Más-o-menos: a simple sign averaging method for discrimination in genomic data analysis

Sihai Dave Zhao 1; Giovanni Parmigiani 2,3, Curtis Huttenhower 2, and Levi Waldron 4

1Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, U.S.A.
2Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, U.S.A.
3Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA 02115, U.S.A.
4City University of New York School of Public Health, Hunter College, New York, NY 10035, U.S.A.

ABSTRACT

Motivation: The successful translation of genomic signatures into clinical settings relies on good discrimination between patient subgroups. Many sophisticated algorithms have been proposed in the statistics and machine learning literature but in practice simpler algorithms are often employed. However, few simple algorithms have been formally described or systematically investigated.

Results: We give a precise definition of a popular simple method we refer to as más-o-menos, which calculates prognostic scores for discrimination by summing standardized predictors, weighted by the signs of their marginal associations with the outcome. We study its behavior theoretically, in simulations, and in an extensive analysis of 27 independent gene expression studies of bladder, breast, and ovarian cancer, altogether totaling 3,833 patients with survival outcomes. We find that despite its simplicity, más-o-menos can achieve quite good discrimination performance. It performs no worse, and sometimes better, than popular and much more CPU-intensive methods for discrimination, including lasso and ridge regression.


Contact: sdzhao@illinois.edu

1 INTRODUCTION

The successful translation of genomic signatures into clinical settings relies on good discrimination between patient subgroups that should receive different clinical management. Relatively sophisticated methods such as penalized regression, support vector machines, random forests, bagging and boosting have seen detailed treatments in the statistics and machine-learning literature (Schölkopf and Smola, 2002; Hastie et al., 2005; Bühlmann and Van De Geer, 2011); however in practice many researcher prefer simpler algorithms (Hand, 2006). A systematic meta-analysis of prognostic models for late-stage ovarian cancer (Waldron et al., 2014) found that the most common methods in the field, and those used to generate the best-performing models on independent datasets, were of the “univariate ensemble” type, where results of univariate regressions are aggregated to formulate a risk score.

The simplest of the univariate ensemble class of prediction methods sets the coefficients of a linear risk score equal to the signs of their univariate associations with the clinical outcome of interest. In other words, for survival analysis it produces a risk score equal to the sum of the “bad prognosis” features minus the sum of the “good prognosis” features. This method and closely related variants can be found in the top clinical, bioinformatic, and general science journals (Colman et al., 2010; Dave et al., 2004; Rème et al., 2013; Bell et al., 2011), and in commercially available prognostic gene signatures, such as the MyPRS signature for multiple myeloma prognosis (Shaughnessy et al., 2007). It has even been proposed, in a formula-free article, as a practical algorithm that can be performed in a spreadsheet with the “software and skill sets available to the cancer biologist” (Hallett et al., 2010).

Despite the popularity and apparent effectiveness of this simplest of methods, to our knowledge it has never been formally described or systematically investigated. We give a precise definition of the procedure and study its behavior theoretically, in simulations, and in an extensive analysis of 27 independent gene expression studies of bladder, breast, and ovarian cancer, altogether totaling 3,833 patients with survival outcomes. We provide theoretical arguments that this method has good discrimination power and low variability when positively correlated features tend to have the same directions of marginal association with outcome. In simulations under a variety of sparsity and covariance structures, it performs competitively with lasso and ridge regression under all situations except the unlikely scenario of independent features, and was more than an order of magnitude faster. In application to survival analysis of three large microarray databases, it performed better than lasso and ridge regression in two of three cancer types, and comparably in the third. We refer to the method as más-o-menos, because in Spanish the phrase “más o menos” means both “plus or minus”, describing the method’s implementation, and “so-so”, describing its theoretically non-optimal but still practically useful discrimination ability.
2 METHODS

