{g}, \mathbf{t}_i(t); \Theta),$$

the $\mathcal{E}(\mathcal{T}_i)$-minimax *predictive density function relative to* $\langle \mathcal{G}, w_i \rangle$.

Theorem 3. *Given some* $i \in \{1, \ldots, N\}$ *and* $\mathbf{t} \in \mathcal{T}_1 \times \cdots \times \mathcal{T}_N$*, if* $\int \bar{L}_i \left(\hat{\theta}_i \left(\mathbf{t}_i \left(t \right) \right) ; \mathbf{t}_i \left(t \right) \right) dt < \infty$*, then, for all* $t_i \in \mathcal{T}_i$*, the* $\mathcal{E}\left(\mathcal{T}_i \right)$*-minimax predictive density function relative to* $\langle \mathcal{G}, w_i \rangle$ *satisfies*

$$\bar{g}_i \left(t_i \right) = \frac{\bar{L}_i \left(\hat{\theta}_i \left(\mathbf{t} \right) ; \mathbf{t} \right)}{\int_{\mathcal{T}_i} \bar{L}_i \left(\hat{\theta}_i \left(\mathbf{t}_i \left(t \right) \right) ; \mathbf{t}_i \left(t \right) \right) dt}. \tag{8}$$

Proof. The present argument follows that used to prove Theorem 1. Assume, contrary to the claim, that the density function $\bar{g}_i$ that satisfies equation (8) for all $t_i \in \mathcal{T}_i$ is not the minimax predictive density function relative to $\langle \mathcal{G}, w_i \rangle$. The substitution $\bar{L}_i \left(\hat{\theta}_i \left(\mathbf{t} \right) ; \mathbf{t} \right) = \bar{L}_i \left(\hat{\theta}_i \left(\mathbf{t}_i \left(t_i \right) \right) ; \mathbf{t}_i \left(t_i \right) \right)$ demonstrates that the ratio $\bar{g}_i \left(t_i \right) / \bar{L}_i \left(\hat{\theta}_i \left(\mathbf{t}_i \left(t_i \right) \right) ; \mathbf{t}_i \left(t_i \right) \right)$ does not depend on t_i. It follows that, for any $\breve{g}_i \in \mathcal{E}\left(\mathcal{T}_i \right) \setminus \{ \bar{g}_i \}$, there is a $t_i \in \mathcal{T}_i$ such that

$$\breve{g}_i \left(t_i \right) / \bar{L}_i \left(\hat{\theta}_i \left(\mathbf{t}_i \left(t_i \right) \right) ; \mathbf{t}_i \left(t_i \right) \right) < \bar{g}_i \left(t_i \right) / \bar{L}_i \left(\hat{\theta}_i \left(\mathbf{t}_i \left(t_i \right) \right) ; \mathbf{t}_i \left(t_i \right) \right).$$

Therefore, given any $\breve{g}_i \in \mathcal{E}\left(\mathcal{T}_i \right) \setminus \{ \bar{g}_i \}$, there is a $t_i \in \mathcal{T}_i$ such that $\operatorname{regret}_i \left(\bar{g}_i, \mathbf{t}_i \left(t_i \right) ; \Theta \right) < \operatorname{regret}_i \left(\breve{g}_i, \mathbf{t}_i \left(t_i \right) ; \Theta \right)$, which contradicts the assumption. □

For any $x_i \in \mathcal{X}^n$, the quantity $\bar{g}_i \left(\tau \left(x_i \right) \right) = \bar{g}_i \left(\tau \left(x_i \right) ; \Theta \right)$ is the *normalized maximum weighted likelihood (NMWL)* with respect to Θ or, more precisely, with respect to $\langle \mathcal{G}, w_i \rangle$. Important special cases of the NMWL appeared in Section 2.2.1.

Example 4. When the constraint that $\sum_{j=1}^{N} w_{ij} = 1$ is relaxed, NMWL generalizes various previous NMLs as follows. If $\tau \left(x_1 \right) = x_1$ for some $x_1 \in \mathcal{X}^{n_1}$ and if $N = 1$, the NMWL reduces to the probability density $\bar{g}_1 \left(x_1 \right)$ with $w_{1,1} = 1$ and thus to $\bar{f}\left(x \right)$, the NML of equation (4). For an observed vector $\langle y_1, \ldots, y_{n_1} \rangle \in \mathcal{X}^{n_1}$, assigning $N = 2$, $\theta_1 = \theta_2$, $t_1 = \langle y_1, \ldots, y_{n_1 - 1} \rangle$, and $t_2 = y_{n_1}$ demonstrates that the following conditional NMLs are NMWLs in the case of IID data. In particular, Grünwald (2007, §11.4.2) considered $\bar{g}_1 \left(\langle t_1, t_2 \rangle \right)$ with $w_{1,1} = w_{1,2} = 1$. Conversely, Rissanen (2007, §5.2.4) and Rissanen and Roos (2007) studied $\bar{g}_2 \left(\langle t_1, t_2 \rangle \right)$ with $w_{2,1} = w_{2,2} = 1$, thereby facilitating computation of the normalizing constant in equation (8) since the integration is only over a scalar. The main drawback of applying conditional NMLs

to the IID setting is the arbitrary nature of choosing an observation x_2 to leave out since the observations are not ordered in time (Grünwald, 2007, §11.4.3). The same issue arises in Bayesian model selection when an improper prior is conditioned on a minimal training sample before computing the Bayes factor. One solution is to take geometric or arithmetic averages over all possible minimal training samples (Berger and Pericchi, 2004). Analogous approaches to IID applications of conditional NMLs would likewise depend on arbitrary choices of averages and of sizes of the training samples (§1.1).

3.2.2 Approximate predictive distribution

Recall from Section 2.2.1 that the ith feature or comparison is said to be in *focus* when evaluating the strength of evidence in favor of the alternative hypothesis about θ_i over the null hypothesis about θ_i and that the $N-1$ comparisons not in focus are called *incidental*. A computationally efficient approximation to the NMWL is available if these *equal weight conditions* hold: (i) the weight of any comparison in focus does not depend on the index of the comparison, i.e., $w_{ii} = w_{1,1}$ for all i; (ii) the weight of any incidental comparison does not depend on the index of the comparison, i.e., $w_{ij} = w_{1,2}$ for all $i \neq j$; (iii) the sample sizes and sample spaces are equal, i.e., $n_i = n_1$ and $\mathcal{T}_i = \mathcal{T}_1$ for all i; (iv) all comparisons share a single family of distributions, i.e., $\mathcal{G}_{n_1,i} = \mathcal{G}_{n_1,1}$ and $L_1 = L_i$ for all i. Those four conditions are met in the special cases of Section 2.2.1.

The approximation simultaneously treats the focus comparison as the focus comparison and as an incidental comparison, thereby allowing reuse of the normalization constant when the focus shifts to another comparison. Thus, only a single computation of the normalization constant is required for the approximation of all N discrimination information values. Intuitively, the approximation is expected to be close if N is sufficiently high. That intuition will be stated precisely and confirmed.

