A MINIMUM DESCRIPTION LENGTH METHOD OF MEDIUM-SCALE SIMULTANEOUS INFERENCE

David R. Bickel

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 the 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.

Keywords: information criteria; minimum description length; model selection; reduced likelihood

1. INTRODUCTION

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). To adequately interpret 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 $x_1, \ldots, x_N$ modeled as realizations of the random variables $X_1, \ldots, X_N$ and for test statistics $T(X_1), \ldots, T(X_N)$ corresponding to null hypotheses $\psi_1 = \psi_0, \ldots, \psi_N = \psi_0$, under which each of the test statistics has a common probability density function g_{ψ_0}, the local false discovery rate for the ith feature is the posterior probability

$$\begin{aligned} \mathrm{LFDR}(x_i) &= P(\psi_i = \psi_0 | T(X_i) = T(x_i)) \\ &= \frac{P(\psi_i = \psi_0)\, g_{\psi_0}(T(x_i))}{g(T(x_i))}, \end{aligned}$$

where $\pi_0 = P(\psi_i = \psi_0)$ is the proportion of null hypotheses that are true and where $g = \pi_0 g_{\psi_0} + (1 - \pi_0)\, g_{\text{alt.}}$ is a mixture density with $g_{\text{alt.}}$ as the probability density function of the test statistic marginal over all alternative features, those for which the null hypothesis $\psi_i = \psi_0$ 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 $\widehat{\mathrm{LFDR}}(x_i)$, the estimated local false discovery rate. Nonparametric methods of estimating π_0 and g include matching the central region of a histogram of statistics [1] or maximizing the likelihood over a truncated normal family of distributions [2]. The empirically demonstrated success of such methods [3, 4] lies in the reliability of nonparametric density estimation in the presence of several thousands of features.

This work was partially supported by the Faculty of Medicine of the University of Ottawa.

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 solutions 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.

Consider the log-likelihood ratio

$$\log\left(\frac{g_{\psi_0}\left(T\left(x_i\right)\right)}{g_{\psi_i}\left(T\left(x_i\right)\right)}\right)$$

as the information in $X_i = x_i$ for discrimination favoring the hypothesis that $\psi = \psi_0$ over the hypothesis that $\psi = \psi_1$ for some $\psi_i \neq \psi_0$ [5, pp. 4-7]. Since ψ_i is unknown, the next section 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] for introductions to the MDL principle of model selection.) The theory is then illustrated with a medium-dimensional biological data set. The paper concludes with a discussion of the general applicability of the methods proposed.

2. THEORY

2.1. Data reduction and distributions

The observed data vector $x \in \mathcal{X}$ is considered a realization of the random variable X of probability distribution $P_{\psi,\lambda}$ that admits a probability density function $f_{\psi,\lambda}$ with respect to some dominating measure, where $\psi \in \Psi$ 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 $\psi_0 \in \Psi$, we have $\psi = \psi_0$ under the null hypothesis or narrow model and $\psi \neq \psi_0$ under the alternative hypothesis or wide model.

Let T be a measurable function on $\mathcal{X}$. If, for each $\psi \in \Psi$, the probability density function g_ψ of the *statistic* or *reduced data* $T\left(X\right)$ does not depend on the value of λ, then

$$\ell\left(\psi\right) = g_\psi\left(T\left(x\right)\right)$$

defines the *marginal likelihood function* ℓ.

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

$$\begin{aligned} f_{\psi,\lambda}\left(x\right) &= f_{\psi,\lambda}\left(T\left(x\right), x | T\left(X\right) = T\left(x\right)\right) \\ &= g_\psi\left(T\left(x\right)\right) f_{\psi,\lambda}\left(x | T\left(X\right) = T\left(x\right)\right) \\ &= g_\psi\left(T\left(x\right)\right) f_\lambda\left(x | T\left(X\right) = T\left(x\right)\right) \quad (1) \\ &= C g_\psi\left(T\left(x\right)\right), \end{aligned}$$

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

$$\frac{f_{\psi_1,\lambda}\left(x\right)}{f_{\psi_0,\lambda}\left(x\right)} = \frac{C g_{\psi_1}\left(T\left(x\right)\right)}{C g_{\psi_0}\left(T\left(x\right)\right)} = \frac{\ell\left(\psi_1\right)}{\ell\left(\psi_0\right)}$$

