BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 21 2014, pages 3062–3069
doi:10.1093/bioinformatics/btu488

Gene expression

Advance Access publication July 23, 2014

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]

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

Associate Editor: Ziv Bar-Joseph

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 used. 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 3833 patients with survival outcomes. We find that despite its simplicity, más-o-menos can achieve 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.

Availability and Implementation: Más-o-menos is implemented for survival analysis as an option in the survHD package, available from http://www.bitbucket.org/lwaldron/survhd and submitted to Bioconductor.

Contact: sdzhao@illinois.edu

Received on April 19, 2014; revised on June 23, 2014; accepted on July 11, 2014

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 (Bühlmann and Van De Geer, 2011; Hastie *et al*., 2005; Schölkopf and Smola, 2002); however, in practice many researchers 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 for standardized covariates 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 (Bell *et al*., 2011; Colman *et al*., 2010; Dave *et al*., 2004; Rème *et al*., 2013) 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 3833 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,

*To whom correspondence should be addressed.

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 $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$ 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 $\mathbf{X}_i$ to calculate a signed sum of that patient's covariate values. The procedure is as follows:

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

The factor of $p^{1/2}$ in the definition of the $\hat{v}_j$ merely serves to ensure the arbitrary scaling $||\hat{\mathbf{v}}||_2 = 1$. 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 $\mathbf{X}_i^T\hat{\mathbf{v}}$ can be quantified using correlation for continuous outcomes, the area under the receiver operating characteristic curve 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 experiments using standard datasets with uncensored outcomes. Some also use marginal regression to identify good and bad prognosis covariates, which are then used to rank patients by risk. Ranking methods include the t-statistic for difference in expression of good versus bad prognosis genes (Bell *et al.*, 2011; Verhaak *et al.*, 2013) and signed averaging of discretized or continuous expression values (Colman *et al.*, 2010; Dave *et al.*, 2004; Hallett *et al.*, 2010; Kang *et al.*, 2012; Rème *et al.*, 2013). Replacing lasso coefficients by their signs has been proposed for summarizing gene pathway activity (Eng *et al.*, 2013).

It may sometimes be helpful to perform an initial feature selection step before implementing más-o-menos, as we argue in Section 2.3. Feature selection has been the subject of a great deal of research, and a detailed discussion is beyond the scope of this article. In our data analysis in Section 3.3, we found that selection had little effect on the discrimination ability of más-o-menos. However, selection can provide more interpretable models by dramatically reducing the number of genes required for prediction.

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

$$C_\tau(\boldsymbol{\beta}) = P\{g(\mathbf{X}_i) > g(\mathbf{X}_j) \mid T_i < T_j, T_i < \tau\}, \quad (1)$$

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

$$\boldsymbol{\beta}_0 = \underset{\boldsymbol{\beta}:||\boldsymbol{\beta}_0||_2=1}{\arg\max}\, P(\mathbf{X}_i^T\boldsymbol{\beta} > \mathbf{X}_j^T\boldsymbol{\beta} | T_i < T_j, T_i < \tau),$$

where we have arbitrarily scaled $\boldsymbol{\beta}_0$ to have norm 1 because $C_\tau(\boldsymbol{\beta})$ is invariant to scaling of $\boldsymbol{\beta}$.

To implement más-o-menos in this setting, we will obtain the $\hat{\alpha}_j$ 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 $\hat{\alpha}_j$ converge to well-defined α_{0j} even when the Cox model is not correctly specified (Lin and Wei, 1989; Struthers and Kalbfleisch, 1986), 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_τ and should be a scalar multiple of the optimal $\boldsymbol{\beta}_0$.

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 directions 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, often little advantage can be gained in prediction accuracy over simple models. Here, we do not assume that the directions of effect are known.