2.1 Más-o-menos

Let each component $X_{ij}$ of the $p \times 1$ covariate vector $X_i = (X_{i1}, \ldots, X_{ip})^T$ be a quantitative measurement of the $j^{th}$ gene from the $i^{th}$ subject. The $X_{ij}$ could represent various types of genomic information, such as expression levels from microarrays or next-generation sequencing experiments, or non-genomic data. Más-o-menos uses a patient’s $X_i$ to calculate a signed sum of that patient’s covariate values. The procedure is:

1. Standardize the covariates such that
   \[
   (n - 1)^{-1} \sum (X_{ij} - \bar{X}_j)^2 = 1, \quad j = 1, \ldots, p,
   \]
   where $\bar{X}_j = n^{-1} \sum X_{ij}$.
2. Perform univariate regressions of the outcome on each gene to obtain marginal estimates of the regression coefficient $\hat{\alpha}_j$.
3. Let $\hat{\beta}_j = \text{sgn}(\hat{\alpha}_j)/p^{1/2}$, where $\text{sgn}(c) = 2I(c > 0) - 1$ for $c \neq 0$ and $\text{sgn}(0) = 0$ for $c = 0$.
4. The risk score for the $i^{th}$ patient is calculated as $X_i^T \hat{\nu}$, where $\hat{\nu} = (\hat{\nu}_1, \ldots, \hat{\nu}_p)^T$.

The factor of $p^{1/2}$ in the definition of the $\hat{\nu}_i$ merely serves to ensure the arbitrary scaling $||\hat{\nu}||_2 = 1$. By changing the regression model used in step (3), más-o-menos can be used with clinical outcomes of any type, such as continuous, binary, or censored data. The discrimination performance of $X_i^T \hat{\nu}$ can be quantified using correlation for continuous outcomes, the area under the receiver operator characteristic curve (AUC) for binary outcomes (Bamber, 1975), or the C-statistic for censored outcomes (Uno et al., 2011).

Más-o-menos, and procedures similar to it, is already in use for analyzing genomic data. For example, Donoho and Jin (2008) introduced a family of classifiers, one of which, called HCT-clip, is equivalent to más-o-menos. They found that HCT-clip performed surprisingly well in cross-validation classifiers, one of which, called HCT-clip, is equivalent to más-o-menos.

2.2 Discrimination for survival outcomes

We focus on survival outcomes, because they are typically the most difficult to deal with and the most clinically relevant, and are the outcomes collected in our real data. Let $T_i$ be the survival time of the $i^{th}$ subject. To measure discrimination in the survival setting, we use the C-statistic over the follow-up period $(0, \tau)$, defined by Uno et al. (2011) as

\[
C_v(\beta) = P(g(X_i) > g(X_j) \mid T_i < T_j, T_i < \tau),
\]

where $g(X)$ is the risk score for a subject with covariate vector $X$. We consider linear risk scores of the form $g(X) = X^T \beta$ for $\beta = (\beta_1, \ldots, \beta_p)^T$. We define the optimal weight vector to be

\[
\beta_0 = \arg \max_{\beta} P(X_i^T \beta > X_j^T \beta \mid T_i < T_j, T_i < \tau),
\]

where we have arbitrarily scaled $\beta_0$ to have norm 1 because $C_v(\beta)$ is invariant to scaling of $\beta$.

To implement más-o-menos in this setting, we will obtain the $\hat{\beta}_0$ by fitting univariate Cox models. We choose the Cox model because it is a well-established and well-understood procedure in clinical research. In addition, the estimators $\hat{\beta}_j$ converge to well-defined $\alpha_0$, even when the Cox model is not correctly specified (Struthers and Kalbfleisch, 1986; Lin and Wei, 1989), as is likely to be the case in our marginal regressions. Finally, if the data truly come from a Cox model, the true parameter vector should maximize $C_v$, and should be a scalar multiple of the optimal $\beta_0$.

2.3 Statistical properties

We show that under certain conditions, the más-o-menos weights can have good discrimination power along with low variability. Hand (2006) provided similar arguments to justify equalization of regression coefficients when all covariates have the same direction of effect on the outcome, and this direction is known a priori. Hand describes this in terms of the “flat maximum effect”: that in the context of classifiers, little advantage can be gained in prediction accuracy over very simple models. Here we do not assume that the directions of effect are known.

