

Minimum description length methods of medium-scale simultaneous inference

David R. Bickel

3 September 2010

See [arXiv:1009.5981](https://arxiv.org/abs/1009.5981) for revisions of this paper.

I. INTRODUCTION

Ottawa Institute of Systems Biology
 Department of Biochemistry, Microbiology, and Immunology
 Department of Mathematics and Statistics
 University of Ottawa
 451 Smyth Rd.
 Ottawa, Ontario K1H 8M5
 dbickel@uottawa.ca

Abstract—Nonparametric statistical methods developed for analyzing data for high numbers of genes, SNPs, or other biological features tend to overfit data with smaller numbers of features such as proteins, metabolites, or, when expression is measured with conventional instruments, genes. For this medium-scale inference problem, the minimum description length (MDL) framework quantifies the amount of information in the data supporting a null or alternative hypothesis for each feature in terms of parametric model selection. Two new MDL techniques are proposed. First, using test statistics that are highly informative about the parameter of interest, the data are reduced to a single statistic per feature. This simplifying step is already implicit in conventional hypothesis testing and has been found effective in empirical Bayes applications to genomics data. Second, the codelength difference between the alternative and null hypotheses of any given feature can take advantage of information in the measurements from all other features by using those measurements to find the overall code of minimum length summed over those features. The techniques are applied to protein abundance data, demonstrating that a computationally efficient approximation that is close for a sufficiently large number of features works well even when the number of features is as low as 20. More generally, the MDL-based information for discrimination does not suffer from the asymmetry of the p -value as a measure of evidence for one hypothesis over another.

Keywords: information criteria; information for discrimination; minimum description length; model selection; reduced likelihood; universal source coding

The author thanks Ye Yang and Zhengmin Zhang for relevant discussions, Zhenyu Yang for comments on the data reduction section, and Corey Yanofsky for both. This work was partially supported by the Faculty of Medicine of the University of Ottawa.

A draft of this technical report became available from www.medicine.uottawa.ca on 14 August 2010. Condensed versions were presented at the 2010 Workshop on Information Theoretic Methods in Science and Engineering in Tampere, Finland on 16 August 2010 and to the Faculty of Agricultural Science of the University of Aarhus on 19 August 2010. This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

In high-dimensional biology, there are one or more measurements for each of thousands or even hundreds of thousands of biological features such as genes, locations in the brain, and, in genome-wide association studies, single-nucleotide polymorphisms (SNPs). For the interpretation of such data, statistical methods of large-scale inference have enjoyed a phase of rapid development over the last decade.

In particular, recent empirical Bayes methodology tests a null hypothesis for each of N features by making use of information in the measurements of the other features. For measurement vectors x_1, \dots, x_N modeled as realizations of the random variables X_1, \dots, X_N and for test statistics $T(X_1), \dots, T(X_N)$ corresponding to null hypotheses $\theta_1 = \theta_0, \dots, \theta_N = \theta_0$, under which each of the test statistics has a common probability density function g_{θ_0} , the local false discovery rate (LFDR) for the i th feature is the posterior probability

$$\begin{aligned} \text{LFDR}(x_i) &= P(\theta_i = \theta_0 | T(x_i)) \\ &= \frac{P(\theta_i = \theta_0) g_{\theta_0}(T(x_i))}{g(T(x_i))}, \end{aligned} \quad (1)$$

where $\pi_0 = P(\theta_i = \theta_0)$ is the proportion of null hypotheses that are true and where

$$g = \pi_0 g_{\theta_0} + (1 - \pi_0) g_{\text{alt}}. \quad (2)$$

is a mixture density with g_{alt} as the probability density function of the test statistic marginal over all alternative features, those for which the null hypothesis $\theta_i = \theta_0$ is false. As π_0 and g are unknown, they are estimated by $\hat{\pi}_0$ and \hat{g} according to empirical Bayes methodology to obtain $\widehat{\text{LFDR}}(x_i)$, the estimated LFDR. Commonly used methods of estimating π_0 and g involve the nonparametric estimation of g_{alt} . [e.g., 1, 2]. The empirical success of such methods [3, 4] is largely due to the reliability of nonparametric density estimation in the presence of several thousands of features.

Such dependence on nonparametric estimation makes the methods of large-scale inference less applicable to problems involving more moderate dimensions. Like high-dimensional biology, medium-dimensional biology involves measurements over multiple features but on a scale of tens to hundreds of genes, proteins, metabolites, or other features rather than thousands of features. Thus, special statistical methods are needed when the number of features is too large for the mere iteration of conventional hypothesis testing and yet too small for the reliable

use of methods developed for extremely large numbers of features. The situation parallels one in classical mechanics: there are exact equations of motion for both sufficiently small and sufficiently large numbers of bodies, but special approximations are needed for medium numbers of bodies. Information theory may play an important role in solving medium-scale inference problems.

Section II addresses technical aspects of eliminating a nuisance parameter by reducing the data of each feature to a statistic of smaller dimension. Following [5, pp. 4-7], Section III presents the log-likelihood ratio

$$\Delta_i(T(x_i)) = \log \left(\frac{g_{\theta_0}(T(x_i))}{g_{\theta_i}(T(x_i))} \right)$$

as the information in $T(x_i)$ for discrimination favoring the hypothesis that $\theta = \theta_0$ over the hypothesis that $\theta = \theta_1$ for some $\theta_i \neq \theta_0$. Since θ_i is unknown, Section III replaces it with a parameter value chosen to minimize the codelength of the data according to the minimum description length (MDL) interpretation of Shannon-Fano coding theory, in which the length of a codeword is the number of independently selected binary digits of equal probability that achieve the joint probability of that codeword [6]. (See [6, 7, 8] for introductions to the MDL principle of model selection.) The availability of measurements pertaining to features other than the inference target enables the construction of a universal codelength function and a close approximation that is more computationally convenient. The theory is then illustrated in Section IV with a medium-dimensional biological data set. Section V concludes the paper with a discussion of the general applicability of the methods proposed.

II. DATA REDUCTION AND LIKELIHOOD

The observed data vector $x \in \mathcal{X}$ is considered a realization of the random variable X of probability distribution $P_{\theta, \lambda}$ that admits a probability density function $f_{\theta, \lambda}$ with respect to some dominating measure, where $\theta \in \Theta$ is the parameter of interest and $\lambda \in \Lambda$ is the nuisance parameter. In the case of discrete X , the density function is defined with respect to the counting measure on \mathcal{X} . For some known $\theta_0 \in \Theta$, we have $\theta = \theta_0$ under the null hypothesis or narrow model and $\theta \neq \theta_0$ under the alternative hypothesis or wide model.

The following two types of likelihood correspond to different ways of reducing a vector x to a scalar statistic and of eliminating the nuisance parameter. Which of the two methods is appropriate depends on the original parametric family $\{f_{\theta, \lambda} : \theta \in \Theta, \lambda \in \Lambda\}$ and on which parameter is of interest.