Let $\mathbf{v}^* = (v_1^*, \ldots, v_p^*)^T$ be the probability limit of $\hat{\mathbf{v}}$, such that $\hat{\mathbf{v}} \to \mathbf{v}^*$. Because $\hat{v}_j = \mathrm{sgn}(\hat{\alpha}_j)$, if $\hat{\alpha}_j \to \alpha_{0j}$ in probability, then by the continuous mapping theorem $v_j^* = \mathrm{sgn}(\alpha_{0j})$. We will analyze the performance of the más-o-menos estimator $\hat{\mathbf{v}}$ in terms of the discrimination ability of $\mathbf{v}^*$ relative to that of $\boldsymbol{\beta}_0$, and the variability of $\hat{\mathbf{v}}$ around $\mathbf{v}^*$. For now, we assume $v_j^* \neq 0$ for all genes j. At the end of the section we discuss the implications if this is not true.

By the definition of C_τ, the discrimination performance of $\mathbf{v}^* = (v_1^*, \ldots, v_p^*)^T$ depends only on the degree of linear association between $\mathbf{X}_i^T\boldsymbol{\beta}_0$ and $\mathbf{X}_i^T\mathbf{v}^*$. In addition,

$$\mathrm{cov}(\mathbf{X}_i^T\boldsymbol{\beta}_0, \mathbf{X}_i^T\mathbf{v}^*) = \sum_{j,k}\beta_{0j}\mathrm{cov}(X_{ij}, X_{ik}v_k^*)$$

$$= \sum_j \beta_{0j}v_j^* \sum_k \mathrm{cov}(X_{ij}v_j^*, X_{ik}v_k^*) \geq \overline{\rho}\sum_j \beta_{0j}v_j^*,$$

where $\overline{\rho} = \min_j p^{-1}\sum_k \mathrm{cov}(X_{ij}v_j^*, X_{ik}v_k^*)$. The second equality follows because $v_j^* \cdot v_j^*$ always equals 1. Thus, $\mathbf{X}_i^T\mathbf{v}^*$ will be highly correlated with $\mathbf{X}_i^T\boldsymbol{\beta}_0$, and will have similar discriminative ability, under the condition that $\sum_j \beta_{0j}v_j^*$ and $\overline{\rho}$ have the same sign.

It is not unreasonable to expect these terms to be positive. First, each β_{0j} quantifies the association between X_{ij} and T_i conditional on all genes in $\mathbf{X}_i$, while each v_j^* reflects its univariate association. If a gene has the same direction of effect in both the conditional and marginal models, then $\beta_{0j}v_j^* > 0$. This is plausible for at least some genes, and even if it does not hold for all genes $\sum_j \beta_{0j}v_j^*$ can still be positive. Second, the $\overline{\rho}$ term is the minimum average covariance between $X_{ij}v_j^*$ and $X_{ik}v_k^*$. 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, $\overline{\rho}$ can be positive even if this covariance 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 $\hat{v}_j$ is given by

$$P(\hat{v}_j \neq v_j^*) = \begin{cases} P(\hat{\alpha}_j < 0) & \text{if } \alpha_{0j} > 0, \\ P(\hat{\alpha}_j > 0) & \text{if } \alpha_{0j} < 0, \\ P(\hat{\alpha}_j \neq 0) & \text{if } \alpha_{0j} = 0. \end{cases}$$

Lin and Wei (1989) showed that $\hat{\alpha}_j \rightarrow N(\alpha_{0j}, \sigma_j^2/n)$ for some α_{0j} and σ_j^2. This approximation, combined with Mill's inequality, gives the approximate relation

$$P(\hat{v}_j \neq v_j^*) \frac{\sigma_j}{n^{1/2}|\alpha_{0j}|\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{\alpha_{0j}^2 n}{\sigma_j^2}\right)$$

when $\alpha_{0j} \neq 0$, which approaches 0 much faster than $\text{var}(\hat{\alpha}_j)$. For large n and/or large $|\alpha_{0j}|$, the variability of $\hat{v}_j$ will be close to zero. Thus, $\hat{\mathbf{v}}$ is likely to be less susceptible to overfitting and, as a result, can have better out-of-sample discrimination performance.

