Harvard University
Harvard University Biostatistics Working Paper Series

Year 2013 *Paper* 158

Más-o-menos: A Simple Sign Averaging Method for Discrimination in Genomic Data Analysis

Sihai Dave Zhao[*] Giovanni Parmigiani[†]
Curtis Huttenhower[‡] Levi Waldron[**]

[*]University of Pennsylvania
[†]Harvard School of Public Health and Dana-Farber Cancer Institute, gp@jimmy.harvard.edu
[‡]Harvard School of Public Health, chuttenh@hsph.harvard.edu
[**]Harvard School of Public Health and Dana-Farber Cancer Institute

http://biostats.bepress.com/harvardbiostat/paper158

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[2,3]

[1]Department of Biostatistics and Epidemiology, University of Pennsylvania Perelman School of Medicine

[2]Department of Biostatistics, Harvard School of Public Health

[3]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute

May 1, 2013

Abstract

Simple classification approaches for clinical genomic data can often have discrimination power comparable to that of more complex procedures. In this paper we study an algorithm that has gained traction in practice but has not been thoroughly investigated in the statistical literature. This method calculates prognostic scores for discrimination by summing standardized predictors, weighted by the signs of their marginal associations with the outcome. We provide a formal definition of the method and study theoretical properties that explain why it can achieve surprisingly good discrimination in clinical genomics. We refer to this 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 non-optimal but still decent discrimination performance. Through simulations and a comprehensive analysis of gene expression datasets of ovarian tumors with survival information, we confirm that más-o-menos can match and even exceed the discrimination power of more established methods, with a significant advantage in speed, ease of implementation and interpretation, and reproducibility across clinical and technological settings.

Keywords: gene signatures; clinical genomics; prognostic modeling; risk score

1 Introduction

The successful translation of genomic signatures into clinical settings relies on good discrimination between patient subgroups that should receive different clinical management, and on ease of model implementation and interpretation. Whereas relatively sophisticated methods such as penalized regression, support vector machines, random forests, bagging and boosting have received detailed treatments in the statistics and machine learning literature (Hastie et al., 2005), many practitioners prefer simpler algorithms and prediction models (Hand, 2006). However, the operating characteristics of some of these simpler procedures have not yet been thoroughly investigated.

In this article we formalize and systematically investigate a family of simple yet robust classifiers for genomic data, where the coefficients of the linear risk score for standardized covariates are equal to the signs of the univariate associations with the clinical outcome of interest. In other words, the risk score is equal to number of “bad prognosis” features minus the number of “good prognosis” features. The procedure can also be preceded by a feature selection step to exclude irrelevant features from the analysis. We will show that in some circumstances this method can achieve quite good discrimination performance, comparable to that of more complex procedures. For this reason we refer to this 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 non-optimal but still decent performance.

Más-o-menos and closely related variants can be found in the top clinical, bioinformatic, and general science journals (Colman et al., 2010; Dave et al., 2004; Réme et al., 2013; Bell et al., 2011), and in commercially available prognostic gene signatures, such as the MyPRS Plus signature for multiple myeloma prognosis (Shaughnessy et al., 2007). Evidence of the method's effectiveness crosses disciplines: it has been observed in the credit scoring literature that large deviations from optimal weights in a linear predictor do not perform much worse than optimal (Lovie and Lovie, 1986), and in the psychometrics literature that equally weighting predictors in a linear regression gives nearly the same predictive accuracy as the least-squares regression coefficients (Wainer, 1976; Laughlin, 1978; Davis-Stober et al., 2010). 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).

There is, however, limited theoretical and computational investigation of the basis for simple average classifiers in the statistical genomics literature. Hand (2006) provides some theoretical arguments to justify equalization of regression coefficients when all covariates have the same direction of effect with 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 very simple models. Simple averaging has been shown to be useful for reducing variance in gene expression measurements prior to Lasso penalized regression (Park et al., 2007), and the replacement of Lasso coefficients by their signs has been proposed for summarizing gene pathway activity (Eng et al., 2013). However to the best of our knowledge, theoretical properties of más-o-menos for discrimination in high-dimensional data have not been described. Furthermore, situations in which it is or is not effective have not been discussed, and no systematic comparison relative to other prediction methods in clinical genomics has been made.