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

Example 1. 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 $\psi = \left(\xi - \eta\right)/\sigma$ with $\psi = 0$ as the null hypothesis and $\psi \neq 0$ as the alternative hypothesis; the parameter space here is $\Psi = \mathbb{R}^1$. A suitable statistic for data reduction is the two-sample t statistic

$$T\left(x, y\right) = \frac{\hat{\xi}\left(x\right) - \hat{\eta}\left(y\right)}{\hat{\sigma}\left(x, y\right)\sqrt{m^{-1} + n^{-1}}}, \quad (2)$$

where $\hat{\xi}$, $\hat{\eta}$, and $\hat{\sigma}$ are the usual unbiased estimators. Then $g_\psi\left(T\left(x, y\right)\right)$, the probability density of $T\left(X, Y\right)$ 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 $\left(m^{-1} + n^{-1}\right)^{-1}\psi$.

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

Example 2. Example 1 is extended to N genes, proteins, or other biological features such that $X_i \sim \mathrm{N}\left(\xi_i, \Sigma_{i,m}\right)$ and $Y_i \sim \mathrm{N}\left(\eta_i, \Sigma_{i,n}\right)$ correspond to the observed outcome $\langle x_i, y_i\rangle$ for the ith feature, where $i = 1, \ldots, N$ 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 $\left(\xi_i > \eta_i\right)$ is close to the number of features with a negative difference $\left(\xi_i < \eta_i\right)$, the parameter of interest for feature i may be $\psi_i = \left|\xi_i - \eta_i\right|/\sigma_i$, the absolute value of the inverse coefficient of variation, with $\psi_i = 0$ as the null hypothesis, $\psi_i > 0$ as the alternative hypothesis, and $\Psi = [0, \infty)$ as the parameter space. Then $T\left(x_i, y_i\right)$ is the absolute value of the two-sample t statistic for $\langle x_i, y_i\rangle$ according to equation (2), and $T\left(X_i, Y_i\right)$ is distributed as the absolute value of a variate from the noncentral Student t probability density with $m + n - 2$ degrees of freedom and noncentrality parameter $\left(m^{-1} + n^{-1}\right)^{-1}\psi_i$. Thus, the density $g_\psi\left(T\left(x_i, y_i\right)\right)$ for the ith feature is defined such that g_ψ is the probability density function of the same distribution.

[8, §8.3] and [9] 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.

2.2. Codelengths

2.2.1. General concepts

According to the MDL framework, each scheme † for coding the data under the alternative hypothesis corresponds to a codelength function $L^\dagger$ on $\mathcal{X}$ and thus to a probability density function $g^\dagger$ selected from the parametric family $\{g_\psi : \psi \in \Psi\}$ prior to 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 ψ_0 is known, the probability density function of the statistic under the null hypothesis is known as g_{ψ_0}, and thus the codelength function L^0 relative to the null hypothesis is already specified by $L^0(T(x)) = -\log g_{\psi_0}(T(x))$.

Suppose, as in Example 2, that there is a vector x_i of measurements for each of N features and that the data are reduced to the statistics $T(x_1), \ldots, T(x_N)$. With $L_i^\dagger(T(x_i))$ as the codelength of $T(x_i)$ relative to the alternative hypothesis, $I^\dagger(x_i) = L_i^\dagger(T(x_i)) - L^0(T(x_i))$ is the *information in $X_i = x_i$ for discrimination* favoring the null hypothesis over the alternative hypothesis.