The approximation uses the *approximate weights* $\tilde{w}_1, \ldots, \tilde{w}_{N+1}$, defined as follows under the equal weight conditions such that the focus comparison has two approximate weights

($\tilde{w}_{N+1}$ and $\tilde{w}_i$). Assigning two approximate weights to the focus comparison allow it to be considered both as the focus comparison without a change in weight ($\tilde{w}_{N+1} = w_{1,1}$) and as an incidental comparison weighted the same as the others. The sum of all approximate weights is the same as that of all original weights. Thus, when the first comparison is in focus, as it was in Section 2.2.1, $\tilde{w}_1 = (N-1) N^{-1} w_{1,2}$ is its incidental weight. Likewise, $\tilde{w}_j = (N-1) N^{-1} w_{1,2}$ for all $j = 1, \ldots, N$. Then $\sum_{j=1}^{N+1} \tilde{w}_j = N\tilde{w}_1 + \tilde{w}_{N+1} = (N-1) w_{1,2} + w_{1,1} = 1$.

For any $t \in \mathcal{T}_1$, let $\tilde{\mathbf{t}}(t)$ denote $\langle t_1, \ldots, t_N, t \rangle \in \mathcal{T}_1^{N+1}$. For inference about θ_i on the basis of $\mathbf{t}$, the *approximate weighted likelihood function* $\tilde{L}\left(\bullet; \tilde{\mathbf{t}}(t)\right) : \Theta \to [0, \infty)$ is defined by

$$\begin{aligned} \log \tilde{L}\left(\theta_i; \tilde{\mathbf{t}}(t)\right) &= \sum_{j=1}^{N} \tilde{w}_j \log L_j(\theta_i; t_j) + \tilde{w}_{N+1} \log L_j(\theta_i; t) \\ &= \frac{1 - w_{1,1}}{N} \sum_{j=1}^{N} \log L_1(\theta_i; t_j) + w_{1,1} \log L_1(\theta_i; t). \end{aligned}$$

Let $\hat{\theta}\left(\tilde{\mathbf{t}}(t)\right) = \arg\sup_{\theta \in \Theta} \tilde{L}\left(\theta; \tilde{\mathbf{t}}(t)\right)$.

The following theorem indicates that the exact NMWL (8) is approximated by

$$\tilde{g}_i(t_i) = \frac{\bar{L}_i\left(\hat{\theta}_i(\mathbf{t}); \mathbf{t}\right)}{\int_{\mathcal{T}_1} \tilde{L}\left(\hat{\theta}\left(\tilde{\mathbf{t}}(t)\right); \tilde{\mathbf{t}}(t)\right) dt},$$

which may be quickly calculated even for large N since the denominator, not depending on i, need only be computed once. Roughly speaking, Lemma 5 concludes that the weighted likelihood formed by using $\tilde{\mathbf{t}}(t)$ is approximately equal to the weighted likelihood formed by using $\mathbf{t}_i(t)$. Likewise, Lemma 6 means the corresponding maximum weighted likelihood estimates are approximately equal to each other. The practical importance of the lemmas is the establishment of Theorem 7, which justifies approximating $\bar{g}_i(t_i)$ by $\tilde{g}_i(t_i)$.

In the theorem and its supporting lemmas, $\tilde{\mathbf{T}}(t) = \langle T_1, \ldots, T_N, t \rangle$, and $\xrightarrow{\text{a.s.}}$ denotes almost sure convergence as N increases with n_1 fixed.

Lemma 5. *If the equal weight conditions hold and if $T_1, \ldots, T_N$ are drawn independently from a mixture distribution, then, for all $\theta \in \Theta$ and $t \in \mathcal{T}_i$,*

$$\log \tilde{L}\left(\theta; \tilde{\mathbf{T}}(t)\right) - \log \bar{L}_i\left(\theta; \mathbf{T}_i(t)\right) \overset{a.s.}{\to} 0. \tag{9}$$

Proof. According to the equal weight conditions,

$$\log \tilde{L}\left(\theta; \tilde{\mathbf{T}}(t)\right) - \log \bar{L}_i\left(\theta; \mathbf{T}_i(t)\right) \quad =$$

$$\left(\frac{1-\tilde{w}_{N+1}}{N}\sum_{j=1}^{N}\log L_1(\theta; T_j) + \tilde{w}_{N+1}\log L_1(\theta; T_i)\right) - \left(\frac{1-w_{ii}}{N-1}\sum_{j\neq i; i=1}^{N}\log L_1(\theta; T_j) + w_{ii}\log L_1(\theta; T_i)\right)$$

$$= (1-\tilde{w}_{N+1})\left(\frac{1}{N}\sum_{j=1}^{N}\log L_1(\theta; T_j) - \frac{1}{N-1}\sum_{j\neq i; i=1}^{N}\log L_1(\theta; T_j)\right).$$

The second factor almost surely vanishes by the law of large numbers. □

Lemma 6. *Under the assumptions of Lemma 5, the stipulations that $\hat{\theta}\left(\tilde{\mathbf{T}}(t)\right)$ and $\hat{\theta}_i\left(\mathbf{T}_i(t)\right)$ are almost always unique for all $t \in \mathcal{T}_i$ and that $L_1(\bullet; T_i)$ is almost surely continuous on Θ for all $i = 1, \ldots, N$ imply that, for all $i = 1, \ldots, N$ and $t \in \mathcal{T}_i$, $\hat{\theta}\left(\tilde{\mathbf{T}}(t)\right) - \hat{\theta}_i\left(\mathbf{T}_i(t)\right) \overset{a.s.}{\to} 0$.*

Proof. By Lemma 5, equation (9) holds for all $\theta \in \Theta$. Thus, since almost sure convergence is preserved under almost surely continuous transformations (Serfling, 1980, §1.7), $\arg\sup_{\theta\in\Theta} \tilde{L}\left(\theta; \tilde{\mathbf{T}}(t)\right) - \arg\sup_{\theta\in\Theta} \bar{L}_i(\theta; \mathbf{T}) \overset{\text{a.s.}}{\to} 0$. □

Theorem 7. *Under the assumptions of Lemma 6, the difference between the approximate and exact normalization constants almost surely vanishes:*

$$\int_{\mathcal{T}_1} \tilde{L}\left(\hat{\theta}\left(\tilde{\mathbf{T}}(t)\right); \tilde{\mathbf{T}}(t)\right) dt - \int_{\mathcal{T}_i} \bar{L}_i\left(\hat{\theta}_i\left(\mathbf{T}_i(t)\right); \mathbf{T}_i(t)\right) dt \overset{a.s.}{\to} 0. \tag{10}$$

Proof. Combining the results of Lemmas 5 and 6 gives

$$\tilde{L}\left(\hat{\theta}\left(\tilde{\mathbf{T}}(t)\right); \tilde{\mathbf{T}}(t)\right) - \bar{L}_i\left(\hat{\theta}_i\left(\mathbf{T}_i(t)\right); \mathbf{T}_i(t)\right) \overset{\text{a.s.}}{\to} 0$$

for all $t \in \mathcal{T}_i$ since $\mathcal{T}_i = \mathcal{T}_1$ and the functions are almost surely continuous by assumption. $\square$