We aim to provide a formal definition of más-o-menos, investigate its theoretical properties, provide a systematic assessment relative to more well-established alternative methods, and establish the pre-conditions for effectiveness of this method. The paper is organized as follows: we first formally define the method in Section 2. We then study its theoretical properties for discrimination performance and variability in Section 3, which explain the results we see in our simulation studies in Section 4. Finally, we show that más-o-menos out-performs ridge and lasso regression, in terms of both speed and accuracy, for prediction of ovarian cancer prognosis in a meta-analysis of 14 datasets totalling 1,455 patients in Section 5. We discuss the implications of these findings in Section 6.

2 The más-o-menos method

Let X_{ij} be a quantitative measurement of the j^{th} gene from the i^{th} subject. The X_{ij} form the $p \times 1$ covariate vector $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$. Covariates could represent various types of genomic information, such as expression levels from microarrays or next-generation sequencing experiments, or non-genomic data.

The más-o-menos procedure a way to use a patient's $\mathbf{X}_i$ to calculate a linear risk score, a weighted sum of that patient's covariate values. The weights are generally estimated from a training dataset, and different methods for calculating linear risk scores result in different weights. Más-o-menos estimates the weights as follows:

1. Perform feature selection, for example using marginal screening (see remark below).
2. Standardize the retained covariates such that $(n-1)^{-1}\sum(X_{ij} - \bar{X}_j)^2 = 1, j = 1, \ldots, p$.
3. Perform univariate regressions, for each X_{ij}, on the outcome to obtain marginal estimates of the regression coefficient $\hat{\alpha}_j$.
4. Let $\hat{v}_j = \text{sgn}(\hat{\alpha}_j)/p^{1/2}$, where $\text{sgn}(c) = 2I(c > 0) - 1$ for $c \neq 0$ and $\text{sgn}(c) = 0$ for $c = 0$. For the j not retained by feature selection, $\hat{v}_j = 0$.

5. 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 discriminative 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 (AUC) for binary outcomes (Bamber, 1975), or the C-statistic for censored outcomes (Uno et al., 2011).

Remark 1. The feature selection step has been the subject of a great deal of recent research and a detailed discussion is beyond the scope of this paper. In the remainder of the paper we perform feature via marginal screening (Fan and Lv, 2008; Fan and Song, 2010; Zhao and Li, 2012) by selecting the X_{ij} with the lowest univariate regression p-values. It has been shown that under certain conditions, this screening procedure will retain all important covariates with high probability. We choose the optimal number of variables to retain by maximizing a cross-validated estimate of discrimination ability, such as the AUC or the C-statistic.

A variety of procedures closely related to más-o-menos have been previously proposed. These 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 vs. bad prognosis genes (Bell et al., 2011; Verhaak and Tamayo, 2013) and signed averaging of discretized or continuous expression values (Dave et al., 2004; Colman et al., 2010; Hallett et al., 2010; Kang et al., 2012; Réme et al., 2013). In contrast to these studies, we provide a more statistical analysis of the properties of más-o-menos.

3 Theoretical properties

In this paper we focus on survival outcomes, because they are typically the most difficult to deal with and the most clinically relevant. We show that under certain conditions, the más-o-menos weights can have fairly high discrimination power along with very low variability.

Let $\mathbf{v}^* = (v_1^*, \ldots, v_p^*)^T$ be the probability limit of $\hat{\mathbf{v}}$, such that $\hat{\mathbf{v}} \to \mathbf{v}^*$. Since $\hat{v}_j = \text{sgn}(\hat{\alpha}_j)$, if $\hat{\alpha}_j \to \alpha_{0j}$ in probability, then by the continuous mapping theorem $v_j^* = \text{sgn}(\alpha_{0j})$. Define $\mathcal{M} = \{j : \alpha_{0j} \neq 0\}$ to be the set of covariates marginally associated with T_i, and let $s = |\mathcal{M}|$. We assume that $\{j : \beta_{0j} = 0\} \subseteq \mathcal{M}$, which is crucial for any marginal regression-based procedure and is commonly made in the marginal screening literature.