A. Conditional likelihood

Consider the functions S and T such that $S(X)$ and $T(X)$ are statistics that together contain all the information in X . If $S(X)$ does not depend on θ and if the probability density function $g_\theta = f_\theta(\bullet | S(X) = S(x))$ of the data conditional on $S(x)$, the realized value of that statistic, does not depend on λ , then the function ℓ defined by

$$\ell(\theta) = g_\theta(T(x)) = f_\theta(T(x) | S(X) = S(x))$$

is called the *conditional likelihood function* given $S(x)$. In analogy with equation (3), [9, §8.2.1] has

$$\begin{aligned} f_{\theta, \lambda}(x) &= f_{\theta, \lambda}(S(x), T(x)) \\ &= g_\theta(T(x)) f_{\theta, \lambda}(S(x)), \end{aligned}$$

where $f_{\theta, \lambda}$ can denote the probably density function of X , $\langle S(X), T(X) \rangle$, or $S(X)$, depending on the context.

Example 1. [9, Example 8.47]. Suppose X_1 is binomial $\langle n_1, \pi_1 \rangle$, that X_2 is binomial $\langle n_2, \pi_2 \rangle$, and that X_1 is independent of X_2 . The parameter of interest is

$$\theta = \log \frac{\pi_1}{1 - \pi_1} - \lambda,$$

where λ is the nuisance parameter

$$\lambda = \log \frac{\pi_2}{1 - \pi_2}.$$

Then $\log L(\theta, \lambda) = x_1 \theta + S(x_1, x_2) \lambda - n_1 \log(1 + e^{\theta + \lambda}) - n_2 \log(1 + e^\lambda)$, where $S(x_1, x_2) = x_1 + x_2 = s$ is sufficient. Then, taking $T(x_1, x_2) = x_1$, the conditional log-likelihood function given $S(x_1, x_2)$ is

$$\log \ell(\theta) = \log g_\theta(x_1) = \theta x_1 - \log K(\theta),$$

where

$$K(\theta) = \sum_{j=\max(0, s-n_2)}^{\min(n_1, s)} \binom{n_1}{j} \binom{n_2}{s-j} e^{j\theta}.$$

Conditional likelihoods are generally available whenever the parameter of interest is a natural parameter of an exponential family [10, §10.3]. For details, see [9, §8.2.4].

B. Marginal likelihood

Let T be a measurable function on \mathcal{X} . If, for each $\theta \in \Theta$, the probability density function g_θ of the *statistic* or *reduced data* $T(X)$ does not depend on the value of λ , then $\ell(\theta) = g_\theta(T(x))$ defines the *marginal likelihood function* ℓ .

If, in addition, the conditional distribution of X given $T(X) = T(x)$ does not depend on θ , then $T(X)$ is called *sufficient* for θ . In that case, no information about θ is lost in replacing X with $T(X)$:

$$\begin{aligned} f_{\theta, \lambda}(x) &= g_\theta(T(x)) f_{\theta, \lambda}(x | T(X) = T(x)) \\ &= g_\theta(T(x)) f_\lambda(x | T(X) = T(x)) \\ &= C g_\theta(T(x)), \end{aligned} \quad (3)$$

where C is constant in θ . The constant is unimportant since it drops out of likelihood ratios:

$$\frac{f_{\theta_1, \lambda}(x)}{f_{\theta_0, \lambda}(x)} = \frac{C g_{\theta_1}(T(x))}{C g_{\theta_0}(T(x))} = \frac{\ell(\theta_1)}{\ell(\theta_0)}$$

for any value of $\lambda \in \Lambda$.

Example 2. Suppose x and y are vectors of m and n values that realize the random variables X and Y of independent components drawn from a normal distributions of unknown means ξ and η , respectively, and of a common unknown

standard deviation σ . The parameter of interest is the inverse coefficient of variation defined by $\theta = (\xi - \eta)/\sigma$ with $\theta = 0$ as the null hypothesis and $\theta \neq 0$ as the alternative hypothesis; the parameter space here is $\Theta = \mathbb{R}^1$. A suitable statistic for data reduction is the two-sample t statistic

$$T(x, y) = \frac{\hat{\xi}(x) - \hat{\eta}(y)}{\hat{\sigma}(x, y) \sqrt{m^{-1} + n^{-1}}}, \quad (4)$$

where $\hat{\xi}$, $\hat{\eta}$, and $\hat{\sigma}^2$ are the usual unbiased estimators. Then $g_\theta(T(x, y))$, the probability density of $T(X, Y)$ evaluated at the observation $\langle x, y \rangle$, is the noncentral Student t probability density with $m + n - 2$ degrees of freedom and noncentrality parameter $(m^{-1} + n^{-1})^{-1/2} \theta$.

The next example encompasses data of medium-dimensional and high-dimensional biology.

Example 3. Example 2 is extended to N genes, proteins, or other biological features such that $X_i \sim N(\xi_i, \Sigma_{i,m})$ and $Y_i \sim N(\eta_i, \Sigma_{i,n})$ correspond to the observed outcome $\langle x_i, y_i \rangle$ for the i th feature, where $i = 1, \dots, N$ and $\Sigma_{i,k}$ is the diagonal covariance matrix of determinant σ_i^{2k} ; thus, σ_i is the standard deviation of independent measurements of feature i . If the number of features with a positive difference ($\xi_i > \eta_i$) is close to the number of features with a negative difference ($\xi_i < \eta_i$), the parameter of interest for feature i may be $\theta_i = |\xi_i - \eta_i|/\sigma_i$, the absolute value of the inverse coefficient of variation, with $\theta_i = 0$ as the null hypothesis, $\theta_i > 0$ as the alternative hypothesis, and $\Theta = [0, \infty)$ as the parameter space. Then $T(x_i, y_i)$ is the absolute value of the two-sample t statistic for $\langle x_i, y_i \rangle$ according to equation (4), and $T(X_i, Y_i)$ is distributed as the absolute value of a variate from the noncentral Student t distribution with $m + n - 2$ degrees of freedom and noncentrality parameter $(m^{-1} + n^{-1})^{-1/2} \theta_i$. Thus, the density $g_\theta(T(x_i, y_i))$ for the i th feature is the probability density of $T(X_i, Y_i)$ evaluated at $\langle x_i, y_i \rangle$.

[9, §8.3] and [11] provide additional examples of the marginal likelihood, also called the *reduced likelihood* and not to be confused with the likelihood integrated with respect to a prior distribution.

III. UNIVERSAL CODING

A. General concepts and notation

The theory of this section is presented in terms of a parametric family that is free of nuisance parameters. In many cases, such a family can be derived using one of the data reduction methods of Section II.

Under the MDL framework, each scheme \dagger for coding the data under the alternative hypothesis corresponds to a codelength function L^\dagger on \mathcal{X} and thus to a *compressing probability density function* g^\dagger selected from the parametric family $\{g_\theta : \theta \in \Theta\}$ before observing $T(x)$, the realized value of the statistic, with the goal of minimizing the codelength $L^\dagger(T(x)) = -\log g^\dagger(T(x))$. Since θ_0 is known, the probability density function of the statistic under the null hypothesis is known to be g_{θ_0} , which compresses the data with respect

to the null model. Accordingly, the codelength function L^0 relative to the null hypothesis is that specified by $L^0(T(x)) = -\log g_{\theta_0}(T(x))$.

Since the base of the logarithm is arbitrary, the inverse logarithm is denoted by $\log^{-1} \bullet$ rather than by $\exp \bullet$ or by 2^\bullet . Plots will follow the base-2 convention.