Difficulties arise when $v_j^* = 0$ for some marginally unimportant genes j. First, $\text{cov}(\mathbf{X}_i^T\boldsymbol{\beta}_0, \mathbf{X}_i^T\mathbf{v}^*)$ 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, as $\hat{\alpha}_j$ 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 $\alpha_{0j} = 0$. 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 non-zero 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

Table 1. Average simulation runtimes

Method	$P = 100$	$P = 10\,000$
Lasso	8.914	47.238
Ridge	0.645	30.124
Marginal	0.016	2.209
Más-o-menos	0.023	2.408
Single	0.017	1.674
Random	0.001	0.004

Runtimes are reported in seconds

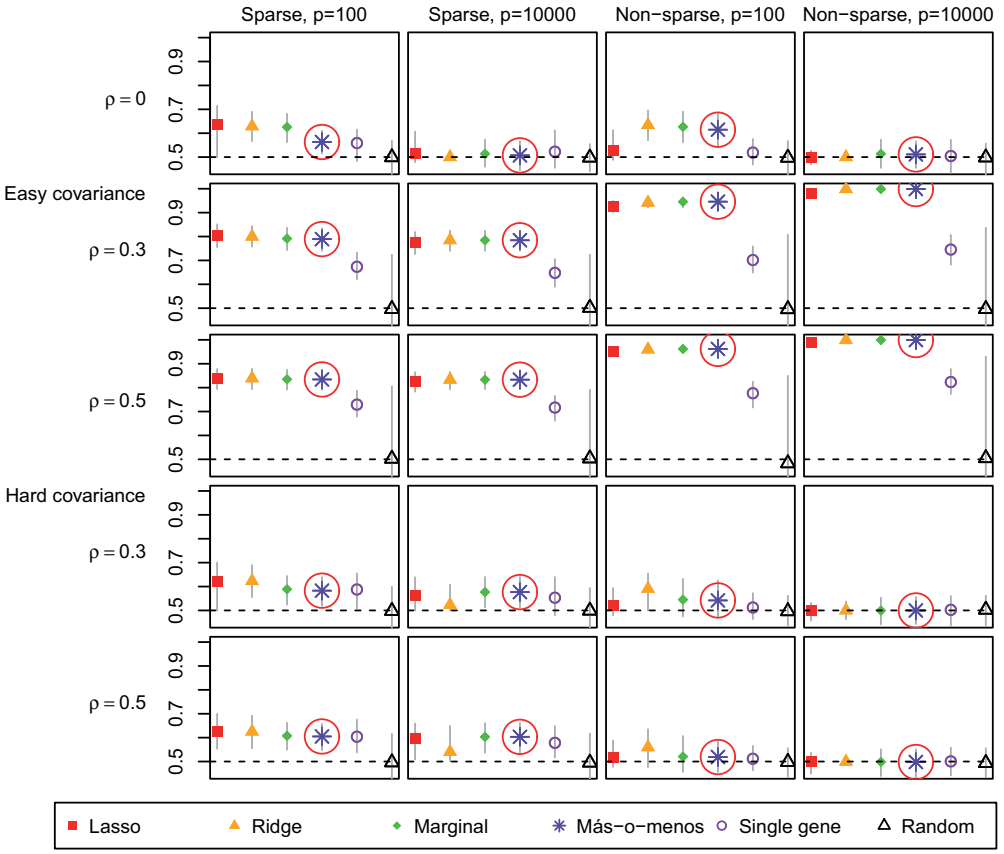

Fig. 1. Average validation C-statistics of different discrimination methods in simulated data. Más-o-menos results highlighted by circle. Vertical bars represent confidence intervals

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: (i) the single gene with the largest $\hat{\alpha}_j$ estimated from the training set and (ii) 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 3-fold cross-validation using the built-in function. Marginal Cox regressions and más-o-menos are implemented in the package survHD (Bernau *et al*., 2012).