Let $\nu^* = (\nu^*_1, \ldots, \nu^*_p)^T$ be the probability limit of $\nu$, such that $\nu \rightarrow \nu^*$. Since $\nu^*_j = \text{sgn}(\hat{\alpha}_j)$, if $\hat{\alpha}_j \rightarrow \alpha_0$, in probability, then by the continuous mapping theorem $\nu^*_j = \text{sgn}(\alpha_0)$.

We will analyze the performance of the más-o-menos estimator $\nu$ in terms of the discrimination ability of $\nu^*$ relative to that of $\beta_0$, and the variability of $\nu$ around $\nu^*$. For now we assume $\nu^*_j \neq 0$ for all genes $j$. At the end of the section we discuss the implications if this is not true.

By the definition of $C_v$, the discrimination performance of $\nu^* = (\nu^*_1, \ldots, \nu^*_p)^T$ depends only on the degree of linear association between $X_i^T \beta_0$ and $X_i^T \nu^*$. In addition,

\[
\text{cov}(X_i^T \beta_0, X_i^T \nu^*) = \sum_{j,k} \beta_{0j} \text{cov}(X_{ij}, X_{ik}\nu^*_k) = \sum_{j,k} \beta_{0j} \nu^*_j \nu^*_k \geq \bar{\rho} \sum_{j,k} \beta_{0j} \nu^*_j \nu^*_k,
\]

where $\bar{\rho} = \min p^{-1} \sum_k \text{cov}(X_{ij}\nu^*_j, X_{ik}\nu^*_k)$. The second equality follows because $\nu^*_j \cdot \nu^*_j$ always equals 1. Thus $X_i^T \nu^*$ will be highly correlated with $X_i^T \beta_0$, and will have similar discriminative ability, under the condition that $\sum \beta_{0j} \nu^*_j = \bar{\rho}$ and $\nu^*_j$ have the same sign.

It is not unreasonable to expect these terms to be positive. First, each $\beta_{0j}$ quantifies the association between $X_{ij}$ and $T_i$ conditional on all genes in $X_i$, while each $\nu^*_j$ reflects its univariate association. If a gene has the same direction of effect in both the conditional and marginal models, then $\beta_{0j} \nu^*_j > 0$. This is plausible for at least some genes, and even if it does not hold for all genes $\sum \beta_{0j} \nu^*_j$ can still be positive. Second, the $\bar{\rho}$ term is the minimum average covariance between $X_{ij}\nu^*_j$ and $X_{ik}\nu^*_k$. This will be positive if genes with the same marginal directions of effect tend to be positively correlated, while genes with different marginal directions of effect tend to be negatively correlated. Indeed, the encoded proteins of conserved co-expressed gene pairs are likely to be part of the same biological pathway (van Noort et al., 2003). Again, $\bar{\rho}$ can be positive even if this covariate condition holds only for some pairs of genes, as we merely need the average covariance to be positive.

Restricting the más-o-menos weights to be either $+1$ or $-1$ endows it with low variability, which has been shown to be especially important in classification (Friedman, 1997). The variability of $\nu_j$ is given by

\[
P(\nu_j \neq \nu^*_j) = \left\{ \begin{array}{l} P(\hat{\alpha}_j < 0) \quad \text{if } \alpha_0 > 0, \\
                  P(\hat{\alpha}_j > 0) \quad \text{if } \alpha_0 < 0, \\
                  P(\hat{\alpha}_j = 0) \quad \text{if } \alpha_0 = 0. 
\end{array} \right.
\]

Lin and Wei (1989) showed that $\hat{\alpha}_j \rightarrow N(\alpha_0, \sigma^2/n)$ for some $\alpha_0$ and $\sigma^2$. This approximation, combined with Mill’s inequality, gives the approximate relation

\[
P(\nu_j \neq \nu^*_j) \leq \frac{\sigma^2_j}{n \sqrt{2} |\alpha_0| \sqrt{2n}} \exp \left(-\frac{1}{2} \frac{\sigma^2_j}{\alpha_0^2} \right).
\]
when $\alpha_{ij} \neq 0$, which approaches 0 much faster than \(\text{Var}(\hat{\alpha}_j)\). For large \(n\) and/or large \(|\alpha_{ij}|\), the variability of $\hat{\alpha}_j$ will be very close to zero. Thus \(\hat{v}\) is likely to be less susceptible to overfitting, and as a result can have better out-of-sample discrimination performance.