Suppose, as in Example 3, that there is a vector x_i of measurements for each of N features and that the data are reduced to the statistics $T(x_1), \dots, T(x_N)$. With $L_i^\dagger(T(x_i))$ as the codelength of $T(x_i)$ relative to the alternative hypothesis, $\Delta_i^\dagger(T(x_i)) = L_i^\dagger(T(x_i)) - L^0(T(x_i))$ is the *information in $T(x_i)$ for discrimination* favoring the null hypothesis over the alternative hypothesis. A difference in null and alternative codelengths has been called a “universal test statistic” [12], but that term can cause confusion with $T(X_i)$.

Example 4. If the restriction to a parametric family were relaxed,

$$-\log \frac{\hat{g}_{\text{alt}}(T(x_i))}{g_{\theta_0}(T(x_i))} = -\log \frac{1 - \widehat{\text{LFDR}}(x_i)}{\widehat{\text{LFDR}}(x_i)} + \log \frac{1 - \hat{\pi}_0}{\hat{\pi}_0} \quad (5)$$

would be the information for discrimination according to the empirical Bayes methodology of the Introduction.

The *regret* [8] of the codelength function L_i^\dagger given by $L_i^\dagger(T(x_i)) = -\log g_i^\dagger(T(x_i))$ is

$$\begin{aligned} \text{reg}(g_i^\dagger, x_i) &= L_i^\dagger(T(x_i)) \\ &\quad - \inf_{\theta \in \Theta} (-\log g_\theta(T(x_i))) \\ &= -\log \frac{g_i^\dagger(T(x_i))}{g_{\hat{\theta}}(T(x_i))}, \end{aligned}$$

where $\hat{\theta} = \arg \sup_{\theta \in \Theta} g_\theta(T(x))$. Likewise, the regret of the codelength function relative to the null hypothesis is $\text{reg}(g_{\theta_0}, x_i) = -\log(g_{\theta_0}(T(x_i))/g_{\hat{\theta}}(T(x_i)))$.

While the sign of $\Delta_i^\dagger(T(x_i))$ indicates which hypothesis is favored [12], it can also be compared to a threshold J of the minimum amount of information considered sufficient for selecting one hypothesis over the other. In that case, the probability of observing misleading information for discrimination has an upper bound for any distributions g_{θ_0} and g_i^\dagger . Specifically, for any $J > 0$, $P_{\theta_0, \lambda}(\Delta_i^\dagger(T(X_i)) \leq -J) = P_{\theta_0, \lambda}(g_i^\dagger(T(X_i))/g_{\theta_0}(T(X_i)) \geq \log^{-1} J) \leq 1/\log^{-1} J$. A proof of the inequality and applications to the probability of observing misleading evidence appear in [13]. Since g_{θ_0} and g_i^\dagger must be genuine probability distributions, the inequality does not necessarily hold for pseudo-likelihoods such as profile likelihoods and likelihoods integrated with respect to an improper prior but does hold for all marginal and conditional likelihoods.

The following two schemes (\dagger and \ddagger) for coding the reduced data give essentially identical regrets for sufficiently large N .

B. Exact codelength

While the codelength function L_i^\dagger for the i th feature cannot depend on x_i , it may depend on x_j for all $j \neq i$ as follows. For all $i = 1, \dots, N$, define L_i^\dagger such that the corresponding probability density function g_i^\dagger is equal to $g_{\theta_i^\dagger}$ for the value θ_i^\dagger such that

$$\theta_i^\dagger = \arg \inf_{\theta \in \Theta} \sum_{j \neq i} \min(\text{reg}(g_\theta, x_j), \text{reg}(g_{\theta_0}, x_j)). \quad (6)$$

In words, the code for a given feature uses the distribution in the parametric family that minimizes the regret summed over all other features. This code is considered *exact* since it has the universality property of Section III-D.

Proportional to N^2 , the computation time can prohibit the use of the universal compression method for large N . For example, N can be in the tens of thousands for gene expression microarrays or in the hundreds of thousands for genome-wide association studies. The next coding scheme overcomes this problem since its computational time is proportional to N .

C. Approximate codelength

The \dagger coding scheme is efficiently approximated by a slightly illegal scheme denoted by \ddagger . It determines the codelength for statistic $T(x_i)$ under the alternative hypothesis by use of a common probability density function g^\ddagger that is in the parametric family, i.e., $g^\ddagger = g_{\theta^\ddagger}$ for some $\theta^\ddagger \in \Theta$. This is accomplished by minimizing the regret over all features:

$$\theta^\ddagger = \arg \inf_{\theta \in \Theta} \sum_{j=1}^N \min(\text{reg}(g_\theta, x_j), \text{reg}(g_{\theta_0}, x_j)). \quad (7)$$

This coding scheme is technically illegal in the sense that g^\ddagger , as a function of the observed data for each feature, depends on hindsight. However, under general conditions, θ^\ddagger approximates θ_i^\dagger for all $i = 1, \dots, N$ given sufficiently large N since the selection of the distribution depends on all features without giving undue weight to any single feature. The approximation is supported by the fact that both θ^\dagger and θ^\ddagger are maximum likelihood estimates of θ under the alternative hypothesis:

Theorem 5. *Assume that for some $\theta_0 \in \Theta$ and $\theta_{\text{alt.}} \in \Theta$ such that $\theta_0 \neq \theta_{\text{alt.}}$ and that for all $j \in \{1, \dots, N\}$, each statistic $T(X_j)$ has probability density g_{θ_j} with $\theta_j \in \{\theta_0, \theta_{\text{alt.}}\}$ and is independent of every $T(X_k)$ with $k \in \{1, \dots, N\} \setminus \{j\}$. It follows that θ^\ddagger , if unique, is the maximum likelihood estimate of $\theta_{\text{alt.}}$.*

Proof: By equation (7),

$$\begin{aligned} \theta^\ddagger &= \arg \inf_{\theta} \sum_{j=1}^N \min \left(-\log g_\theta(T(x_j)), -\log g_{\theta_0}(T(x_j)) \right) \\ &= \arg \sup_{\theta \in \Theta} \sum_{j=1}^N \max \left(\log g_\theta(T(x_j)), \log g_{\theta_0}(T(x_j)) \right) \\ &= \arg \sup_{\theta \in \Theta} \sup_{\theta \in \{\theta_0, \theta_{\text{alt.}}\}^N} \sum_{j=1}^N \log g_{\theta_j}(T(x_j)) \\ &= \arg \sup_{\theta \in \Theta} \sup_{\theta \in \{\theta_0, \theta_{\text{alt.}}\}^N} \prod_{j=1}^N g_{\theta_j}(T(x_j)), \end{aligned}$$

where $\theta = \langle \theta_1, \dots, \theta_N \rangle$ and $\{\theta_0, \theta_{\text{alt.}}\}^N$ is the N -factor Cartesian product $\{\theta_0, \theta_{\text{alt.}}\} \times \dots \times \{\theta_0, \theta_{\text{alt.}}\}$. ■

Corollary 6. *Under the assumptions of Theorem 5, $i \in \{1, \dots, N\}$, if θ_i^\dagger is unique, then it is the maximum likelihood estimate of $\theta_{\text{alt.}}$ on the basis of the outcomes $X_j = x_j$ for all $j \in \{1, \dots, N\} \setminus \{i\}$.*

Proof: The claim reduces to that of Theorem 5 since the data are equivalent except for the presence or absence of the outcome $T(X_i) = T(x_i)$ and since θ_i^\dagger and θ^\ddagger are equivalent except for the presence or absence of the term involving that outcome. Thus, for all $i \in \{1, \dots, N\}$,

$$\theta_i^\dagger = \arg \sup_{\theta \in \Theta} \sup_{\theta \in \{\theta_0, \theta_{\text{alt.}}\}^N} \prod_{j \neq i} g_{\theta_j}(T(x_j)). \quad \blacksquare$$

Appendix A specifies sufficient conditions for the convergence of $\theta^\ddagger - \theta_i^\dagger$ to 0 as N increases. Section IV shows that the approximation of \ddagger to \dagger can be quite close even for N as small as 20.