3.2 Simulations

To simulate training data, we generated $p \times 1$ covariate vectors $\mathbf{X}_i$ and survival times from a Cox model with a $p \times 1$ true parameter vector $\boldsymbol{\beta}_0$. We let the true Cox regression coefficient vector $\boldsymbol{\beta}_0$ have s non-zero components all with magnitude $s^{-1/2}$, such that $\|\boldsymbol{\beta}_0\|_2 = 1$. The first $s/2$ non-zero components were positive and the rest were negative. We generated censoring times from an independent exponential distribution to give ~50% censoring. In each testing dataset, we replaced the positive entries of $\boldsymbol{\beta}_0$ 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 $P = 10\,000$. To generate sparse $\boldsymbol{\beta}_0$, we let $s = 10$, and for non-sparse $\boldsymbol{\beta}_0$ we let $s = P$. We drew $\mathbf{X}_i$ 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 $\mathbf{X}_i$. In an 'easy' setting, the covariates were divided into two blocks, with X_{ij} positively correlated within blocks and negatively correlated between blocks. Those X_{ij} with $\beta_{0j} > 0$ were assigned to one block, those with $\beta_{0j} < 0$ were assigned to the other and those with $\beta_{0j} = 0$ were assigned equally between the blocks. In a 'hard' setting, we let $\mathrm{cov}(X_{ij}, X_{ik}) > 0$ for j and k both even or both odd, and $\mathrm{cov}(X_{ij}, X_{ik}) < 0$ otherwise. We let $|\mathrm{cov}(X_{ij}, X_{ik})| = 0$, 0.3 or 0.5 for all j and k and ran 200 simulations.

The computations in this article were run on the Odyssey cluster supported by the Faculty of Arts and Sciences (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 τ 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: $\mathrm{cov}(X_{ij}v_j^*, T_i) > 0$ and $\mathrm{cov}(X_{ij}v_j^*, X_{ik}v_k^*) > 0$ 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, $\mathrm{cov}(X_{i1}, T_i) > 0$ and $\mathrm{cov}(X_{i2}, T_i) > 0$, but $\mathrm{cov}(X_{i1}, X_{i2}) < 0$. In other words, the signs of the β_{0j} 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 $\boldsymbol{\beta}_0$, 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 β_{0j} and the covariate 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 owing 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

Table 2. Cancer gene expression datasets

Reference	Sample size	Events
Bladder, 2463 common probesets		
Als *et al*. (2007)	30	25
Blaveri *et al*. (2005)	80	44
Kim *et al*. (2010)	165	69
Lindgren *et al*. (2010)	87	26
Riester *et al*. (2012)	93	65
Sjödahl *et al*. (2012)	224	25
Breast, 9768 common probesets		
Desmedt *et al*. (2007)	134	35
Foekens *et al*. (2006)	710	191
Minn *et al*. (2005)	245	76
Minn *et al*. (2007); Wang *et al*. (2005)	209	80
Schmidt *et al*. (2008)	162	33
Sotiriou *et al*. (2006)	85	19
Symmans *et al*. (2010)	164	38
Ovarian, 6138 common probesets		
Bentink *et al*. (2012)	128	73
Crijns *et al*. (2009)	98	72
Bonome *et al*. (2008)	185	129
Denkert *et al*. (2009)	41	13
Dressman *et al*. (2007)	59	36
Ferriss *et al*. (2012)	30	22
Konstantinopoulos *et al*. (2010)	28	17
Konstantinopoulos *et al*. (2010)	42	23
Mok *et al*. (2009)	53	41
Bell *et al*. (2011)	452	239
Tothill *et al*. (2008)	140	72
Yoshihara *et al*. (2010)	43	22
Yoshihara *et al*. (2012)	129	60
Yoshihara *et al*. (2012)	17	10