Difficulties arise when $v_j^2 = 0$ for some marginally unimportant genes \(j\). First, \(\text{cov}(X_i^T \beta_0, X_j^T \nu^*)\) will depend in part on the covariances between these genes and the marginally important ones, and it is unclear how these covariances will behave. Second, since $\hat{\alpha}_j$ is a continuous estimator, \(P(\hat{\alpha}_j \neq 0)\) will equal 1 for any sample size. In other words, más-o-menos may be less predictive and more variable when used on data where many of the covariates are not marginally associated with the outcome. An initial feature screening step may remove many such covariates so that there are few \(j\) such that $\alpha_{ij} = 0$. On the other hand, because gene expression levels tend to be correlated, even genes not involved in the disease process may be correlated to important genes and may have nonzero marginal associations.

3 RESULTS

3.1 Competing methods

We compared más-o-menos to three popular analysis methods that also generate linear risk scores: lasso (Tibshirani, 1996, 1997), ridge regression (Hoerl and Kennard, 1970; Verweij and Van Houwelingen, 1994), and marginal regression (Emura et al., 2012), which gives risk scores of the form $\sum_j X_{ij} \hat{\alpha}_j$. For all methods we first standardized all covariates to have unit variance. We also included two negative controls: 1) the single gene with the largest \(|\alpha_{ij}|\) estimated from the training set, and 2) randomly generated risk scores $\sum_j X_{ij} Z_j$, where the $Z_j$ were drawn independently from a standard normal.

We implemented lasso and ridge regression for the Cox model using the package glmnet (Friedman et al., 2010), selecting the penalty parameter using three-fold cross-validation using the built-in function. Marginal Cox regressions and más-o-menos are implemented in the package survHDD (Bernau et al., 2012).

3.2 Simulations

To simulate training data we generated \(p \times 1\) covariate vectors $X_i$ and survival times from a Cox model with a \(p \times 1\) true parameter vector $\beta_0$. We let the true Cox regression coefficient vector $\beta_0$ have \(s\) non-zero components all with magnitude $s^{-1/2}$, such that $\|\beta_0\|_1 = 1$. The first \(s/2\) nonzero components were positive and the rest were negative. We generated censoring times from an independent exponential distribution to give approximately 50% censoring. In each testing dataset, we replaced the positive entries of $\beta_0$ by random uniform draws from (0, 4/$s^{1/2}$), and the negative entries by random draws from (−4/$s^{1/2}$, 0). Each training and testing dataset contained \(n = 200\) observations.

We considered the low-dimensional case of \(p = 100\) and the high-dimensional one of \(p = 10000\). To generate sparse $\beta_0$, we let \(s = 10\), and for non-sparse $\beta_0$ we let \(s = p\). We drew $X_i$ from a multivariate normal with mean zero and unit marginal variance. From Section 2.3, the discrimination ability of más-o-menos depends on the covariance structure of the $X_i$. In an “easy” setting, the covariates were divided into two blocks, with $X_i$ positively correlated within blocks and negatively correlated between blocks. Those $X_{ij}$ with $\beta_{ij} > 0$ were assigned to one block, those with $\beta_{ij} < 0$ were assigned to the other, and those with $\beta_{ij} = 0$ were assigned equally between the blocks. In a “hard” setting, we let $\text{cov}(X_{ij}, X_{ik}) > 0$ for \(j\) and \(k\) both even or both odd, and $\text{cov}(X_{ij}, X_{ik}) < 0$ otherwise. We let $|\text{cov}(X_{ij}, X_{ik})| = 0$, 0.3, or 0.5 for all \(j\) and \(k\) and ran 200 simulations.