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

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

where $g(\mathbf{X})$ is the risk score for a subject with covariate vector $\mathbf{X}$. While in general g can have any functional form, it is frequently taken to be linear in $\mathbf{X}$. We consider 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

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

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

While the C-statistic is a popular metric for quantifying discrimination, many authors have recognized that it is not very sensitive. In other words, there are cases where the difference in discriminative ability between two competing models must be very large in order to see a meaningful difference in their C-statistics. One increasingly popular alternative is the the Integrated Discrimination Improvement Index (IDI) (Pencina et al., 2008; Uno et al., 2009). On the other hand, Hilden and Gerds (2013) recently argued that using the IDI is not always safe. In the absence of a well-accepted alternative, we focus on the C-statistic in this paper.

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 (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_τ, and should be a scalar multiple of the optimal $\boldsymbol{\beta}_0$.

3.1 Discrimination performance

By the definition of C_τ, the discrimination performance of $\mathbf{v}^* = (v_1^*, \ldots, v_p^*)^T$ depends only on $\text{cor}(\mathbf{X}_i^T\boldsymbol{\beta}_0, \mathbf{X}_i^T\mathbf{v}^*)$. Under certain conditions, this correlation can be fairly high, so the discrimination performance of the $\mathbf{v}^*$ is not much worse than that of the optimal $\boldsymbol{\beta}_0$.

In particular,

$$\begin{aligned}
\text{cov}(\mathbf{X}_i^T\boldsymbol{\beta}_0, \mathbf{X}_i^T\mathbf{v}^*) &= \sum_{j,k\in\mathcal{M}} \beta_{0j}\,\text{cov}(X_{ij}, X_{ik}v_k^*) \\
&= \sum_{j\in\mathcal{M}} \beta_{0j}v_j^* \sum_{k\in\mathcal{M}} \text{cov}(X_{ij}v_j^*, X_{ik}v_k^*) \\
&\geq \bar{\rho} \sum_{j\in\mathcal{M}} \beta_{0j}v_j^*,
\end{aligned}$$

where $\bar{\rho} = \min_{j\in\mathcal{M}} |\mathcal{M}|^{-1} \sum_{k\in\mathcal{M}} \text{cov}(X_{ij}v_j^*, X_{ik}v_k^*)$. The second equality follows because $v_j^{*2} = 1$ for $j \in \mathcal{M}$. 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 $\sum_{j\in\mathcal{M}} \beta_{0j}v_j^*$ and $\bar{\rho}$ have the same sign.

It is not unreasonable to expect these terms to be positive. We first consider the term $\sum_{j\in\mathcal{M}} \beta_{0j}v_j^*$. 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 $\sum_{j\in\mathcal{M}} \beta_{0j}v_j^*$ can still be positive.

The $\bar{\rho}$ term is the minimum average correlation between $X_{ij}v_j^*$ and $X_{ik}v_k^*$ for $j, k \in \mathcal{M}$.

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. This is also not unreasonable, as the encoded proteins of conserved co-expressed gene pairs are likely to be part of the same biological pathway (van Noort et al., 2003). As before, $\bar{\rho}$ can be positive even if this correlation condition holds only for some pairs of genes, as we merely need the average correlation to be positive.

3.2 Variability

An additional appealing feature of más-o-menos is its low variability, which makes it more robust against overfitting and more reproducible across different datasets and technologies. 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 $\hat{\alpha}_j \to^{\mathcal{D}} N(\alpha_{0j}, \sigma_j^2/n)$ for some α_{0j} and σ_j^2. This approximation, combined with Mill's inequality, gives the approximate relation

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

for $\alpha_{0j} \neq 0$, 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 much closer to zero than the variability of marginal regression. Thus $\hat{\mathbf{v}}$ is likely to be more robust and less susceptible to overfitting, and as a result can have better out-of-sample discrimination performance.