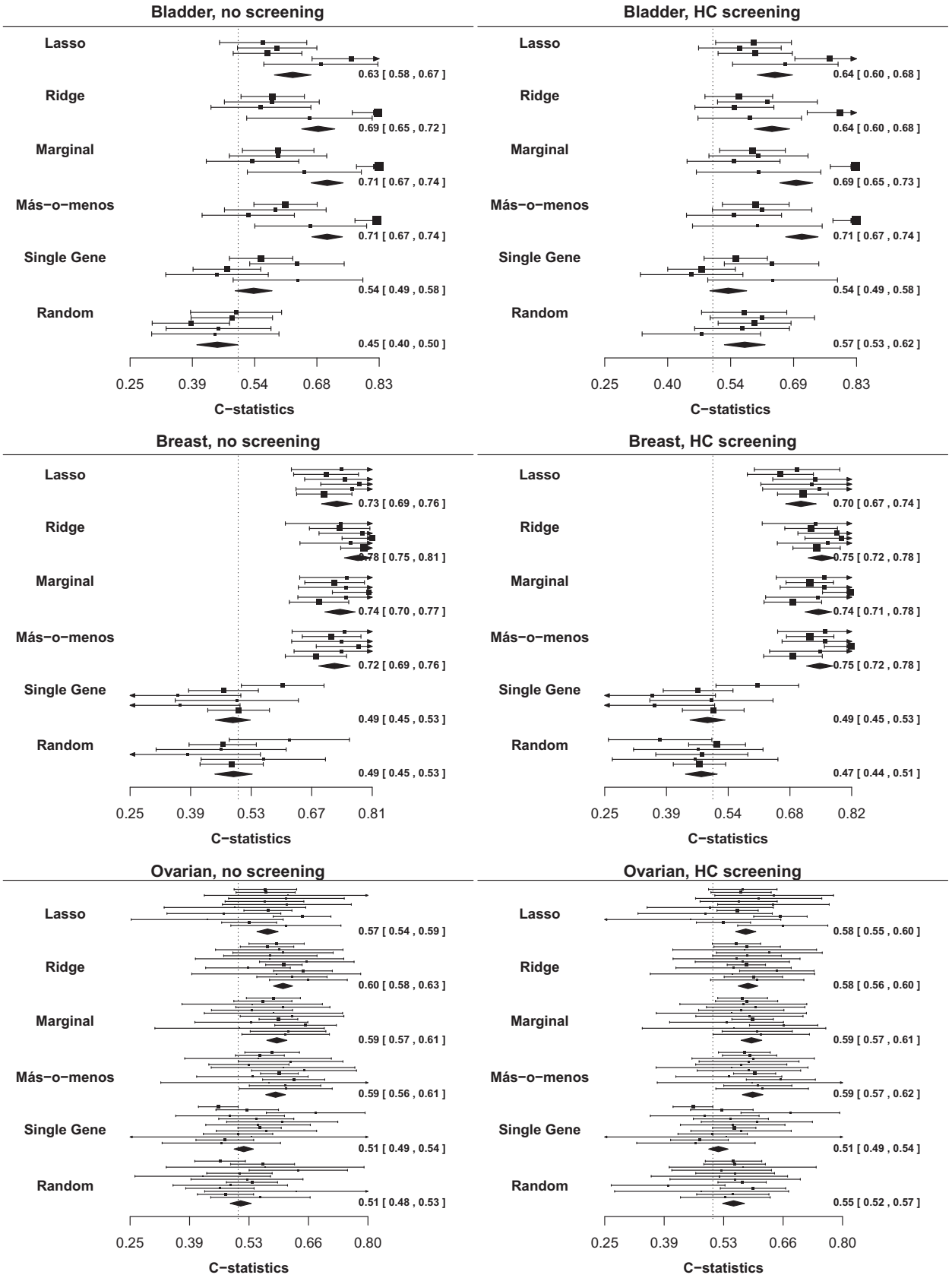

Fig. 2. Validation C-statistics at $\tau = 5$ years using different discrimination methods