The computations in this paper were run on the Odyssey cluster supported by the FAS Science Division Research Computing Group at Harvard University. Table 1 illustrates the speed advantage enjoyed by más-o-menos.

In general, más-o-menos kept pace with lasso, ridge, and marginal regression. Each of these performed better than the single best gene and the randomly generated negative control. Figure 1 reports the average out-of-sample C-statistics obtained by the different methods. The C-statistics were calculated at $\tau = 2$, where $\tau$ is defined in (1). Confidence intervals represent the empirical 2.5% and 97.5% quantiles. The results clearly illustrate the importance of the covariance structure. All of the methods except for the negative control performed better under the easy covariance setting than under the hard one. The easy covariance satisfies the assumptions of the theoretical discussion in Section 2.3: $\text{cov}(X_{ij} v_j^*, X_{ik} v_k^*) > 0$ and $\text{cov}(X_{ij} v_j^*, X_{ik} v_k^*) < 0$ for all \(j, k\). The difficulty of the hard covariance structure arises from the fact that it is impossible to meet this condition. For example, by construction, $\text{cov}(X_{ij}, T_j) > 0$ and $\text{cov}(X_{ij}, T_k) < 0$, but $\text{cov}(X_{ij}, X_{kj}) < 0$. In other words, the signs of the $\beta_{ij}$ and the covariances are incoherent in the hard covariance case.

When the covariates were independent, higher dimensionality made discrimination harder regardless of sparsity, perhaps because there was no way to borrow information across the covariates. Under the easy covariance structure with a dense $\beta_0$, however, high dimensionality was actually beneficial, perhaps because if the effects of some covariates were by chance incorrectly estimated, there were many other correlated ones that could be used as surrogates. On the other hand, with a hard covariance matrix, dimensionality added difficulty even in the non-sparse case because of the incoherence between the $\beta_{ij}$ and the covariance correlations.

With no correlation, sparsity allowed for easier discrimination. When correlation was introduced in the easy covariance setting, sparsity was detrimental to prediction. This might have been due to the screening step, because univariate screening is liable to retain unimportant covariates simply because they are correlated with important ones. These incorrectly retained covariates can degrade performance. In the hard covariance setting, however, sparsity was helpful regardless of the level of correlation. This may be because in the sparse case, there were fewer important covariates with which the hard covariance structure could cause difficulty.

<table>
<thead>
<tr>
<th>Method</th>
<th>$p = 100$</th>
<th>$p = 10000$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lasso</td>
<td>8.914</td>
<td>47.238</td>
</tr>
<tr>
<td>Ridge</td>
<td>0.645</td>
<td>30.124</td>
</tr>
<tr>
<td>Marginal</td>
<td>0.016</td>
<td>2.209</td>
</tr>
<tr>
<td>Máis-o-menos</td>
<td>0.023</td>
<td>2.408</td>
</tr>
<tr>
<td>Single</td>
<td>0.017</td>
<td>1.674</td>
</tr>
<tr>
<td>Random</td>
<td>0.001</td>
<td>0.004</td>
</tr>
</tbody>
</table>

Runtimes are reported in seconds

| Table 1. Average simulation runtimes |
3.3 Application to bladder, breast, and ovarian cancer

We applied más-o-menos, lasso, ridge, marginal regression, and the two negative control methods to an extensive compendium of real cancer gene expression data (see Table 2). We obtained six bladder cancer datasets totaling 679 patients from Riester et al. (2012), seven breast cancer datasets totaling 1709 patients from Haibe-Kains et al. (2012), and 14 ovarian cancer datasets totaling 1445 patients from Ganzfried et al. (2013). We processed the breast cancer data as in Bernau et al. (2014). The bladder and ovarian cancer data have been manually curated to have standardized clinical annotations, probeset identifiers, and microarray preprocessing, and are available in the Bioconductor packages curatedBladderData and curatedOvarianData, respectively.