One difficulty arises when $\alpha_{0j} = 0$. Since $\hat{\alpha}_j$ is a continuous estimator, $P(\hat{\alpha}_j \neq 0) = 1$ for any sample size. In other words, if más-o-menos is used on data where many of the covariates are not marginally associated with the outcome, it can actually be more variable

than the marginal estimator and may perform worse. This is the reason for the initial feature selection step, which should remove many unimportant covariates so that there are few j such that $\alpha_{0j} = 0$.

4 Simulations

We conducted simulations to compare más-o-menos and 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 generates risk scores of the form $\sum_j X_{ij}\hat{\alpha}_j$. In several configurations of dimensionality, regression parameter sparsity, covariance structure, and correlation strength, we found that más-o-menos had similar discriminative ability but was significantly faster to implement and execute than lasso and ridge regression.

4.1 Setup

For the lasso and ridge we standardized the covariates to have variance 1. We also included two negative controls: 1) the single gene with the largest $\hat{\alpha}_j$ 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 the built-in cross-validation function. Marginal Cox regressions can be performed very quickly with the function `rowCoxTests` found in the R package `survHD` (Bernau and Riester, 2012).

To generate a training dataset, on which we fit our regression methods, 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 $\boldsymbol{\beta}_0$ have s non-zero components all with magnitude $s^{-1/2}$, such that $\|\boldsymbol{\beta}_0\|_2 = 1$. The first $s/2$ nonzero components were positive and the rest were negative. We generated censoring times from an independent exponential distribution to give approximately 50%

censoring. Each simulation contained $n = 200$ independent observations.

To study the effect of dimensionality, we considered the low-dimensional case of $p = 50$ and the high-dimensional one of $p = 10000$. To generate sparse $\boldsymbol{\beta}_0$, we let $s = 10$, and for non-sparse $\boldsymbol{\beta}_0$ we let $s = p$. In Section 3.1 we saw that the bias of más-o-menos depends on the covariance structure of the $\mathbf{X}_i$, so we considered an "easy" structure and a "hard" structure. In the former, 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 the latter, we let $\text{cor}(X_{ij}, X_{ik}) > 0$ for j and k both even or both odd, and $\text{cor}(X_{ij}, X_{ik}) < 0$ otherwise, regardless of the signs of the corresponding β_{0j} and β_{0k}. We then drew $\mathbf{X}_i$ from a multivariate normal with unit marginal variance. Finally, we considered different levels of correlation, with $|\text{cor}(X_{ij}, X_{ik})| = 0$, 0.3, or 0.5 for all j and k.

We preceded each procedure, other than the method using the single best gene as the risk score, with a prescreening step that kept covariates with low marginal Cox regression p-values. We performed 3-fold cross-validation to choose the optimal number of covariates to retain: within each fold we implemented the screening step, fit the risk score, and calculated the C-statistic on the test fold using the method of Uno et al. (2011). We then picked the optimal number of retained covariates to maximize the average C-statistic across the folds.

4.2 Computation time

We first compared computation times for the different methods, averaged over 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. Marginal regression and más-o-menos could be more than 100 times faster than lasso and 15 times faster than ridge when $p = 100$. When $p = 10000$, más-o-menos was still faster by a factor of 10 to 20.

Table 1: Average simulation runtime in seconds

Method	$p = 100$	$p = 10000$
Lasso	98.7534	637.9471
Ridge	10.5475	337.5248
Marginal	0.7037	28.4332
Más-o-menos	0.8052	52.5774
Single	0.0167	1.6741
Random	0.5401	9.5072

4.3 Discriminative Ability by C-statistics

In addition to being very fast, más-o-menos can nearly match the more complicated procedures in discriminative ability. Performance of all methods is impacted by covariance structure, sparsity, and magnitude of associations, but 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. 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 3.1: $\text{cor}(X_{ij}v_j^*, T_i) > 0$ and $\text{cor}(X_{ij}v_j^*, X_{ik}v_k^*) > 0$ 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, $\text{cor}(X_{i1}, T_i) > 0$ and $\text{cor}(X_{i2}, T_i) > 0$, but $\text{cor}(X_{i1}, X_{i2}) < 0$. In other words, the signs of the β_{0j} and the directions of the covariate correlations are incoherent in the hard covariance case.