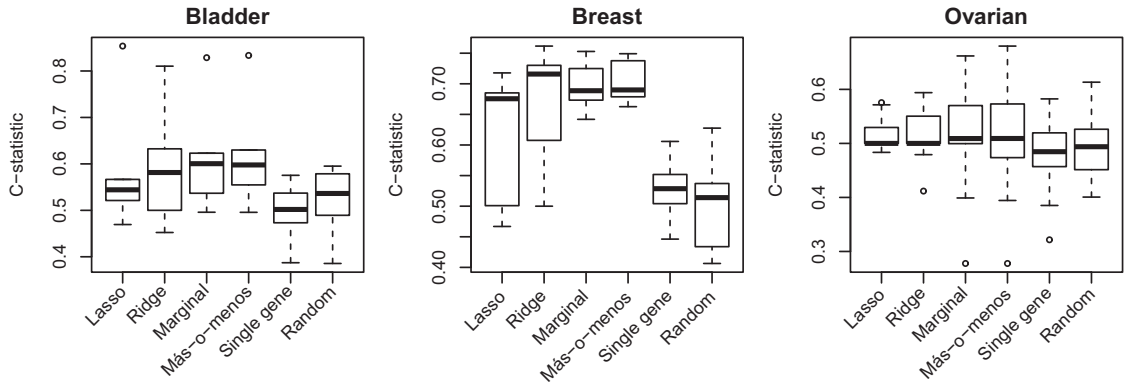

Fig. 3. Average 3-fold cross-validation C-statistics at $\tau = 5$ years, calculated within each dataset of each disease type; no feature screening was implemented

fewer important covariates with which the hard covariance structure could cause difficultly.

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 (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 τ is defined in (1). Roughly 60% of all study participants, combined across all diseases, were still alive after 5 years. The C-statistic is robust to the choice of τ unless few deaths or censoring events occur at times greater than τ (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 3-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.

4 DISCUSSION

We have studied más-o-menos, a simple algorithm for classification and discrimination that has seen popular adoption but 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 achieve good discrimination performance in realistic settings, even compared with 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 with 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 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 overfitting. Más-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 (Davis-Stober *et al*., 2010; Wainer, 1976; Laughlin, 1978; Lovie and Lovie, 1986). It is fast to implement, simple to interpret, comparable in performance with 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.

ACKNOWLEDGEMENT

The authors thank the anonymous referees for comments, which substantially improved this article.

Funding: This work was funded by the National Cancer Institute at the National Institutes of Health (1RC4CA156551-01 and 5P30 CA006516-46 to G.P.) and by the National Science Foundation (CAREER DBI-1053486 to C.H.).

Conflict of Interest: none declared.

REFERENCES

Als,A.B. *et al.* (2007) Emmprin and survivin predict response and survival following cisplatin-containing chemotherapy in patients with advanced bladder cancer. *Clin. Cancer Res.*, **13**, 4407–4414.

Bamber,D. (1975) The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J. Math. Psychol.*, **12**, 387–415.

Bell,D. *et al.* (2011) Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**, 609–615.

Bentink,S. *et al.* (2012) Angiogenic mRNA and microRNA gene expression signature predicts a novel subtype of serous ovarian cancer. *PloS One*, **7**, e30269.

Bernau,C. *et al.* (2012) *survHD: Synthesis of Microarray-based Survival Analysis*. R package version 0.5.0. https://bitbucket.org/lwaldron/survhd.

Bernau,C. *et al.* (2014) Cross-study validation for assessment of prediction models and algorithms. *Bioinformatics*, **30**, i105–i112.

Blaveri,E. *et al.* (2005) Bladder cancer outcome and subtype classification by gene expression. *Clin. Cancer Res.*, **11**, 4044–4055.

Bonome,T. *et al.* (2008) A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer. *Cancer Res.*, **68**, 5478–5486.

Bühlmann,P. and Van De Geer,S. (2011) *Statistics for High-dimensional Data: Methods, Theory and Applications*. Springer, Heidelberg.

Colman,H. *et al.* (2010) A multigene predictor of outcome in glioblastoma. *Neuro-oncology*, **12**, 49–57.