D. Universality of the \dagger code

The coding method of Section III-B is *universal* in the sense that it asymptotically compresses the data as much as the noiseless coding theorem allows for any distribution in the parametric family (cf. [6, §3.7] and [8, §6.5]). Sufficient conditions for universality are stated in the following lemma, in which *strong consistency* means almost sure convergence to a parameter value as $n \rightarrow \infty$ if each $T(X_i)$ is stationary and, at fixed n , of a density function in $\{g_\theta : \theta \in \Theta\}$. (The dependence of g_θ on n is suppressed.) Such convergence will be denoted by \xrightarrow{n} .

Lemma 7 (Consistency). *Suppose that for some $\theta_0 \in \Theta$ and $\theta_{\text{alt.}} \in \Theta$ such that $\theta_0 \neq \theta_{\text{alt.}}$ and that for all $j \in \{1, \dots, N\}$, each statistic $T(X_j)$ has probability density g_{θ_j} with $\theta_j \in \{\theta_0, \theta_{\text{alt.}}\}$ such that $\theta_j = \theta_{\text{alt.}}$ for at least two values of j in $\{1, \dots, N\}$. Suppose further that $g_\bullet(T(X_j))$ is almost surely continuous on Θ for all $j \in \{1, \dots, N\}$. If, for some $i \in \{1, \dots, N\}$, θ_i^\dagger is unique and $\hat{\theta}_j = \arg \sup_{\theta \in \Theta} g_\theta(T(X_j))$ is a strongly consistent estimate of θ_j for all $j \in \{1, \dots, N\} \setminus \{i\}$, then θ_i^\ddagger is a strongly consistent estimate of $\theta_{\text{alt.}}$.*

Proof: Let $\mathfrak{J} = \{j : \theta_j = \theta_{\text{alt.}}, j \in \{1, \dots, N\} \setminus \{i\}\}$, which by assumption is nonempty. By the consistency condition, $\hat{\theta}_j \xrightarrow{n} \theta_{\text{alt.}}$ for all $j \in \mathfrak{J}$ and $\hat{\theta}_j \xrightarrow{n} \theta_0$ for all $j \in \{1, \dots, N\} \setminus \mathfrak{J}$. Thus, with probability 1,

$$\begin{aligned} \prod_{j \neq i} g_{\theta_j}(T(X_j)) &= \prod_{j \in \mathfrak{J}} g_{\theta_{\text{alt.}}}(T(X_j)) \prod_{j \notin \mathfrak{J} \cup \{i\}} g_{\theta_0}(T(X_j)) \\ &= \prod_{j \neq i} g_{\hat{\theta}_j}(T(X_j)) \\ &= \prod_{j \neq i} \max(g_{\theta_{\text{alt.}}}(T(X_j)), g_{\theta_0}(T(X_j))) \\ &= \sup_{\theta \in \Theta} \prod_{j \neq i} \max(g_{\theta}(T(X_j)), g_{\theta_0}(T(X_j))) \end{aligned}$$

in the limit as $n \rightarrow \infty$, with the equalities holding by the almost-sure continuity of $g_{\bullet}(T(X_j))$ as a function on Θ [14, §1.7]. Since by equation (6),

$$\theta_i^{\dagger} = \arg \sup_{\theta \in \Theta} \sum_{j \neq i} \max(g_{\theta}(T(X_j)), g_{\theta_0}(T(X_j))),$$

it follows that $\theta_i^{\dagger} \xrightarrow{n} \theta_i$. ■

Heuristically, the key observation of the proof is that whether θ is constrained to have one of two values has no asymptotic effect on the estimates of θ_j . The universality of the codelength function is a consequence.

Theorem 8 (Universality). *Under the conditions of Lemma 7,*

$$\lim_{n \rightarrow \infty} E_{\theta_{\text{alt.}}} \left(\frac{L_i^{\dagger}(T(X_i))}{n} \right) = \lim_{n \rightarrow \infty} E_{\theta_{\text{alt.}}} \left(\frac{-\log g_{\theta_{\text{alt.}}}(T(X_i))}{n} \right)$$

for all $i \in \{1, \dots, N\}$ such that $\theta_i = \theta_{\text{alt.}}$, where $E_{\theta_{\text{alt.}}}$ signifies the expectation value with respect to $g_{\theta_{\text{alt.}}}$, i.e., $E_{\theta_{\text{alt.}}}(\bullet) = \int \bullet dP_{\theta_{\text{alt.}}}$.

Proof: $P_{\theta_{\text{alt.}}}(\lim_{n \rightarrow \infty} \theta_i^{\dagger} \in \{\theta_0, \theta_{\text{alt.}}\}) = 1$ for all $i \in \{1, \dots, N\}$ since $\theta_i^{\dagger} \xrightarrow{n} \theta_i$ by the lemma and $\theta_i \in \{\theta_0, \theta_{\text{alt.}}\}$ by assumption. Hence, $\theta_i^{\dagger} \xrightarrow{n} \theta_{\text{alt.}}$ for all $i \in \{1, \dots, N\}$ such that $\theta_i = \theta_{\text{alt.}}$. Thus, for those values of i ,

$$\lim_{n \rightarrow \infty} E_{\theta_{\text{alt.}}} \left(\frac{-\log(g_{\theta_{\text{alt.}}}(T(X_i))/g_{\theta_i^{\dagger}}(T(X_i)))}{n} \right) = 0$$

since $g_{\theta_{\text{alt.}}}(T(X_i))/g_{\theta_i^{\dagger}}(T(X_i)) \xrightarrow{n} 1$ by the almost-sure continuity of $g_{\bullet}(T(X_i))$ as a function on Θ [14, §1.7]. ■

The $N \rightarrow \infty$ universality of a related mixture code is established in Appendix B.