Example 3. If the restriction to a parametric family were relaxed,

$$-\log \frac{\hat{g}_{\text{alt.}}(T(x_i))}{g_{\psi_0}(T(x_i))} = -\log \frac{1-\widehat{\text{LFDR}}(x_i)}{\widehat{\text{LFDR}}(x_i)} + \log \frac{1-\hat{\pi}_0}{\hat{\pi}_0} \tag{3}$$

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

The *regret* [7] of the codelength function $L_i^\dagger$ given by $L_i^\dagger(T(x_i)) = -\log g_i^\dagger(T(x_i))$ is

$$\begin{aligned} \text{reg}\left(g_i^\dagger, x_i\right) &= L_i^\dagger(T(x_i)) - \inf_{\psi\in\Psi}\left(-\log g_\psi(T(x_i))\right) \\ &= -\log \frac{g_i^\dagger(T(x_i))}{g_{\hat{\psi}}(T(x_i))}, \end{aligned}$$

where $\hat{\psi} = \arg\sup_{\psi\in\Psi} g_\psi(T(x))$. Likewise, the regret of the codelength function relative to the null hypothesis is $\text{reg}(g_{\psi_0}, x_i) = -\log\left(g_{\psi_0}(T(x_i))/g_{\hat{\psi}}(T(x_i))\right)$.

The probability of observing misleading information for discrimination has an upper bound for any distributions g_{ψ_0} and $g_i^\dagger$. Specifically, for any $J > 0$,

$$\begin{aligned} P_{\psi_0,\lambda}\left(I^\dagger(X_i) \le -J\right) &= P_{\psi_0,\lambda}\left(\frac{g_i^\dagger(T(x))}{g_{\psi_0}(T(x))} \ge 2^J\right) \\ &\le 2^{-J}. \end{aligned}$$

A proof of the inequality and applications to the probability of observing misleading evidence appear in [10].

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

2.2.2. Exact codelength

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

$$\psi_i^\dagger = \arg\inf_\psi \sum_{j\ne i} \min\left(\text{reg}(g_\psi, x_j), \text{reg}(g_{\psi_0}, x_j)\right).$$

In words, the code for a given feature uses the distribution in the parametric family that minimizes the regret summed over all other features.

Proportional to N^2, the computation time can prohibit the use of this 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.

2.2.3. Approximate codelength

The † coding scheme is efficiently approximated by a slightly illegal scheme denoted by ‡. 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., $g^\ddagger = g_{\psi^\ddagger}$ for some $\psi^\ddagger \in \Psi$. This is accomplished by minimizing the regret over all features:

$$\psi^\ddagger = \arg\inf_\psi \sum_{j=1}^{N} \min\left(\text{reg}(g_\psi, x_j), \text{reg}(g_{\psi_0}, x_j)\right).$$

This coding scheme is illegal in the sense that $g^\ddagger$ depends on hindsight in that it is a function of the observed data for each feature. However, under general conditions, $g^\ddagger$ approximates $g_i^\dagger$ for all $i = 1, \ldots, N$ given sufficiently large N since the selection of the distribution depends on all features without giving undue weight to any single feature.The next section shows that the approximation can be quite close even for N as small as 20.

3. 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 [11]. The respective data vectors $x_1^{\text{HER2}}, \ldots, x_{20}^{\text{HER2}}, x_1^{\text{ER/PR}}, \ldots, x_{20}^{\text{ER/PR}}, y_1, \ldots, y_{20}$ 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 prepossessing steps have worked well with gene expression data [12].

The preprocessed data were modeled as normally distributed as per Example 2. 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 $\psi_i^{\text{HER2}} = \left|\xi_i^{\text{HER2}} - \eta_i\right| / \sigma_i$ and $\psi_i^{\text{ER/PR}} = \left|\xi_i^{\text{ER/PR}} - \eta_i\right| / \sigma_i$, the standardized levels of the ith protein's abundance relative to the healthy controls.

The data were analyzed according to the distributions of $T\left(X_i^{\text{HER2}}, Y_i\right)$ and $T\left(X_i^{\text{ER/PR}}, Y_i\right)$ given in Example 2 using the coding schemes of Section 2.2. The results are displayed as Figures 1, 2, and 3.

4. DISCUSSION

This paper proposes a general method that quantifies the information in data supporting the null hypothesis over the alternative hypothesis and vice versa. Since the method was specifically designed for medium-scale inference, it does not suffer from the tendency of nonparametric methods of large-scale inference to overfit the data of medium-dimensional biology.