3.3 Minimax discrimination information

The discrimination information presented in Section 1 for quantifying the strength of evidence will now be extended on the basis of the more general NMWL defined in Section 3.2.1. For any $\Theta' \subseteq \Theta$, let $\bar{g}_i(t_i; \Theta')$ denote the minimax predictive density function relative to $\langle \{g_\theta : \theta \in \Theta'\}, w_i \rangle$ as defined in Section 3.2.1. For any $\Theta_0, \Theta_1 \subseteq \Theta$, the *minimax information in x for discrimination* in favor of the hypothesis that $\theta_i \in \Theta_1$ over the hypothesis that $\theta_i \in \Theta_0$ is

$$\bar{I}_i(\Theta_1, \Theta_0) = -\log \bar{g}_i(t_i; \Theta_0) - (-\log \bar{g}_i(t_i; \Theta_1)), \tag{11}$$

generalizing quantities in Kullback (1968), Rissanen (1987), Bickel (2010b), and Bickel (2010a). The approximate minimax information $\tilde{I}_i(\Theta_1, \Theta_0)$ is defined identically except with $\tilde{g}_i$ of Section 3.2.2 in place of $\bar{g}_i$:

$$\tilde{I}_i(\Theta_1, \Theta_0) = -\log \tilde{g}_i(t_i; \Theta_0) - (-\log \tilde{g}_i(t_i; \Theta_1)). \tag{12}$$

Not restricted to the case of smoothness conditions on $\{g_\theta : \theta \in \Theta\}$, $\bar{I}_i(\Theta_1, \Theta_0)$ applies to any problem of selecting one of two models. Appendix A discusses the appropriateness of the information-theoretic terminology.

Since $\bar{g}_i(t_i; \Theta_1) / \bar{g}_i(t_i; \{\theta_0\})$ for any $\theta_0 \in \Theta$ is a likelihood ratio, the discrimination information has the Chebyshev bound on the probability under $\theta = \theta_0$ of observing discrimination information above any positive threshold k (Royall, 2000; Bickel, 2010a):

$$P\left(\bar{I}_i(\Theta_1, \{\theta_0\}) > k\right) \leq 1/k.$$

In addition, Section 1.1 suggests that, in order to interpret the discrimination information as the strength of evidence, the probability of observing discrimination information in favor

of the alternative hypothesis should asymptotically vanish if the null hypothesis is true:

$$\lim_{n_i \to \infty} P\left(\bar{I}_i\left(\Theta_1, \{\theta_0\}\right) > 0\right) = 0. \tag{13}$$

The next lemma and theorem bear on whether the minimax information for discrimination is an interpretable measure of evidence in that the probability of observing misleading information converges to 0 as $n_i \to \infty$. Let $\hat{\theta}_i\left(\mathbf{T}; \Theta'\right) = \arg\sup_{\theta \in \Theta'} \bar{L}_i\left(\theta; \mathbf{T}\right)$ for $i = 1, \ldots, N$ and $\hat{\theta}_0\left(t; \Theta'\right) = \arg\sup_{\theta \in \Theta'} L_i\left(\theta; t\right)$ given any $\Theta' \subseteq \Theta$.

Lemma 8. *Suppose* $\Theta = \mathbb{R}^D$, $\theta = \langle \theta_1, \ldots, \theta_D \rangle^{\mathrm{T}}$, $n_j = O\left(n_k\right)$ *for all* $j, k \in \{1, \ldots, N\}$, $\mathcal{G}_{n,i} = \mathcal{G}_{n,1}$ *and* $L_i = L_1$ *for all* $i \in \{1, \ldots, N\}$ *and sufficiently large* n, *and* $\tau\left(x\right) = x$ *for all* $x \in \mathcal{X}^n$, *which implies that* $\mathcal{T}_i = \mathcal{X}^n$, $t_i = x_i$, *and* $T_i = X_i$. *Assume also that for some* $i \in \{1, \ldots, N\}$, *there exists an open, bounded set* $\Theta' \subseteq \Theta$ *on which* $L\left(\bullet; X_i\right)$ *is almost surely continuous and such that*

$$\log \int_{\mathcal{X}^n} \bar{L}_i\left(\hat{\theta}_0\left(t; \Theta'\right); t\right) dt = \frac{D}{2} \log \frac{n_i}{2\pi} + \log \int_{\Theta'} \sqrt{\frac{1}{n_i} \left| E \frac{\partial^2 \ln g_\theta\left(X_i\right)}{\partial \theta \partial \theta^{\mathrm{T}}} \right|} d\theta + o\left(1\right). \tag{14}$$

$$\therefore \log \int_{\mathcal{T}_i} \bar{L}_i\left(\hat{\theta}_i\left(\mathbf{T}_i\left(t\right); \Theta'\right); \mathbf{T}_i\left(t\right)\right) dt = \frac{D}{2} \log \frac{n_i}{2\pi} + \log \int_{\Theta'} \sqrt{\frac{1}{n_i} \left| E \frac{\partial^2 \ln g_\theta\left(X_i\right)}{\partial \theta \partial \theta^{\mathrm{T}}} \right|} d\theta + o\left(1\right)$$

almost surely holds for any weights that satisfy $P\left(\lim_{n_i \to \infty} w_{ii} = 1\right) = 1$ *and* $\sum_{j=1}^{N} w_{ij} = 1$.

Proof. The continuity condition and the constraints on the weights and sample sizes ensure that $\bar{L}_i\left(\hat{\theta}_i\left(\mathbf{T}_i\left(t\right); \Theta'\right); \mathbf{T}_i\left(t\right)\right) \overset{\text{a.s.}}{\to} \bar{L}_i\left(\hat{\theta}_0\left(t; \Theta'\right); t\right)$ as $n_i \to \infty$ for all $t \in \mathcal{T}_i$. □

The assumptions of Lemma 8 are broadly applicable since equation (14) holds under general regularity conditions (Rissanen, 1996). The result will now be extended to non-bounded parameter spaces.

Theorem 9. *Suppose that* $\Theta_1 \subseteq \Theta$, *that* $n_j = O\left(n_k\right)$ *for all* $j, k \in \{1, \ldots, N\}$, *and that* P *is the sampling distribution of* $\mathbf{T}$. *Assume also that for any* $i \in \{1, \ldots, N\}$, *there exists an*

open, bounded set $\Theta' \subseteq \Theta_1$ *such that*

$$P\left(\lim_{n_i \to \infty} \log \int_{\mathcal{T}_i} \bar{L}_i \left(\hat{\theta}_i \left(\mathbf{T}_i \left(t\right); \Theta'\right); \mathbf{T}_i \left(t\right)\right) dt = \infty\right) = 1. \tag{15}$$

$$\therefore P\left(\lim_{n_i \to \infty} \log \int_{\mathcal{T}_i} \bar{L}_i \left(\hat{\theta}_i \left(\mathbf{T}_i \left(t\right); \Theta_1\right); \mathbf{T}_i \left(t\right)\right) dt = \infty\right) = 1.$$