E. Model selection (hypothesis testing)

Let P_i^{\dagger} denote the probability distribution satisfying

$$P_i^{\dagger}(\theta_0) = (N-1)^{-1} \sum_{j \neq i} 1 \left(\Delta_i^{\dagger}(T(x_i)) \geq 0 \right)$$

and $P_i^{\dagger}(\theta_i^{\dagger}) = 1 - P_i^{\dagger}(\theta_0)$, where $1(\bullet)$ is the indicator function equal to 1 if its argument is true and equal to 0 otherwise. Then the *codelength* of P_i^{\dagger} and $T(x_i)$ is the codelength of the

parameter value plus the codelength of the reduced data and is specifically

$$L_i^{\dagger}(P_i^{\dagger}; T(x_i)) = -\log P_i^{\dagger}(\theta_i^{\dagger}) + L_i^{\dagger}(T(x_i)) \quad (8)$$

relative to $\theta_i \neq \theta_0$, the alternative hypothesis, or

$$L^0(P_i^{\dagger}; T(x_i)) = -\log P_i^{\dagger}(\theta_0) + L^0(T(x_i)) \quad (9)$$

relative to $\theta_i = \theta_0$, the null hypothesis.

The MDL model selection code [8, pp. 409, 423-424] is that for which the codelength of the test statistic of the i th feature is

$$\min \left(L_i^{\dagger}(P_i^{\dagger}; T(x_i)), L^0(P_i^{\dagger}; T(x_i)) \right),$$

That code selects the same hypotheses as does

$$\Delta_i^{\dagger}(P_i^{\dagger}; T(x_i)) = L_i^{\dagger}(P_i^{\dagger}; T(x_i)) - L^0(P_i^{\dagger}; T(x_i)),$$

the *information* in P_i^{\dagger} and $T(x_i)$ for *discrimination* favoring the null hypothesis that $\theta_i = \theta_0$ over the alternative hypothesis that $\theta_i \neq \theta_0$. If P_i^{\dagger} is extended to define a random variable $\hat{\theta}$ such that $g_{\theta}(T(x_i))$ is the probability density of the i th statistic conditional on $\hat{\theta} = \theta$ for both $\theta = \theta_0$ and $\theta = \theta_i^{\dagger}$, then, according to equations (8) and (9), the information for discrimination is the logarithm of the posterior odds:

$$\Delta_i^{\dagger}(P_i^{\dagger}; T(x_i)) = \log \frac{P_i^{\dagger}(\hat{\theta} = \theta_0 | T(X_i) = T(x_i))}{P_i^{\dagger}(\hat{\theta} = \theta_i^{\dagger} | T(X_i) = T(x_i))}.$$

Solving for $P_i^{\dagger}(\hat{\theta} = \theta_0 | T(X_i) = T(x_i))$ gives

$$P_i^{\dagger}(\theta_0 | T(x_i)) = \frac{1}{1 + 1/\log^{-1} \Delta_i^{\dagger}(P_i^{\dagger}; T(x_i))}. \quad (10)$$

Defining the distribution P_i^{\ddagger} by

$$P_i^{\ddagger}(\theta_0) = N^{-1} \sum_{j=1}^N 1 \left(\Delta_i^{\ddagger}(T(x_i)) \geq 0 \right)$$

and otherwise replacing coding system \dagger with coding system \ddagger in the equations of this subsection yields $\Delta_i^{\ddagger}(P_i^{\ddagger}; T(x_i))$ as an approximation of $\Delta_i^{\dagger}(P_i^{\dagger}; T(x_i))$. The proportion $P_i^{\ddagger}(\theta_0)$ is the maximum likelihood estimate of the proportion of features that correspond to true null hypotheses.

In the empirical Bayes framework constrained to the parametric family, both $P_i^{\dagger}(\theta_0)$ and $P_i^{\ddagger}(\theta_0)$ may serve as estimates of π_0 . Likewise, the $P_i^{\dagger}(\theta_0 | T(x_i))$ of equation (10) and its approximate probability $P_i^{\ddagger}(\theta_0 | T(x_i))$ are suitable as estimates of the LFDR.

IV. APPLICATION

Alex Miron's lab at the Dana-Farber Cancer Institute measured abundance levels of each of 20 plasma proteins of each of 55 women with HER2-positive breast cancer, 35 women mostly with ER/PR-positive breast cancer, and 64 healthy women [15]. The respective data vectors $x_1^{\text{HER2}}, \dots, x_{20}^{\text{HER2}}$,

$x_1^{\text{ER/PR}}, \dots, x_{20}^{\text{ER/PR}}, y_1, \dots, y_{20}$ were created by adding the first quartile of the abundance levels (over the 64 healthy women and over all proteins) to each abundance level and by taking natural logarithms of the resulting sums; similar conservative preprocessing steps have worked well with gene expression data [16].

The preprocessed data were modeled as normally distributed per Example 3. Following the notation of the example, ξ_i^{HER2} , $\xi_i^{\text{ER/PR}}$, and η_i are the expectation values of X_i^{HER2} , $X_i^{\text{ER/PR}}$, and Y_i , respectively, and are as such interpretable as population levels of the abundance of protein i . The parameters of interest are $\theta_i^{\text{HER2}} = |\xi_i^{\text{HER2}} - \eta_i|/\sigma_i$ and $\theta_i^{\text{ER/PR}} = |\xi_i^{\text{ER/PR}} - \eta_i|/\sigma_i$, the standardized levels of the i th protein's abundance relative to the healthy controls.

The data were analyzed according to the distributions of $T(X_i^{\text{HER2}}, Y_i)$ and $T(X_i^{\text{ER/PR}}, Y_i)$ given in Example 3 using the coding schemes of Section III. The results are displayed as Figures 1-4. In Figure 1, the left panel indicates that the information in the data favors differential protein abundance over equivalent protein abundance for most proteins, and the near-zero minimum regret in right panel at an estimated parameter value of about -0.5 shows that θ^\ddagger is close to the parameter value that maximizes the likelihood functions of some of the proteins. Figure 2 is similar but includes a comparison to a more nonparametric method that can achieve negative regret because its estimate of the alternative distribution need not lie in $\{g_\theta : \theta \in \Theta\}$, potentially leading to overfitting. Figure 3 demonstrates that the number of features need not be high to effectively achieve the asymptotic equivalence between the approximate and exact universal distributions. Figure 4 shows a shift relative to Figures 1-2 by an amount of bits equal to the discrimination information in the parameter distribution.

V. DISCUSSION

A. Medium-scale inference

This paper proposes methods for quantifying the information in data supporting the null hypothesis over the alternative hypothesis and vice versa. The proposed use of a parametric family of distributions does not suffer from the tendency of nonparametric methods of large-scale inference to overfit the data of medium-dimensional biology.

B. Severability of the general methods

Three of the methods highlighted are severable from each other in the sense that each applies to settings in which the others do not. The remainder of this section addresses more general applications of each method.

First, the technique of coding statistics rather than the original data can be used in MDL model selection contexts more general than those involving multiple features. Second, whether or not the data are reduced to test statistics, the strategy of designing a code based on data from features other than the feature of the data currently coded can be generalized beyond the specific method proposed. (An analogous strategy is also finding application to the estimation of mutual information [17].)

Finally, the information for discrimination as in effect generalized by [12] is a viable substitute for the p -value as a measure of evidence concerning hypotheses under consideration. The problematic nature of the asymmetry of conventional hypothesis testing is well known. On one hand, the p -value tends to favor the null hypothesis given insufficient data even if the alternative hypothesis is true. On the other hand, it tends to insufficiently favor of the null hypothesis asymptotically if it is true since the p -value distribution in that case remains uniform irrespective of the amount of data accumulated. As a result, the p -value is not interpretable as a measure of evidence apart from sample size adjustments [18, 19]. By contrast, the information for discrimination, by virtue of its basis on universal distributions, tends to favor whichever hypothesis is true by an amount proportional to the size of the sample.