For each disease, we limited our analyses to the probesets common to all studies. We trained each algorithm on the largest available study and evaluated its performance on each of the remaining datasets using the C-statistic calculated at $\tau = 5$ years, where $\tau$ is defined in (1). Roughly 60% of all study participants, combined across all diseases, were still alive after five years. The C-statistic is robust to the choice of $\tau$ unless very few deaths or censoring events occur at times greater than $\tau$ (Uno et al., 2011).

We generated 100 bootstrap samples of each validation dataset to obtain 95% confidence intervals. In addition to applying the methods without feature selection, we also implemented higher criticism thresholding (Donoho and Jin, 2008), which screens out covariates with high marginal Cox regression p-values but is entirely data-driven and automatically chooses the number of covariates to retain. Summary statistics were calculated by fixed effects meta-analysis with the metafor package (Viechtbauer, 2010).

Figure 2 reports the results. Selecting only a single gene or using random weights gave the lowest performance, confirming the appropriateness of our negative controls. Más-o-menos was consistently on par with lasso, and even outperformed lasso in several cases. Its performance was much more similar to those of ridge and marginal regression. Screening did not dramatically affect the performances of any of the methods.

A referee noted that it is unclear how well más-o-menos performs within a single dataset, as opposed to across datasets. To answer this question, we evaluated the performance of each risk prediction algorithm within each dataset of each disease type by calculating the average three-fold cross-validated C-statistic at $\tau = 5$ years. No feature screening was implemented. Figure 3 reports the results and shows that más-o-menos was again on par or better than the other methods. It appears that in addition to being robust across studies, más-o-menos is also simply a good predictor.
Fig. 2. Validation C-statistics at τ = 5 years using different discrimination methods.
Fig. 3. Average three-fold cross-validation C-statistics at \( \tau = 5 \) years, calculated within each dataset of each disease type; no feature screening was implemented.

Table 2. Cancer gene expression datasets

<table>
<thead>
<tr>
<th>Reference</th>
<th>Sample size</th>
<th>Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bladder, 2463 common probesets</td>
<td>Als et al. (2007)</td>
<td>30</td>
</tr>
<tr>
<td>Blaveri et al. (2005)</td>
<td>80</td>
<td>44</td>
</tr>
<tr>
<td>Kim et al. (2010)</td>
<td>165</td>
<td>69</td>
</tr>
<tr>
<td>Lindgren et al. (2010)</td>
<td>87</td>
<td>26</td>
</tr>
<tr>
<td>Ræster et al. (2012)</td>
<td>93</td>
<td>65</td>
</tr>
<tr>
<td>Sjödahl et al. (2012)</td>
<td>224</td>
<td>25</td>
</tr>
<tr>
<td>Breast, 9768 common probesets</td>
<td>Desmedt et al. (2007)</td>
<td>134</td>
</tr>
<tr>
<td>Foekens et al. (2006)</td>
<td>710</td>
<td>191</td>
</tr>
<tr>
<td>Minn et al. (2005)</td>
<td>245</td>
<td>76</td>
</tr>
<tr>
<td>Minn et al. (2007); Wang et al. (2005)</td>
<td>209</td>
<td>80</td>
</tr>
<tr>
<td>Schmidt et al. (2008)</td>
<td>162</td>
<td>33</td>
</tr>
<tr>
<td>Sotiriou et al. (2006)</td>
<td>85</td>
<td>19</td>
</tr>
<tr>
<td>Symmans et al. (2010)</td>
<td>164</td>
<td>38</td>
</tr>
<tr>
<td>Ovarian, 6138 common probesets</td>
<td>Bentink et al. (2012)</td>
<td>128</td>
</tr>
<tr>
<td>Crnogorac et al. (2009)</td>
<td>98</td>
<td>72</td>
</tr>
<tr>
<td>Bonome et al. (2008)</td>
<td>185</td>
<td>129</td>
</tr>
<tr>
<td>Denkert et al. (2009)</td>
<td>41</td>
<td>13</td>
</tr>
<tr>
<td>Dressman et al. (2007)</td>
<td>59</td>
<td>36</td>
</tr>
<tr>
<td>Ferriss et al. (2012)</td>
<td>30</td>
<td>22</td>
</tr>
<tr>
<td>Konstantinopoulos et al. (2010)</td>
<td>28</td>
<td>17</td>
</tr>
<tr>
<td>Konstantinopoulos et al. (2010)</td>
<td>42</td>
<td>23</td>
</tr>
<tr>
<td>Mok et al. (2009)</td>
<td>53</td>
<td>41</td>
</tr>
<tr>
<td>Bell et al. (2011)</td>
<td>452</td>
<td>239</td>
</tr>
<tr>
<td>Tothill et al. (2008)</td>
<td>140</td>
<td>72</td>
</tr>
<tr>
<td>Yoshihara et al. (2010)</td>
<td>43</td>
<td>22</td>
</tr>
<tr>
<td>Yoshihara et al. (2012)</td>
<td>129</td>
<td>60</td>
</tr>
<tr>
<td>Yoshihara et al. (2012)</td>
<td>17</td>
<td>10</td>
</tr>
</tbody>
</table>