The effect of dimensionality depended on the sparsity and the covariance structure. 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, high dimensionality was still detrimental under the sparse setting, perhaps due to the difficulty of selecting the important covariates through

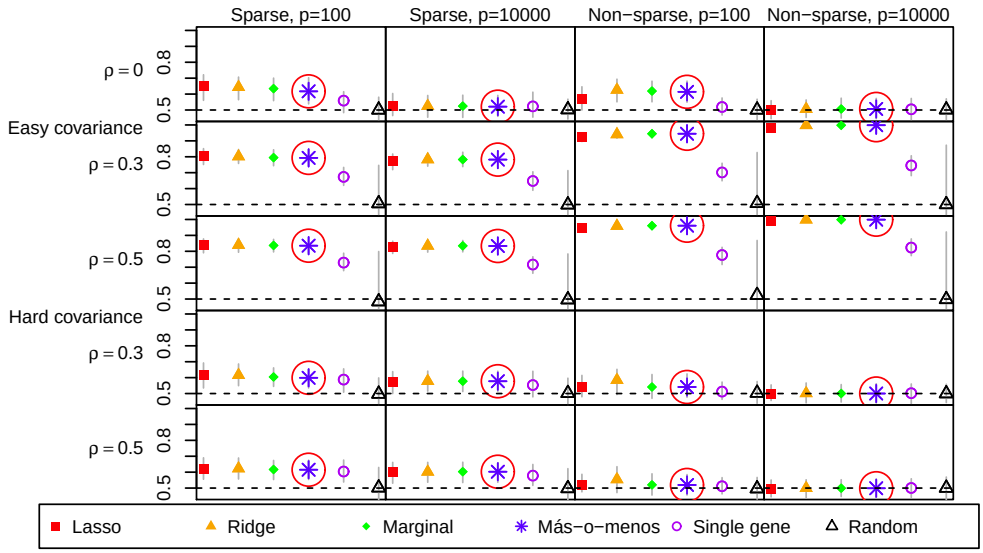

Figure 1: Average validation C-statistics of different discrimination methods in simulated data. Más-o-menos results highlighted by red circle.

the prescreening procedure. 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.

The impact of sparsity depended on the correlation structure. 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 difficultly.

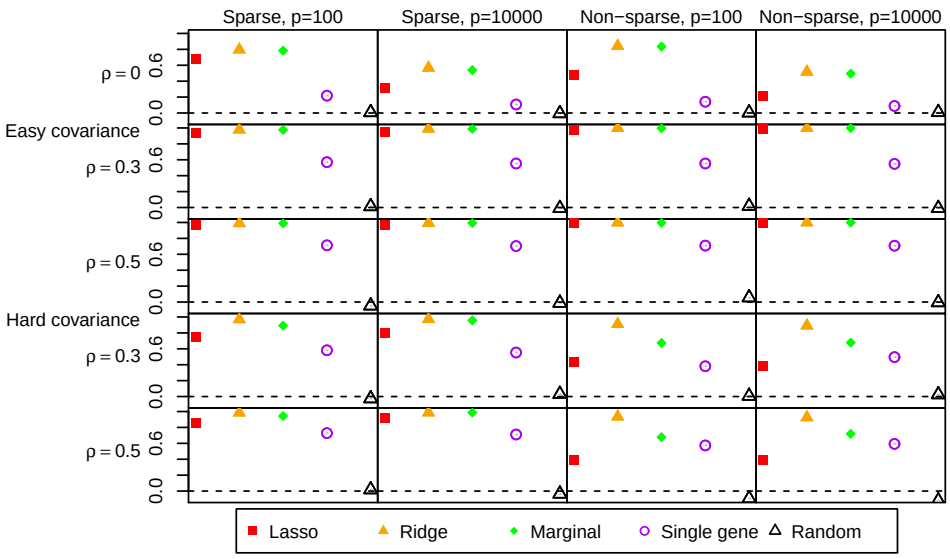

Figure 2: Average correlations of different discrimination methods with más-o-menos risk score in simulated data.