APPENDIX A

Approximate coding

Assume X_1, X_2, \dots are independent and each of identical distribution P_\star . For example, P_\star could be a K -component mixture distribution $P_\star = \sum_{k=1}^K \pi_k P_{\star k}$, where π_k is the probability that some X_j has distribution $P_{\star k}$, which is not necessarily in $\{P_{\theta, \lambda} : \theta \in \Theta, \lambda \in \Lambda\}$. Let $E_\star(\bullet)$ and \xrightarrow{N} denote the expectation value and almost sure convergence as $N \rightarrow \infty$ with respect to P_\star .

Theorem 9. *Suppose that, for all $i \in \{1, \dots, N\}$, $E_\star(\log g_\theta(T(X_j))) < \infty$ for all $\theta \in \Theta$ and that θ^\ddagger and θ_i^\ddagger are unique with P_\star -probability 1. Then $\theta^\ddagger - \theta_i^\ddagger \xrightarrow{N} 0$ for all $i \in \{1, \dots, N\}$.*

Proof: For any $\theta \in \Theta$, let $\hat{\theta}_j(\theta) = \arg \max_{\tilde{\theta} \in \{\theta_0, \theta\}} g_{\tilde{\theta}}(T(x_j))$. Since $\log g_{\hat{\theta}_j(\theta)}(T(X_j))$ is IID for all $j \in \{1, \dots, N\}$, the strong law of large numbers implies that, for all $\mathcal{J}_N \in \{\{1, \dots, N\}, \{1, \dots, N\} \setminus \{1\}, \dots, \{1, \dots, N\} \setminus \{N\}\}$,

$$\begin{aligned} & \frac{1}{|\mathcal{J}_N|} \sum_{j \in \mathcal{J}_N} \log g_{\hat{\theta}_j(\theta)}(T(X_j)) \xrightarrow{N} E_\star \left(\log g_{\hat{\theta}_j(\theta)}(T(X_j)) \right) \\ & = P_\star \left(\hat{\theta}_j(\theta) = \theta_0 \right) E_\star \left(\log g_{\hat{\theta}_j(\theta)}(T(X_j)) \mid \hat{\theta}_j(\theta) = \theta_0 \right) \\ & \quad + P_\star \left(\hat{\theta}_j(\theta) = \theta \right) E_\star \left(\log g_{\hat{\theta}_j(\theta)}(T(X_j)) \mid \hat{\theta}_j(\theta) = \theta \right), \end{aligned}$$

the finiteness of which follows from that of $E_\star(\log g_\theta(T(X_j)))$. As the result holds for arbitrary $\theta \in \Theta$,

$$\begin{aligned} & \arg \sup_{\theta \in \Theta} \frac{1}{|\mathcal{J}_N|} \sum_{j \in \mathcal{J}_N} \log g_{\hat{\theta}_j(\theta)}(T(X_j)) \xrightarrow{N} \\ & \arg \sup_{\theta \in \Theta} E_\star \left(\log g_{\hat{\theta}_i(\theta)}(T(X_j)) \right) \end{aligned}$$

irrespective of whether the sum on the left-hand-side is over $\{1, \dots, N\}$ or over $\{1, \dots, N\} \setminus \{i\}$ for some $i \in \{1, \dots, N\}$. (The uniqueness of the maximizing value of θ on the left-hand-side is guaranteed by the postulated uniqueness of θ^\ddagger and θ_i^\ddagger .) Therefore, the difference in the maximum likelihood

estimate of θ under the alternative hypothesis using X_1, \dots, X_N and that using $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_N$ converges almost surely to 0, but such maximum likelihood estimates are θ_i^\ddagger and θ_i^\ddagger , respectively, according to Theorem 5 and Corollary 6. ■

APPENDIX B

Mixture coding

This section extends the the fixed-component results of Section III-D to the two-component mixture density of equation (2) with the constraint that $g_{\text{alt.}} = g_{\theta_{\text{alt.}}}$ for some $\theta_{\text{alt.}} \in \Theta$. In this setting, the universal density g_i^\ddagger and its approximation g_i^\ddagger are replaced with $g_i^* = g_{\theta_i^*}$ and its approximation $g^{**} = g_{\theta^{**}}$, where

$$\langle \theta_i^*, \pi_{0i}^* \rangle = \arg \sup_{(\theta, \pi_0) \in \Theta \times [0,1]} \prod_{j \neq i} \left(\frac{\pi_0 g_{\theta_0}(T(X_j)) +}{(1 - \pi_0) g_{\theta}(T(X_j))} \right),$$

with θ^{**} and π_0^{**} defined analogously except with the product over all $\{1, \dots, N\}$.

Assuming the statistics are independent, $\langle \theta_i^*, \pi_{0i}^* \rangle$ and $\langle \theta^{**}, \pi_0^{**} \rangle$ are clearly maximum likelihood estimates of $\langle \theta_{\text{alt.}}, \pi_0 \rangle$. (In that context, [20] used θ^{**} and π_0^{**} to estimate global false discovery rates, and the mixture codes similarly form LFDR estimates via substituting either θ^* and π_0^* or θ^{**} and π_0^{**} into equations (1) and (2).) Consequently, the steps used to prove Theorem 9 also demonstrate that $\theta^{**} - \theta_i^* \xrightarrow{N} 0$ and $\pi_0^{**} - \pi_{0i}^* \xrightarrow{N} 0$ for all $i \in \{1, \dots, N\}$ under the independence condition.

Whereas regularity conditions entailing the strong consistency of maximum likelihood estimates for finite-mixture models [21] would apply as $N \rightarrow \infty$, seemingly more pertinent to universality is consistency in the sense of \xrightarrow{n} , almost sure convergence as $n \rightarrow \infty$ under the stationarity of every $T(X_i)$, assumed to have density g for some values of π_0 and $\theta_{\text{alt.}}$. However, such \xrightarrow{n} consistency does not hold if N is finite and if $\pi_0 > 0$, for in that case, there is fixed, nonzero probability π_0^N that all N statistics have probability density function g_{θ_0} rather than $g_{\theta_{\text{alt.}}}$. For that reason, \xrightarrow{N} consistency will be used instead.

Theorem 10. *If the maximum likelihood estimate θ^{**} almost surely converges to $\theta_{\text{alt.}}$ as $N \rightarrow \infty$ and if $g_{\bullet}(T(X_i))$ is almost surely continuous on Θ for all $i \in \{1, 2, \dots\}$, then*

$$\lim_{N \rightarrow \infty} E_{\theta_{\text{alt.}}} (L_i^*(T(X_i)) / n) = E_{\theta_{\text{alt.}}} (-\log g_{\theta_{\text{alt.}}}(T(X_i)) / n)$$

for all $i \in \{1, 2, \dots\}$ such that $\theta_i = \theta_{\text{alt.}}$, where $L_i^*(T(X_i)) = -\log g_{\theta_i^*}(T(X_j))$ and $E_{\theta_{\text{alt.}}}$ signifies the expectation value with respect to $g_{\theta_{\text{alt.}}}$, i.e., $E_{\theta_{\text{alt.}}}(\bullet) = \int \bullet dP_{\theta_{\text{alt.}}}$.