The coding of a distribution of statistics rather than of the original data can be used in model selection contexts more general than those involving multiple features. The reduction of data to carefully chosen statistics for inference about a parameter of interest can greatly simplify applications of the MDL principle to otherwise complex data and models. For example, some strategies for applying nonlinear maximum likelihood (NML) to models with infinite parametric complexity involve imposing restrictions on the data space or on the parameter space, sometimes involving hyperprior distributions [7], and data reduction can facilitate such efforts by in effect reducing the dimension of the parameter as well as that of the data.

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. Returning to the problem of using NML with models of infinite parametric complexity, the parameter space or data space for the ith feature may be restricted according to measurements of other features. However, further research is needed before this approach can be successfully applied. Although limits of the parameter space could in principle be set by the data of the most extreme feature, the results would have high variance due to such dependence on a sample maximum unless the number of features is sufficiently high.

5. 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] P. D. Grünwald, *The Minimum Description Length Principle*. London: The MIT Press, 2007.

[8] T. Severini. Oxford: Oxford University Press, 2000.

[9] T. Schweder and N. L. Hjort, "Confidence and likelihood," *Scandinavian Journal of Statistics*, vol. 29, no. 2, pp. 309–332, 2002.

[10] R. Royall, "On the probability of observing misleading statistical evidence," *Journal of the American Statistical Association*, vol. 95, no. 451, pp. 760–768, 2000.

[11] X. Li, "ProData," *Bioconductor.org documentation for the ProData package*, 2009.

[12] D. R. Bickel, "Microarray gene expression analysis: Data transformation and multiple-comparison bootstrapping," *Computing Science and Statistics*, vol. 34, pp. 383–400, 2002.

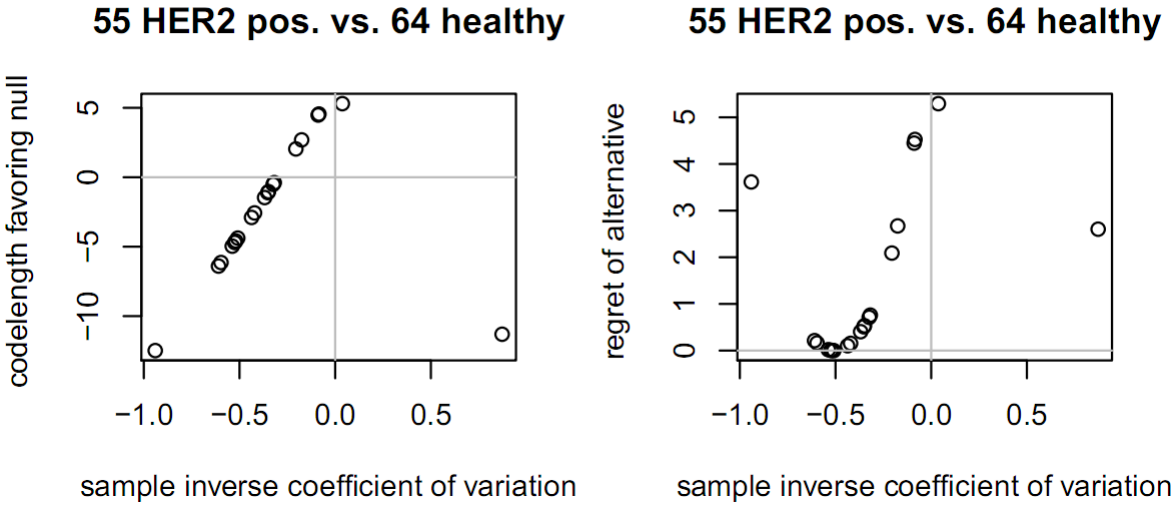

Figure 1. Codelengths and regrets for protein abundance of women with HER2-positive breast cancer relative to healthy women. Each circle corresponds to a different protein. Left panel: $L^{\ddagger}\left(T\left(x_i\right)\right) - L^{0}\left(T\left(x_i\right)\right)$, the approximate information for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