Crijns,A. *et al.* (2009) Survival-related profile, pathways, and transcription factors in ovarian cancer. *PLoS Med.*, **6**, e1000024.

Dave,S.S. *et al.* (2004) Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. *N. Engl. J. Med.*, **351**, 2159–2169.

Davis-Stober,C. *et al.* (2010) A constrained linear estimator for multiple regression. *Psychometrika*, **75**, 521–541.

Denkert,C. *et al.* (2009) A prognostic gene expression index in ovarian cancer-validation across different independent data sets. *J. Pathol.*, **218**, 273–280.

Desmedt,C. *et al.* (2007) Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series. *Clin. Cancer Res.*, **13**, 3207–3214.

Donoho,D. and Jin,J. (2008) Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proc. Nat Acad. Sci. USA*, **105**, 14790–14795.

Dressman,H. *et al.* (2007) An integrated genomic-based approach to individualized treatment of patients with advanced-stage ovarian cancer. *J. Clin. Oncol.*, **25**, 517–525.

Emura,T. *et al.* (2012) Survival prediction based on compound covariate under cox proportional hazard models. *PLoS One*, **7**, e47627.

Eng,K.H. *et al.* (2013) Pathway index models for construction of patient-specific risk profiles. *Stat. Med.*, **32**, 1524–1535.

Ferriss,J.S. *et al.* (2012) Multi-gene expression predictors of single drug responses to adjuvant chemotherapy in ovarian carcinoma: predicting platinum resistance. *PloS One*, **7**, e30550.

Foekens,J.A. *et al.* (2006) Multicenter validation of a gene expression–based prognostic signature in lymph node–negative primary breast cancer. *J. Clin. Oncol.*, **24**, 1665–1671.

Friedman,J.H. (1997) On bias, variance, 0/1loss, and the curse-of-dimensionality. *Data Min. Knowl. Discov.*, **1**, 55–77.

Friedman,J.H. *et al.* (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**, 1–22.

Ganzfried,B.F. *et al.* (2013) curatedOvarianData: clinically annotated data for the ovarian cancer transcriptome. *Database*, **2013**, bat013.

Haibe-Kains,B. *et al.* (2012) A three-gene model to robustly identify breast cancer molecular subtypes. *J. Nat Cancer Inst.*, **104**, 311–325.

Hallett,R.M. *et al.* (2010) An algorithm to discover gene signatures with predictive potential. *J. Exp. Clin. Cancer Res.*, **29**, 120.

Hand,D.J. (2006) Classifier technology and the illusion of progress. *Stat. Sci.*, **21**, 1–14.

Hastie,T. *et al.* (2005) The elements of statistical learning: data mining, inference and prediction. *Math. Intell.*, **27**, 83–85.

Hoerl,A. and Kennard,R. (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Kang,J. *et al.* (2012) A dna repair pathway-focused score for prediction of outcomes in ovarian cancer treated with platinum-based chemotherapy. *J. Nat Cancer Inst.*, **104**, 670–681.

Kim,W.J. *et al.* (2010) Predictive value of progression-related gene classifier in primary non-muscle invasive bladder cancer. *Mol. Cancer*, **9**, 3.

Konstantinopoulos,P. *et al.* (2010) Gene expression profile of BRCAness that correlates with responsiveness to chemotherapy and with outcome in patients with epithelial ovarian cancer. *J. Clin. Oncol.*, **28**, 3555–3561.

Laughlin,J.E. (1978) Comment on Estimating coefficients in linear models: it don't make no nevermind. *Psychol. Bull.*, **85**, 247–253.

Leek,J.T. and Storey,J.D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.*, **3**, e161–e161.

Leek,J.T. *et al.* (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.*, **11**, 733–739.

Li,C. and Rabinovic,A. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118–127.

Lin,D. and Wei,L. (1989) The robust inference for the Cox proportional hazards model. *J. Am. Stat. Assoc.*, **84**, 1074–1078.