Proof: Since θ_i^* is the maximum likelihood estimate for the $N - 1$ statistics other than $T(X_i)$, $\theta_i^* \xrightarrow{N} \theta_{\text{alt.}}$. Thus, the claim follows from reasoning analogous to that used to prove Theorem 8. ■

Corollary 11 (Asymptotic universality). *Given the conditions of Theorem 10,*

$$\lim_{n \rightarrow \infty} \lim_{N \rightarrow \infty} E_{\theta_{\text{alt.}}} \left(\frac{L_i^*(T(X_i))}{n} \right)$$

$$= \lim_{n \rightarrow \infty} E_{\theta_{\text{alt.}}} \left(\frac{-\log g_{\theta_{\text{alt.}}}(T(X_j))}{n} \right)$$

for all $i \in \{1, 2, \dots\}$ such that $\theta_i = \theta_{\text{alt.}}$.

The proof is trivial. The corollary means

$$(L_i^*(T(x_i)) - \log(1 - \pi_{0i}^*)) - (L^0(T(x_i)) - \log \pi_{0i}^*)$$

may be regarded as approaching the information for discrimination under the mixture model as $N \rightarrow \infty$. Since $\theta^{**} - \theta_i^* \xrightarrow{N} 0$ and $\pi_0^{**} - \pi_{0i}^* \xrightarrow{N} 0$, that information is approximated by substituting the maximum likelihood estimates θ^{**} and π_0^{**} for θ_i^* and π_{0i}^* .

REFERENCES

- [1] B. Efron, "Large-scale simultaneous hypothesis testing: The choice of a null hypothesis," *Journal of the American Statistical Association*, vol. 99, no. 465, pp. 96–104, 2004.
- [2] B. Efron, "Size, power and false discovery rates," *Annals of Statistics*, vol. 35, pp. 1351–1377, 2007.
- [3] C. M. Yanofsky and D. R. Bickel, "Validation of differential gene expression algorithms: Application comparing fold-change estimation to hypothesis testing," *BMC Bioinformatics*, vol. 11, p. 63, 2010.
- [4] Z. Montazeri, C. M. Yanofsky, and D. R. Bickel, "Shrinkage estimation of effect sizes as an alternative to hypothesis testing followed by estimation in high-dimensional biology: Applications to differential gene expression," *Statistical Applications in Genetics and Molecular Biology*, vol. 9, p. 23, 2010.
- [5] S. Kullback, *Information Theory and Statistics*. New York: Dover, 1968.
- [6] J. Rissanen, *Information and Complexity in Statistical Modeling*. New York: Springer, 2007.
- [7] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2743–2760, 1998.
- [8] P. D. Grünwald, *The Minimum Description Length Principle*. London: The MIT Press, 2007.
- [9] T. Severini. Oxford: Oxford University Press, 2000.
- [10] Y. Pawitan, *In All Likelihood: Statistical Modeling and Inference Using Likelihood*. Oxford: Clarendon Press, 2001.
- [11] T. Schweder and N. L. Hjort, "Confidence and likelihood," *Scandinavian Journal of Statistics*, vol. 29, no. 2, pp. 309–332, 2002.
- [12] J. Rissanen, "Stochastic complexity," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 49, no. 3, pp. 223–239, 1987.
- [13] R. Royall, "On the probability of observing misleading statistical evidence," *Journal of the American Statistical Association*, vol. 95, no. 451, pp. 760–768, 2000.
- [14] R. J. Serfling, *Approximation theorems of mathematical statistics*. New York: Wiley, 1980.
- [15] X. Li, "ProData," *Bioconductor.org documentation for the ProData package*, 2009.

- [16] D. R. Bickel, "Microarray gene expression analysis: Data transformation and multiple-comparison bootstrapping," *Computing Science and Statistics*, vol. 34, pp. 383–400, 2002.
- [17] M. Gastpar, P. Gill, A. Huth, and F. Theunissen, "Anthropic correction of information estimates and its application to neural coding," *IEEE Transactions on Information Theory*, vol. 56, no. 2, pp. 890–900, 2010.
- [18] R. Royall, *Statistical Evidence: A Likelihood Paradigm*. New York: CRC Press, 1997.
- [19] D. R. Bickel, "The strength of statistical evidence for composite hypotheses: Inference to the best explanation," *Technical Report, Ottawa Institute of Systems Biology, 2008 draft available at biostats.bepress.com/cobra/ps/art49*, 2010.
- [20] Y. Pawitan, K. Murthy, S. Michiels, and A. Ploner, "Bias in the estimation of false discovery rate in microarray studies," *Bioinformatics*, vol. 21, no. 20, pp. 3865–3872, 2005.
- [21] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Review*, vol. 26, no. 2, pp. 195–239, 1984.

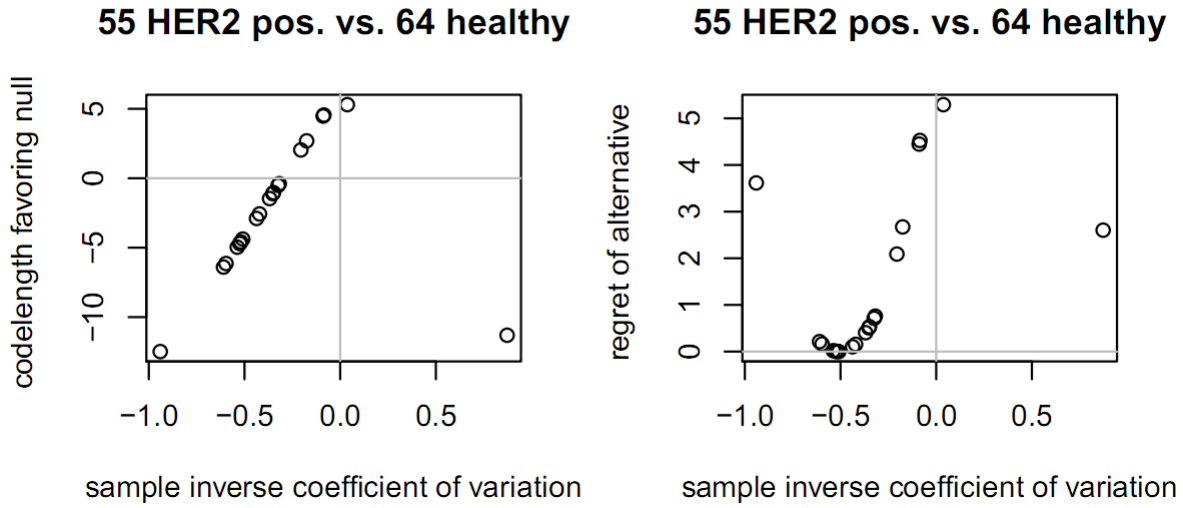


Figure 1. Codelengths and regrets for protein abundance of women with HER2-positive breast cancer relative to healthy women. Each circle corresponds to a different protein. Left panel: $L^\dagger(T(x_i)) - L^0(T(x_i))$, the approximate information for discrimination in favor of the null hypothesis for the i th protein; negative values favor the alternative hypothesis. Right panel: $\text{reg}(g^\dagger, x_i)$, the regret relative to the alternative hypothesis.