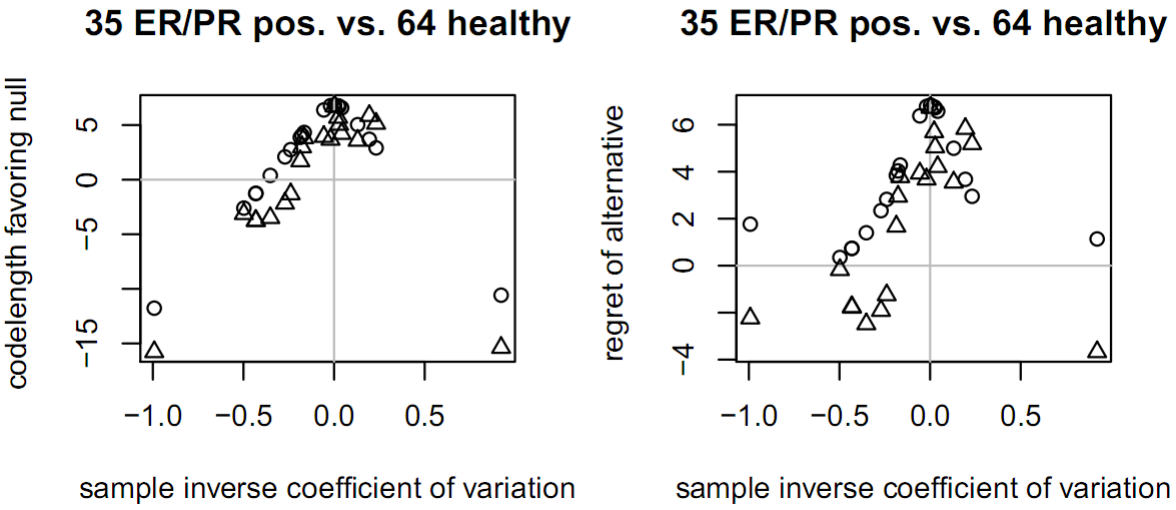

Figure 2. The circles are as in Figure 1, except for women mostly with ER/PR-positive breast cancer. The triangles represent the analogous results (3) of a "theoretical null" empirical Bayes method [2] as implemented in the `locfdr` R package, which failed to assign different codelengths to different proteins for the women with HER2-positive breast cancer.

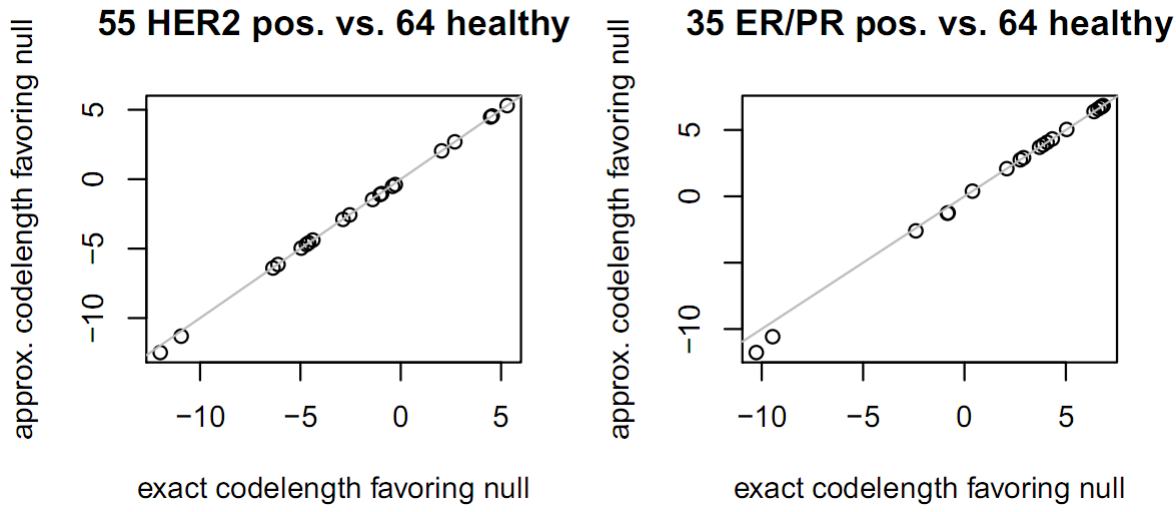

Figure 3. Approximate information $L^{\ddagger}(T(x_i)) - L^{0}(T(x_i))$ versus exact information $L_i^{\dagger}(T(x_i)) - L^{0}(T(x_i))$. Each circle corresponds to a different protein.