Proof. Let $\mathfrak{T} = \left\{t \in \mathcal{T}_i : \hat{\theta}_i \left(\mathbf{T}_i \left(t\right); \Theta_1\right) \in \Theta'\right\}$ to expand $\int_{\mathcal{T}_i} \bar{L}_i \left(\hat{\theta}_i \left(\mathbf{T}_i \left(t\right); \Theta_1\right); \mathbf{T}_i \left(t\right)\right) dt$ as

$$\int_{\mathfrak{T}} \bar{L}_i \left(\hat{\theta}_i \left(\mathbf{T}_i \left(t\right); \Theta_1\right); \mathbf{T}_i \left(t\right)\right) dt + \int_{\mathcal{T}_i \backslash \mathfrak{T}} \bar{L}_i \left(\hat{\theta}_i \left(\mathbf{T}_i \left(t\right); \Theta_1\right); \mathbf{T}_i \left(t\right)\right) dt.$$

Thus, since $\bar{L}_i \left(\hat{\theta}_i \left(\mathbf{T}_i \left(t\right); \Theta_1\right); \mathbf{T}_i \left(t\right)\right) = \bar{L}_i \left(\hat{\theta}_i \left(\mathbf{T}_i \left(t\right); \Theta'\right); \mathbf{T}_i \left(t\right)\right)$ for all $t \in \mathfrak{T}$ and

$$\bar{L}_i \left(\hat{\theta}_i \left(\mathbf{T}_i \left(t\right); \Theta_1\right); \mathbf{T}_i \left(t\right)\right) > \bar{L}_i \left(\hat{\theta}_i \left(\mathbf{T}_i \left(t\right); \Theta'\right); \mathbf{T}_i \left(t\right)\right)$$

for all t in non-empty $\mathcal{T}_i \backslash \mathfrak{T}$ given any sufficiently large n_i,

$$\int_{\mathcal{T}_i} \bar{L}_i \left(\hat{\theta}_i \left(\mathbf{T}_i \left(t\right); \Theta_1\right); \mathbf{T}_i \left(t\right)\right) dt \geq \int_{\mathcal{T}_i} \bar{L}_i \left(\hat{\theta}_i \left(\mathbf{T}_i \left(t\right); \Theta'\right); \mathbf{T}_i \left(t\right)\right) dt$$

follows, where the equality and both inequalities hold with P-probability 1. □

Since the claim of Lemma 8 implies equation (15), Theorem 9 applies to the wide class of models satisfying the regularity conditions of Rissanen (1996). The largely overlapping regularity conditions of Sin and White (1996) then imply equation (13) when there is no $\theta \in \Theta$ such that g_θ is closer in Kullback-Leibler divergence than g_{θ_0} to the marginal distribution $P\left(T_i \in \bullet\right)$. In the special case of correct model specification considered in Royall (2000) and Bickel (2010a), equation (13) holds for all P admitting g_{θ_0} as the marginal density of T_i.

4 Weighting the predictive density

4.1 Data-dependent weights

The weighted likelihood was originally proposed for bias-variance trade-offs given relatively small n_i but potentially large N (Hu and Zidek, 2002). More formally, Wang and Zidek (2005a) derived the weighted likelihood from the minimization of Kullback-Leibler loss.

Many methods of adaptively assigning data-dependent weights have been proposed. For example, Wang and Zidek (2005b) and Wang (2006b) set the weights by cross validation, whereas Plante (2008) and Plante (2009) instead set them by constrained minimization.

Further research is needed to determine which existing methods are suitable for computing the minimax discrimination information and whether modifications or entirely new methods are needed. While the framework of Section 3 encompasses data-dependent weights, the conditions it imposes on them are weak enough to permit a wide variety of adaptive methods. From this viewpoint, the non-adaptive weights of the next subsection represent a simple starting point.

4.2 Single-observation weights

The preset weights of Section 2.2.1 will now be extended to allow each comparison to have a different sample size. Therein, the concept of a single-observation weight is heuristically motivated by starting with a single pseudo-observation y_0.

More generally, *single-observation weights* are the components of w_i such that (i) for every $i \in 1, \ldots, N$, all incidental data (all x_j with $j \neq i$) together have the weight of one observation in the focus vector x_i $\left(\sum_{j \neq i} w_{ij} = w_{ii}/n_i\right)$ and (ii) each comparison other than the ith has equal weight $(\forall j, k \neq i \, w_{ij} = w_{ik})$.

Solving those equations and $\sum_{j=1}^{N} w_{ij} = 1$ uniquely gives $w_{ii} = 1 - (n_i + 1)^{-1}$ and $i \neq j \implies w_{ij} = (n_i + 1)^{-1} (N - 1)^{-1}$. If there is only a single comparison, then its observed statistic $t_1 = \tau(x_1)$ is supplemented by a pseudo-statistic t_0, a scientifically meaningful value

in $\mathcal{T}_1$ that does not depend on x_1. For example, t_0 might be the mode ($\arg\max_{t\in\mathcal{T}_1} g_{\theta_0}(t)$) or the expectation value $\left(\int t g_{\theta_0}(t)\,dt\right)$ of T_1 under $\theta = \theta_0$. (Similarly, Kass and Wasserman (1995) considered the use of a prior with the Fisher information of a single pseudo-observation.) The use of $\mathbf{t} = \langle t_0, t_1\rangle$ and $N = 2$ with single-observation weights then entails that

$$\log \bar{L}_1(\theta_1; \mathbf{t}) = (n_1+1)^{-1} \log L_1(\theta_1; t_0) + \left(1 - (n_1+1)^{-1}\right) \log L_1(\theta_1; t_1). \tag{16}$$

For a smoother transition from a single comparison to multiple comparisons, the pseudo-statistic may be assigned the same weight as each of the $N-1$ incidental statistics among $t_1, \ldots, t_N$, i.e., $w_{ij} = (n_i+1)^{-1} N^{-1}$ for all $j \in \{0, 1, \ldots, N\} \setminus \{i\}$.

The following corollary indicates that the use of single-observation weights is sufficient for application of the practical approximation made in Section 3.2.2. The result applies whether there is a single comparison or multiple comparisons.

Corollary 10. *Assume the components of w_i are single-observation weights, that $n_i = n_1$ for all $i = 1, \ldots, N$, and that $\mathcal{G}_{n_1,i} = \mathcal{G}_{n_1,1}$ for all $i = 1, \ldots, N$. If $T_1, \ldots, T_N$ are independent and drawn from a mixture distribution, if $\hat{\theta}\left(\tilde{\mathbf{T}}(t)\right)$ and $\hat{\theta}_i(\mathbf{T}_i(t))$ are almost always unique for all $t \in \mathcal{T}_i$, and if $L_i(\bullet; T_i)$ is almost surely continuous on Θ for all $i = 1, \ldots, N$, then equation (10) holds.*

Proof. All the conditions of Theorem 7 are given except for the equal weights condition, which follows from the single-observation weights assumption, the equality of the sample sizes, and the commonality of the family of distributions. □

5 Case studies

In the models used in the applications of this section, $\int f_{\hat{\phi}(x)}(x)\,dx = \infty$, rendering the unmodified NML (4) useless. The NMWL (8) can be used instead since numerical integration reveals that $\int_{\mathcal{T}_i} \bar{L}_i\left(\hat{\theta}_i(\mathbf{t}_i(t)); \mathbf{t}_i(t)\right) dt < \infty$ in all cases considered. Thus, the applications will motivate the proposed generalization of NML.