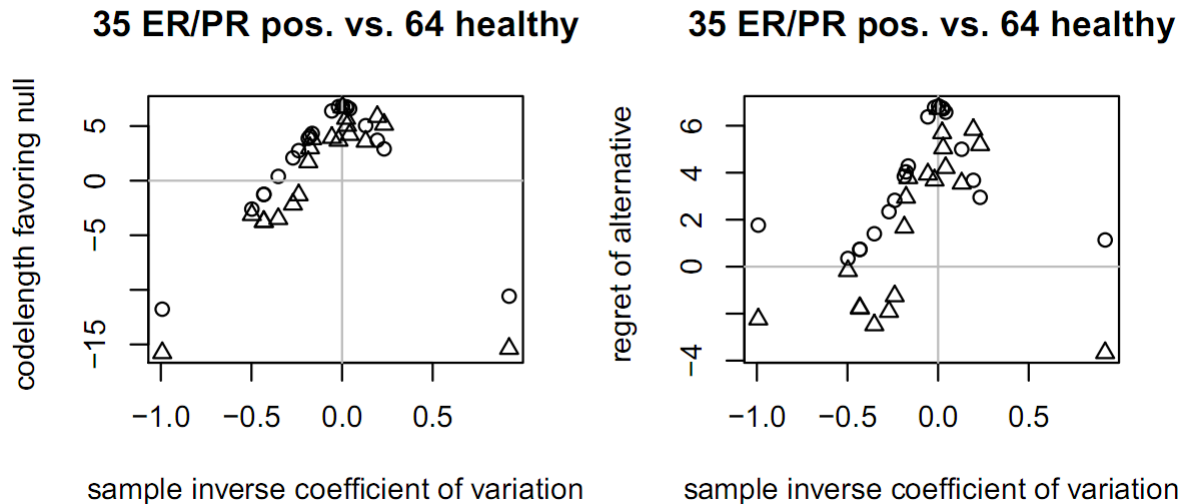


Figure 2. The circles are as in Figure 1, except for women mostly with ER/PR-positive breast cancer. The triangles represent the analogous results (5) of a “theoretical null” empirical Bayes method [2] as implemented in the `locfdr` R package, which failed to assign different codelengths to different proteins for the women with HER2-positive breast cancer.

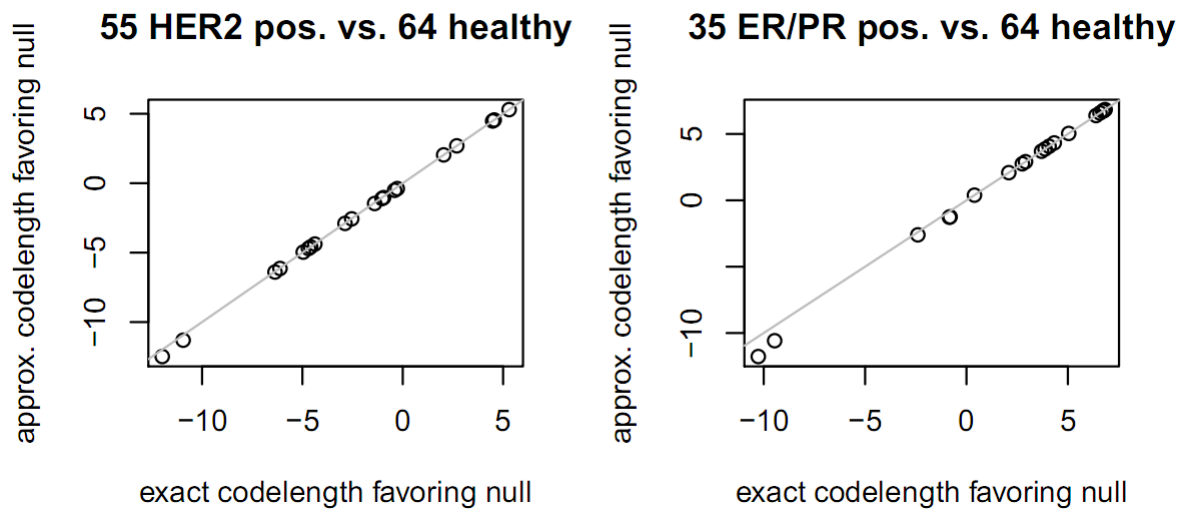


Figure 3. Approximate information $L^{\frac{1}{2}}(T(x_i)) - L^0(T(x_i))$ versus exact information $L_i^{\frac{1}{2}}(T(x_i)) - L^0(T(x_i))$ for the i th protein. Each panel corresponds to a different type of cancer, and each circle corresponds to a different protein.

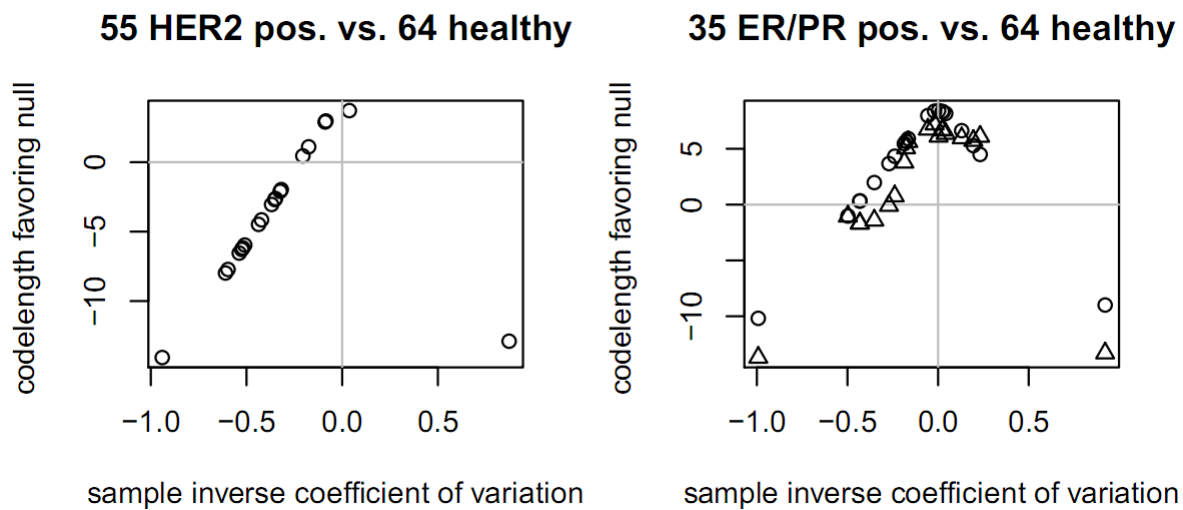


Figure 4. The i th circle of each panel represents the discrimination information $\Delta_i^{\ddagger}(P_i^{\ddagger}; T(x_i))$ of both the parameter distribution P_i^{\ddagger} and the reduced measurement $T(x_i)$ for the i th protein. The triangles in the right panel represent the estimated posterior log-odds derived from the LFDR estimates of a “theoretical null” empirical Bayes method [2] as implemented in the `locfdr` R package.