Lindgren,D. *et al.* (2010) Combined gene expression and genomic profiling define two intrinsic molecular subtypes of urothelial carcinoma and gene signatures for molecular grading and outcome. *Cancer Res.*, **70**, 3463–3472.

Lovie,A. and Lovie,P. (1986) The flat maximum effect and linear scoring models for prediction. *J. Forecast.*, **5**, 159–168.

MAQC Consortium (2010) The microarray quality control (maqc)-ii study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.*, **28**, 827–865.

Minn,A.J. *et al.* (2005) Genes that mediate breast cancer metastasis to lung. *Nature*, **436**, 518–524.

Minn,A.J. *et al.* (2007) Lung metastasis genes couple breast tumor size and metastatic spread. *Proc. Nat Acad. Sci. USA*, **104**, 6740–6745.

Mok,S. *et al.* (2009) A gene signature predictive for outcome in advanced ovarian cancer identifies a survival factor: microfibril-associated glycoprotein 2. *Cancer Cell*, **16**, 521–532.

Rème,T. *et al.* (2013) Modeling risk stratification in human cancer. *Bioinformatics*, **29**, 1149–1157.

Riester,M. *et al.* (2012) Combination of a novel gene expression signature with a clinical nomogram improves the prediction of survival in high-risk bladder cancer. *Clin. Cancer Res.*, **18**, 1323–1333.

Schmidt,M. *et al.* (2008) The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res.*, **68**, 5405–5413.

Schölkopf,B. and Smola,A.J. (2002) *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT press.

Shaughnessy,J. *et al.* (2007) A validated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome 1. *Blood*, **109**, 2276–2284.

Sjödahl,G. *et al.* (2012) A molecular taxonomy for urothelial carcinoma. *Clin. Cancer Res.*, **18**, 3377–3386.

Sotiriou,C. *et al.* (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J. Nat Cancer Inst.*, **98**, 262–272.

Struthers,C. and Kalbfleisch,J. (1986) Misspecified proportional hazard models. *Biometrika*, **73**, 363–369.

Symmans,W.F. *et al.* (2010) Genomic index of sensitivity to endocrine therapy for breast cancer. *J. Clin. Oncol.*, **28**, 4111–4119.

Tibshirani,R.J. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B*, **58**, 267–288.

Tibshirani,R.J. (1997) The lasso method for variable selection in the Cox model. *Stat. Med.*, **16**, 385–395.

Tothill,R. *et al.* (2008) Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin. Cancer Res.*, **14**, 5198–5208.

Uno,H. *et al.* (2011) On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat. Med.*, **30**, 1105–1117.

van Noort,V. *et al.* (2003) Predicting gene function by conserved co-expression. *Trends Genet.*, **19**, 238–242.

Verhaak,R.G. *et al.* (2013) Prognostically relevant gene signatures of high-grade serous ovarian carcinoma. *J. Clin. Invest.*, **123**, 517.

Verweij,P. and Van Houwelingen,H. (1994) Penalized likelihood in cox regression. *Stat. Med.*, **13**, 2427–2436.

Viechtbauer,W. (2010) Conducting meta-analyses in R with the metafor package. *J. Stat. Softw.*, **36**, 1–48.

Wainer,H. (1976) Estimating coefficients in linear models: it don't make no nevermind. *Psychol. Bull.*, **83**, 213–217.

Waldron,L. *et al.* (2014) Comparative meta-analysis of prognostic gene signatures for Late-Stage ovarian cancer. *J. Nat Cancer Inst.*, **106,** pii: dju049.

Wang,Y. *et al.* (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, **365**, 671–679.

Yoshihara,K. *et al.* (2010) Gene expression profile for predicting survival in advanced-stage serous ovarian cancer across two independent datasets. *PloS One*, **5**, e9615.

Yoshihara,K. *et al.* (2012) High-risk ovarian cancer based on 126-gene expression signature Is uniquely characterized by downregulation of antigen presentation pathway. *Clin. Cancer Res.*, **18**, 1374–1385.