Results of two separate NMWL analyses with single observation weights (§4.2) are presented for each application. In the terminology of Section 3.2.2, the first analysis uses multiple incidental comparisons for inference relevant to each comparison in focus (6). The second analysis uses $t_0 = 0$ for each focus comparison in place of data associated with the other comparisons, as if it were the only comparison made. Thus, for the second analysis of the ith of N focus comparisons, equation (16) is replaced with

$$\log \bar{L}_i(\theta_i; \mathbf{t}) = (n_i + 1)^{-1} \log L_i(\theta_i; 0) + \left(1 - (n_i + 1)^{-1}\right) \log L_i(\theta_i; t_i) \tag{17}$$

for $i = 1, \ldots, N$.

All plots use the binary logarithm to express information in bits and display a different value for each comparison in focus.

5.1 Single and multiple populations

Before addressing a problem in contemporary biology, the proposed methodology will be illustrated using a simple data set that has motivated both Bayesian (Rubin, 1981) and weighted likelihood (Wang, 2006a) approaches. The reduced data consist of the estimated average effect of a training program on SAT scores and an estimated standard error of the effect estimate for each of eight test sites. Following the tradition continued by Wang (2006a), the standard errors $\sigma_1, \ldots, \sigma_8$ are considered known, and the effect estimates $\hat{\mu}_1, \ldots, \hat{\mu}_8$ are modeled as normal observations with unknown means $\mu_1, \ldots, \mu_8$. Thus, estimated average effects are naturally reduced to their standardized values, $t_1 = \hat{\mu}_8/\sigma_8, \ldots, t_8 = \hat{\mu}_8/\sigma_8$. In the notation of Section 2.2.1, $N = 8$ and $\{g_\theta : \theta \in \Theta\}$ is the family of distributions common to each test site, where g_θ is the normal density of mean θ and standard deviation 1. (There is no nuisance parameter.) The parameters of interest, $\theta_1 = \mu_1/\sigma_1, \ldots, \theta_8 = \mu_8/\sigma_8$, are the inverse coefficients of variation; $\theta_i \neq 0$ is the alternative hypothesis that the training program affected the mean score at the ith site, and $\theta_i = 0$ is the null hypothesis that it did not.

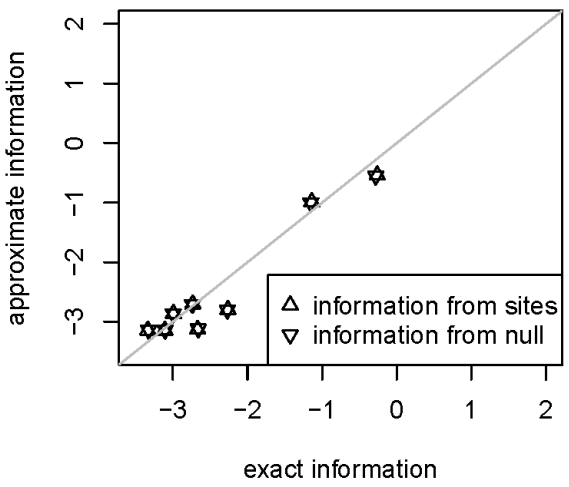

Figure 1: Information (bits) favoring the hypothesis that the test score at a site was affected by the treatment.

Fig. 1 displays $\tilde{I}_i(\mathbb{R}\backslash\{0\},\{0\})$, the resulting approximate discrimination information (12), with $\bar{I}_i(\mathbb{R}\backslash\{0\},\{0\})$, the exact discrimination information (11). As in Section 4.2, the weight of a single observation is assigned either to 0, the null hypothesis value ("information from null"), or to all of the incidental testing sites ("information from sites"). The resulting information values are barely distinguishable.

Fig. 1 indicates that there is evidence favoring an effect of the training program $\left(\bar{I}_i(\mathbb{R}\backslash\{0\},\{0\}) > 0\right)$ at only one exam site. Table 1 classifies the strength of the evidence at that site as weak $\left(1 \leq \bar{I}_i(\mathbb{R}\backslash\{0\},\{0\}) < 2\right)$. All other sites favor their null hypotheses of no effect, and three do so with strong evidence $\left(-5 < \bar{I}_i(\mathbb{R}\backslash\{0\},\{0\}) \leq -3\right)$ according to Table 1. Section 5.2.3 will explain a complementary perspective on interpreting levels of evidence.

5.2 Single and multiple biological features

5.2.1 Differential protein abundance

In typical experiments measuring gene expression or the abundance of proteins or metabolites, the primary question is whether the expectation value of a logarithm of the expression or abundance of each feature is affected by a treatment, disease, or other perturbation. That

question is equivalent to that of whether $\left|\mathrm{CV}_i^{-1}\right| = |1/\mathrm{CV}_i|$ is 0, where CV_i is the coefficient of variation for the ith feature. Thus, the data reduction strategy of Example 2 often proves effective even if the magnitude of $\left|\mathrm{CV}_i^{-1}\right|$ is not of direct interest. Under the model of that example, CV_i has a one-to-one correspondence to the proportion of the feature-feature pairs with abundance ratios greater than 1 (Bickel, 2004, 2008). In addition, $\left|\mathrm{CV}_i^{-1}\right|$ is often of more scientific interest than the mean since small changes in numbers of biomolecules can have a strong influence on downstream processes.

The method of Example 2 is applied to the proteomics data set of Alex Miron's lab at the Dana-Farber Cancer Institute (Li, 2009), with x_{ij} and y_{ij} as the logarithms of the abundance levels of the ith of $N = 20$ proteins in the jth woman with and without breast cancer, respectively, after adding the first quartile of the abundance levels (over the 64 healthy women and over all proteins) to each abundance level (Bickel, 2010a). Likewise, ξ_i and η_i are the expectation values of the random variables X_{ij} and Y_{ij}. Each of two breast cancer groups (one of 55 HER2-positive women and the other of 35 women mostly-ER/PR-positive) was compared to a control group of 64 women. Since $\theta_i = \left|\mathrm{CV}_i^{-1}\right|$ and thus $\Theta = [0, \infty)$, the competing hypotheses for the ith protein are $\theta_i > 0$ and $\theta_i = 0$.

5.2.2 Application of discrimination information

The methodology was first applied to the HER2-control comparison (Fig. 2). The left panel displays the approximate information for discrimination (12) in favor of the alternative hypothesis that $\theta_i \neq 0$ over the null hypothesis that $\theta_i = 0$ by weighing the incidental proteins as a single observation (§4.2). $\tilde{I}_i\left((0, \infty), \{0\}\right)$, the approximate minimax information (12), is compared to $\log\left(g_i\left(\hat{\theta}_{\mathrm{MLE}}; t_i\right) / g_i\left(0; t_i\right)\right)$. Here, $\hat{\theta}_{\mathrm{MLE}}$ is common to all proteins, denoting a numeric approximation to the maximum likelihood estimate (MLE) of the $\theta_{\mathrm{alt.}} > 0$ defined under the assumptions that $\theta_i \in \{0, \theta_{\mathrm{alt.}}\}$ for all i and that the test statistics are independent (Bickel, 2010a). The right panel of Fig. 2 contrasts the widely varying regret of the MLE information with the constant regret of the minimax information.