4 DISCUSSION

We have studied más-o-menos, a simple algorithm for classification and discrimination that has seen popular adoption but which has not been formally investigated. We gave a precise definition of the algorithm, showed theoretically and in simulations that it can perform well, and demonstrated in an extensive analysis of real cancer gene expression studies that it can indeed achieve good discrimination performance in realistic settings, even compared to lasso and ridge regression. Our results provide some justification to support its widespread use in practice. We hope our work will help shift the emphasis of ongoing prediction modeling efforts in genomics from the development of complex models to the more important issues of study design, model interpretation, and independent validation.

One reason why más-o-menos is comparable to more sophisticated methods such as penalized regression may be that we often use a prediction model trained on one set of patients to discriminate between subgroups in an independent sample, usually collected from a slightly different population and processed in a different laboratory. This cross-study variation is not captured by standard theoretical analyses, so theoretically optimal methods may not perform well in real applications (Hand, 2006). Bernau et al. (2014) proposed a method for giving a realistic measure of the practical utility of algorithms in the presence of cross-study variation. At the same time, we found using cross-validation experiments that even within the same dataset, más-o-menos remained competitive with more sophisticated methods.

Batch effects create study-specific measurement bias, and are widespread and often unidentified in genomic data (Leek et al., 2010). They may be responsible for the the cross-study variation that degrades the performance of algorithms such as lasso or ridge regression. Although certain batch-correction techniques have gained widespread use (Leek and Storey, 2007; Li and Rabinovic, 2007), these have been motivated primarily by class comparison rather than class prediction. In a genomic prediction competition batch correction was seen to provide no overall benefit for validation accuracy (MAQC Consortium, 2010). Rather, we propose that the impact of unknown batch effects may be best mitigated by using methods less prone to over-fitting. Més-o-menos risk scores have lower variability, and may be less associated with batch, than those of the other methods, which might explain its robust performance in both cross-validation and cross-study validation in 27 datasets from three cancer types.
While we focused on microarray data and survival endpoints, más-o-menos can be applied to any type of outcome variable, using any regression model, and has precedents for application in diverse settings outside of genomics (Wainer, 1976; Laughlin, 1978; Lovie and Lovie, 1986; Davis-Stober et al., 2010). It is fast to implement, simple to interpret, comparable in performance to more complex methods, and appears robust to cross-study variation. Más-o-menos should be useful for developing prediction models from high-dimensional data in any situation where the covariates are sufficiently correlated, and the true effect is roughly linear.

Funding: We thank the anonymous referees for comments which substantially improved this paper. This work was funded by the National Cancer Institute at the National Institutes of Health [1R01CA156551-01 and SP01 CA006516-46 to G.P.] and by the National Science Foundation [CAREER DBI-1053486 to C.H.]

**REFERENCES**