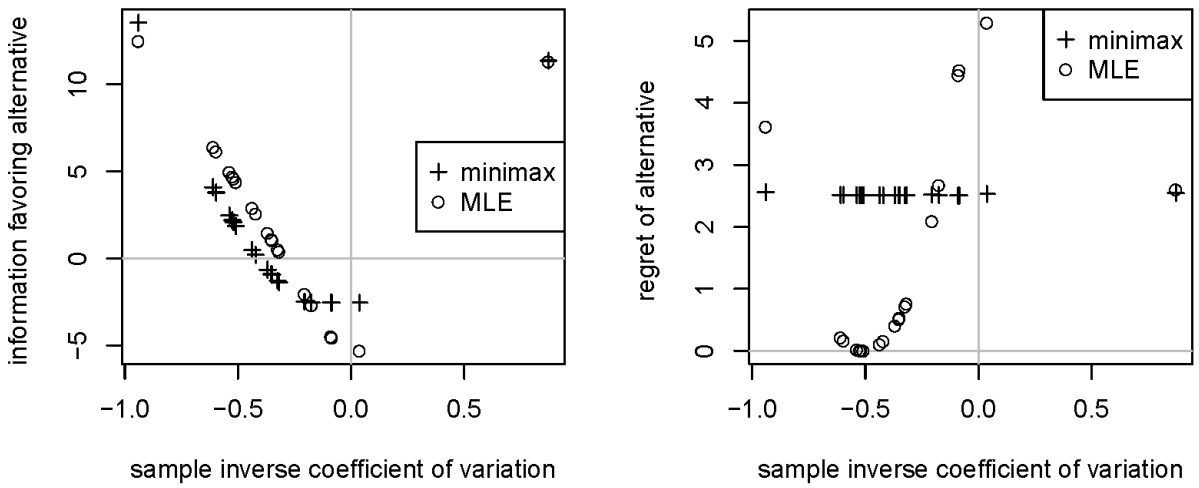

Figure 2: *Left panel:* Discrimination information (bits) favoring the hypothesis that the abundance level of a protein differs by disease status versus $\widehat{\mathrm{CV}}_i^{-1}$. *Right panel:* The corresponding regret versus $\widehat{\mathrm{CV}}_i^{-1}$. ($\widehat{\mathrm{CV}}_i^{-1}$ denotes the difference in sample means divided by the sample standard deviation for the ith protein.)

Giving the null hypothesis the weight of a single observation (17) at $t_0 = 0$, as if the abundance level of only one protein were measured, results in information values that are visually indistinguishable from those of Fig. 2. Nonetheless, some weight dependence is perceptible for much smaller sample sizes. For example, Fig. 3 displays the effect of using the null hypothesis weights instead of the protein weights on $\tilde{I}_i((0,\infty),\{0\})$ for two randomly selected patients from each breast cancer group and from the healthy group. Even in this extreme case, the evidence grade (Table 1) changed with the weighting method for no proteins of the HER2 group (left) and for only one out of 20 proteins of the ER/PR group (right).

5.2.3 Biological interpretation

The results will now be interpreted biologically. Each + of the left-hand side of Figure 2 records the amount of discrimination information favoring the hypothesis of differential abundance of a protein between the cancer and control groups. A protein with a positive discrimination information value exhibits evidence in favor of its differential abundance, whereas one with a negative discrimination information value exhibits evidence in favor of its

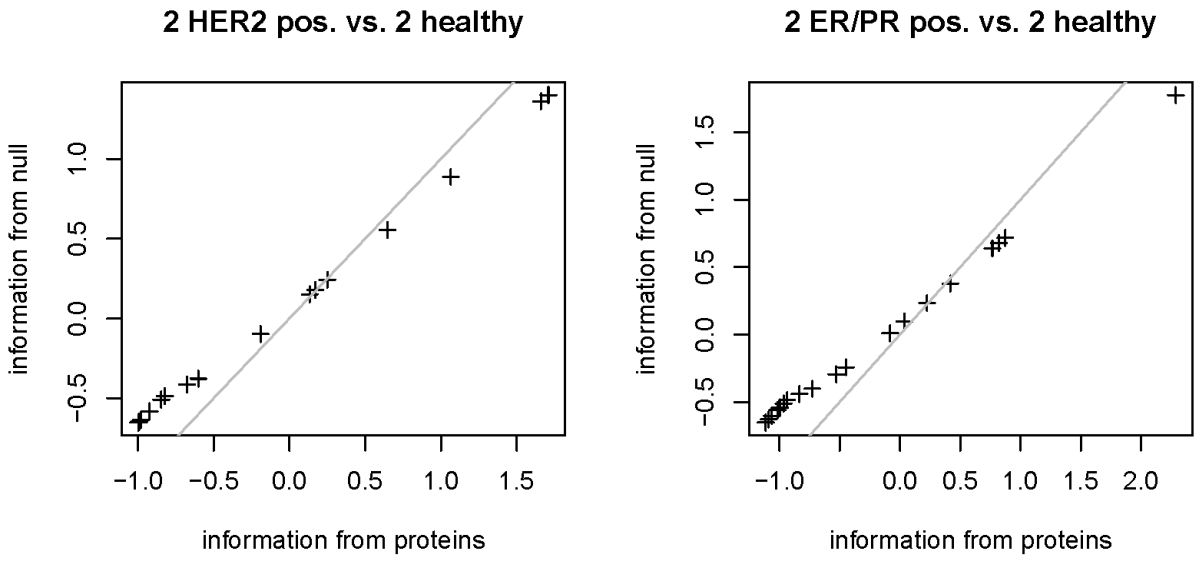

Figure 3: Discrimination information (bits) favoring the hypothesis that the abundance level of a protein differs by disease status ($n = 2$ women per group) using weights from the null hypothesis ($t_0 = 0$) versus that using weights from the incidental proteins. The randomly drawn women are labeled I72 and I14 from the HER2 group; II28 and I23 from the ER/PR group; I67 and I63 from the healthy group.

equivalent abundance between the cancer and control groups. The strength of that evidence may be qualitatively described using the grades of Table 1. Thus, for example, no proteins attain at least strong evidence of having equivalent abundance $\tilde{I}_i\left((0,\infty),\{0\}\right) \leq -3$, but two proteins attain at least strong evidence of differential abundance $\tilde{I}_i\left((0,\infty),\{0\}\right) \geq 3$.

Further, one may use Bayes's theorem to combine the discrimination information with $\hat{p}_1$, an estimated or hypothetical proportion p_1 of proteins with differential abundance or, more generally, the proportion of the N alternatives hypotheses that are true (Bickel, 2011a). Since the true prior probability p_1 refers to an unknown true value determined by protein abundance levels, it should not be confused with the prior level of belief used in pure Bayesian decision theory. The logarithm of the estimated or hypothetical posterior odds that the ith protein has differential abundance is

$$\log\left(\frac{\hat{p}_1}{1-\hat{p}_1}\right) + \tilde{I}_i\left((0,\infty),\{0\}\right), \tag{18}$$

yielding the estimated or hypothetical posterior probability of differential abundance. From this viewpoint, the discrimination information is interpreted as an approximation to the change in log odds induced by the data (Bickel, 2011a). For example, if $\hat{p}_1 = 5\% \approx p_1$, then the protein of discrimination information value $\tilde{I}_i((0,\infty),\{0\}) = 4.1$ bits has an approximate posterior log odds of -4.2 bits $+ 4.1$ bits $= -0.1$ bits according to equation (18). Thus, even though that protein has strong evidence for differential abundance, the posterior probability that it has differential abundance would only be about 50% if indeed only 5% of the proteins have differential abundance. On the other hand, the protein of $\tilde{I}_i((0,\infty),\{0\}) = 11.4$ bits has an approximate posterior log odds of -4.2 bits $+ 11.4$ bits $= 7.2$ bits, which corresponds to a posterior probability of over 99%.

6 Discussion

The main contributions of this paper have been summarized in Section 1.2. The minimax method of quantifying evidence introduced informally in Section 2.2 was generalized in Sections 3-4 and illustrated in Section 5. This section will highlight advantages of the proposed minimax discrimination information over previous evidence measures in addition to the advantages already noted in Sections 1.1, 2.1, and 2.2.

The minimax discrimination information was designed to integrate data of each comparison currently in focus with any available incidental data, i.e., data associated with comparisons other than the comparison currently in focus. In both of the case studies of Section 5, the use of incidental data in place of t_0, an artificial incidental data point that is determined by the null hypothesis, has little effect on the minimax information for discrimination. In the second application (§5.2), the amount of information lost for inference about a single protein in the absence of the other 19 proteins was negligible. Even when the sample size was reduced to $n_i = 4$, the effect of eliminating the incidental data was very small in 39 of 40 $\langle$protein, cancer type$\rangle$ combinations. Thus, the minimax discrimination information

with single-observation weights (§4.2) is robust against changes in the incidental data. Even so, giving the incidental data the weight of a single observation was sufficient to render the normalization constant finite, thereby removing the primary obstacle to applying the NML to statistical data analysis (§2.2).

The insensitivity to the use of incidental data also suggests that the new minimax solution to the incidental-data issue raised in the Section 2.2 is a measure of evidence that has the same interpretation for any number of comparisons. By contrast, p-values adjusted to control error rates tend to vary so greatly between a single comparison and a large number of comparisons that they require researchers to separately build the intuition needed to interpret statistical reports for small numbers of comparisons, medium numbers of comparisons, and large numbers of comparisons (cf. Bickel, 2011b). This shortcoming of traditional approaches to the multiple comparisons problem is especially glaring when the same article reports various degrees of adjusting p-values for data types involving very different numbers of features.

Appendix A: The coding analogy

This appendix has two purposes. First, it is intended to facilitate relating the minimum description length (MDL) literature to the framework of the paper. Second, the appendix explains why *information for discrimination* is an apt term for the $\bar{I}_i\left(\Theta_1, \Theta_0\right)$ defined in Section 3.3. The appendix is not needed to comprehend or implement the methodology of the main text.

A *coding scheme* is an algorithm that faithfully records or describes a message by the assignment of digits or other letters. The string of letters used to record the message is called a *code* or a compressed message. The *codelength* or "description length" is the number of letters in the code. The codelength should be kept to a minimum since the number of letters in the code directly impacts the cost of storing it on a retrieval device such as a

hard disk or sending it across a communications channel such as a telephone line. Thus, just as the performance of a predictive distribution is judged by its error in predicting new data, the performance of a coding scheme is judged by the codelength of a message that was not used in the design of the coding scheme. In fact, the best coding schemes make use of distributions designed to predict new messages since the expected codelength may be minimized by coding schemes that assign shorter codes to more common messages and longer codes to rarer messages. Thus, under some practical constraints and an idealization ignoring the integer nature of literal codelengths, there is a one-to-one correspondence between $p(x)$, the relative frequency of message x, and ℓ, its optimal codelength:

$$\ell(x) = \log(1/p(x)) = -\log p(x). \tag{19}$$

The correspondence may be derived from the Kraft inequality and the information inequality (Cover and Thomas, 2006; Rissanen, 2007; Grünwald, 2007).

Were the probability mass function p known, one would design an optimal coding scheme with codelengths given by equation (19). When it is unknown but assumed to belong to a parametric family of probably mass functions, a coding scheme is devised such that the mean difference per letter between its codelength and $\ell(x)$ asymptotically vanishes. Such a coding scheme is called *universal* since the asymptotic convergence holds for any message-generating mass function in the parametric family (Rissanen, 2007; Grünwald, 2007; Bickel, 2011a). Since universality is a weak condition akin to the consistency of an estimator, various optimality criteria have been imposed to narrow down the number of qualifying coding schemes. For example, Shtarkov (1987) introduced a unique coding scheme by requiring it to minimize the codelength given the worst possible message x and the worst possible mass function p.

The MDL research program launched by Jorma Rissanen (Rissanen, 2007; Grünwald, 2007) identifies the problem of recording a message with the problem of selecting a family of

Universal coding	Model selection
uncompressed message	sample of data
family of message-generating distributions	family of sampling distributions
(universal) coding scheme	predictive probability distribution
codelength or description length	logarithmic prediction error

Table 2: MDL metaphors.

probability distributions on the basis of a sample of data (Table 2). The concepts of Section 2 are metaphorically translated to those of universal coding as follows. The error $-\log \bar{f}(x)$ of the observed sample x incurred by the worst-sample minimax predictive density function $\bar{f}$ is the codelength of the message x as recorded by the worst-message minimax coding scheme of Shtarkov (1987). Likewise, the error $-\log \bar{f}(x; y_0, \Phi)$ incurred by the density function $\bar{f}(\bullet; y_0, \Phi)$ is the codelength of x. The coding scheme corresponding to $\bar{f}(\bullet; y_0, \Phi)$ depends on the incidental message y_0 but not on any part of the message x to be recorded.

In the more general notation of Section 3.3, $-\log \bar{g}_i(t_i; \Theta_0)$ is the codelength of using the null hypothesis that $\theta_i \in \Theta_0$ to record the message t_i, whereas $-\log \bar{g}_i(t_i; \Theta_1)$ is the codelength of using the alternative hypothesis that $\theta_i \in \Theta_1$ to do so. Thus, according to equation (11), using the alternative hypothesis for the coding scheme in place of the null hypothesis reduces the codelength by about $\bar{I}_i(\Theta_1, \Theta_0)$ letters if $\bar{I}_i(\Theta_1, \Theta_0) \geq 0$. In that sense, $\bar{I}_i(\Theta_1, \Theta_0)$ may be regarded as the number of letters recording the statistic t_i that favor the use of the alternative hypothesis over the null hypothesis. Identifying the number of letters with the amount of information justifies calling $\bar{I}_i(\Theta_1, \Theta_0)$ the amount of information in t_i discriminating between the two hypotheses in favor of the alternative hypothesis over the null hypothesis. Likewise, $\bar{I}_i(\Theta_0, \Theta_1) = -\bar{I}_i(\Theta_1, \Theta_0)$ is the number of letters recording the statistic t_i that favor the use of the null hypothesis over the alternative hypothesis if $\bar{I}_i(\Theta_0, \Theta_1) \geq 0$.

Acknowledgments

Mayer Alvo kindly provided many useful suggestions, especially regarding the organization and focus of the paper. I also thank Corey Yanofsky and Dejian Lai for helpful comments on the manuscript. Finally, I am very grateful to two anonymous reviewers, the Associate Editor, and the Editor-In-Chief for remarks that led to much clearer presentation.

`Biobase` (Gentleman et al., 2004) facilitated data management. This research was partially supported by the Canada Foundation for Innovation, by the Ministry of Research and Innovation of Ontario, and by the Faculty of Medicine of the University of Ottawa.

References

Aitkin, M., 2010. Statistical Inference: An Integrated Bayesian/Likelihood Approach (Chapman & Hall/CRC Monographs on Statistics & Applied Probability), 1st Edition. Chapman and Hall/CRC.

Berger, T., Pericchi, L., 2004. Training samples in objective Bayesian model selection. Annals of Statistics 32 (3), 841–869.

Bickel, D. R., 2004. Degrees of differential gene expression: Detecting biologically significant expression differences and estimating their magnitudes. Bioinformatics (Oxford, England) 20, 682–688.

Bickel, D. R., 2008. Correcting the estimated level of differential expression for gene selection bias: Application to a microarray study. Statistical Applications in Genetics and Molecular Biology 7 (1), 10.

Bickel, D. R., 2009. A frequentist framework of inductive reasoning. Technical Report, Ottawa Institute of Systems Biology, arXiv:math.ST/0602377.

Bickel, D. R., 2010a. Minimum description length methods of medium-scale simultaneous inference. Technical Report, Ottawa Institute of Systems Biology, arXiv:1009.5981.

Bickel, D. R., 2010b. The strength of statistical evidence for composite hypotheses: Inference to the best explanation. Technical Report, Ottawa Institute of Systems Biology, COBRA Preprint Series, Article 71, available at biostats.bepress.com/cobra/ps/art71.

Bickel, D. R., 2011a. Measuring support for a hypothesis about a random parameter without estimating its unknown prior. Technical Report, Ottawa Institute of Systems Biology, arXiv:1101.0305.

Bickel, D. R., 2011b. Small-scale inference: Empirical bayes and confidence methods for as few as a single comparison. Technical Report, Ottawa Institute of Systems Biology, arXiv:1104.0341.

Blume, J., Peipert, J., 2003. What your statistician never told you about p-values. Journal of the American Association of Gynecologic Laparoscopists 10 (4), 439–444.

Cover, T., Thomas, J., 2006. Elements of Information Theory. John Wiley and Sons, New York.

Edwards, A. W. F., 1992. Likelihood. Johns Hopkins Press, Baltimore.

Efron, B., 1986. Why Isn't Everyone A Bayesian. American Statistician 40 (1), 1–5.

Efron, B., 2010. The future of indirect evidence. Statistical Science 25 (2), 145–157.

Efron, B., Gous, A., 2001. Scales of evidence for model selection: Fisher versus Jeffreys. Lecture Notes - Monograph Series 38, 208–256.

Fisher, R. A., 1973. Statistical Methods and Scientific Inference. Hafner Press, New York.

Gentleman, R. C., Carey, V. J., Bates, D. M., et al., 2004. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology 5, R80.

Greenland, S., 2006. Bayesian perspectives for epidemiological research: I. Foundations and basic methods. International Journal of Epidemiology 35 (3), 765–775.

Grünwald, P. D., 2007. The Minimum Description Length Principle. The MIT Press, London.

Hu, F. F., Zidek, J., 2002. The weighted likelihood. Canadian Journal of Statistics 30 (3), 347–371.

Jeffreys, H., 1948. Theory of Probability. Oxford University Press, London.

Kass, R. E., Raftery, A. E., 1995. Bayes factors. Journal of the American Statistical Association 90 (430), 773–795.

Kass, R. E., Wasserman, L., 1995. A reference Bayesian test for nested hypotheses and its relationship to the schwarz criterion. Journal of the American Statistical Association 90 (431), 928–934.

Kullback, S., 1968. Information Theory and Statistics. Dover, New York.

Lanterman, A. D., 2005. Advances in Minimum Description Length: Theory and Applications. The MIT Press, London, Ch. Hypothesis testing for Poisson versus geometric distributions using stochastic complexity, pp. 23–79.

Lele, S. R., 2004. Elicit data, not prior: On using expert opinion in ecological studies. The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations. University of Chicago Press, Chicago, pp. 410–436.

Li, X., 2009. ProData. Bioconductor.org documentation for the ProData package.

Plante, J.-F., 2008. Nonparametric adaptive likelihood weights. Canadian Journal of Statistics 36 (3), 443–461.

Plante, J.-F., 2009. Asymptotic properties of the mamse adaptive likelihood weights. Journal of Statistical Planning and Inference 139 (7), 2147–2161.

Rissanen, J., 1987. Stochastic complexity. Journal of the Royal Statistical Society.Series B (Methodological) 49 (3), 223–239.

Rissanen, J., 2007. Information and Complexity in Statistical Modeling. Springer, New York.

Rissanen, J., Roos, T., 2007. Conditional NML universal models. pp. 337–341.

Rissanen, J. J., 1996. Fisher information and stochastic complexity. IEEE Transactions on Information Theory 42 (1), 40–47.

Royall, R., 1997. Statistical Evidence: A Likelihood Paradigm. CRC Press, New York.

Royall, R., 2000. On the probability of observing misleading statistical evidence. Journal of the American Statistical Association 95 (451), 760–768.

Rubin, D. B., 1981. Estimation in parallel randomized experiments. Journal of Educational Statistics 6 (4), pp. 377–401.

Serfling, R. J., 1980. Approximation theorems of mathematical statistics. Wiley, New York.

Severini, T., 2000. Likelihood Methods in Statistics. Oxford University Press, Oxford.

Shtarkov, Y. M., 1987. Universal sequential coding of single messages. Problems of information transmission 23 (3), 175–186.

Sin, C.-Y., White, H., 1996. Information criteria for selecting possibly misspecified parametric models. Journal of Econometrics 71 (1-2), 207–225.

Takimoto, E., Warmuth, M. K., 2000. The last-step minimax algorithm. In: ALT '00: Proceedings of the 11th International Conference on Algorithmic Learning Theory. Springer-Verlag, London, UK, pp. 279–290.

Wang, X., 2006a. Approximating Bayesian inference by weighted likelihood. Canadian Journal of Statistics 34 (2), 279–298.

Wang, X., 2006b. Asymptotic properties of adaptive likelihood weights by cross-validation. Communications in Statistics - Theory and Methods 35 (7), 1257–1270.

Wang, X., Zidek, J. V., 2005a. Derivation of mixture distributions and weighted likelihood function as minimizers of KL-divergence subject to constraints. Annals of the Institute of Statistical Mathematics 57 (4), 687–701.

Wang, X., Zidek, J. V., 2005b. Selecting likelihood weights by cross-validation. Annals of Statistics 33 (2), 463–500.