4.4 Similarity of risk scores

The question arises as to whether lasso, ridge, marginal regression, and más-o-menos give similar risk scores. To address this issue, we report in Figure 2 the average correlations between the más-o-menos risk score and the scores generated by the other methods. In general, we see that the results of all methods except for the negative controls are generally highly correlated. When $\rho = 0$, ridge and marginal regression were decently correlated with más-o-menos, though less so in high-dimensions. Lasso was not as highly correlated, perhaps because of its sparse estimates, but in these situations the discrimination performances of all methods were very low. Under the easy covariance setting, however, correlations between all methods except negative controls were close to 1, as they were with a hard covariance matrix and sparse $\boldsymbol{\beta}_0$. In the most difficult setting, a hard covariance matrix and non-sparse $\boldsymbol{\beta}_0$, the correlations with más-o-menos returned to the levels when $\rho = 0$.

Table 2: Datasets from Ganzfried et al. (2013), 6147 probesets in common

Name	Sample size	Events	Reference
E.MTAB.386	129	73	Bentink et al. (2012)
GSE13876	157	113	Crijns et al. (2009)
GSE14764	41	13	Denkert et al. (2009)
GSE17260	110	46	Yoshihara et al. (2010)
GSE18520	53	41	Mok et al. (2009)
GSE19829.GPL570	28	17	Konstantinopoulos et al. (2010)
GSE19829.GPL8300	42	23	Konstantinopoulos et al. (2010)
GSE26712	185	129	Bonome et al. (2008)
GSE32062.GPL6480	129	60	Yoshihara et al. (2012)
GSE32063	17	10	Yoshihara et al. (2012)
GSE9891	140	72	Tothill et al. (2008)
PMID17290060	59	36	Dressman et al. (2007)
PMID19318476	24	12	Berchuck et al. (2009)
TCGA	452	239	Bell et al. (2011)

5 Discrimination of ovarian cancer outcome

It is of great interest to distinguish between long and shorter-term survivors of ovarian cancer, so that the higher risk patients can be treated with alternative therapies. Ganzfried et al. (2013) have compiled and curated an extensive collection of publicly available ovarian cancer gene expression studies, and have standardized their clinical annotations, probeset identifiers, and microarray preprocessing. We use their datasets to train and validate prognostic scores for overall survival in a total of 1,445 patients from 14 datasets. This collection is available as a BioConductor package (Ganzfried et al., 2012).

We focused on patients with late stage, high grade serous tumors with available survival information, giving us 14 datasets from the Ganzfried et al. (2013) collection to work with (see Table 2). We limited our analysis to the 6138 probesets found in all 14 studies. We selected the largest available study (Bell et al., 2011) ("TCGA") for training, which had more than twice as many samples than any other single dataset. We standardized probesets to unit variance in the TCGA training dataset, then applied the scales determined from the training dataset to each test dataset. As in the simulations, we compared más-o-menos to lasso, ridge, marginal regression, the single best gene, and a randomly generated negative

control. We used these procedures to estimate risk score weights in the TCGA dataset, because it had the largest number of events, and we compared the C-statistics achieved by these different scoring systems on the remaining 13 datasets. For each procedure, again with the exception of the single best gene, we conducted univariate prescreening, choosing the number of covariates to retain using 3-fold cross-validation.

5.1 Validation C-statistics

Figure 3 displays a meta-analysis of the C-statistics from 13 validation datasets, for weights trained using the 452 samples of the TCGA dataset by each method. We used 100 bootstrap samples of each validation datset to obtain 95% confidence intervals. Uno et al. (2011) proposed a perturbation-resampling approach to estimating confidence intervals for the C-statistic, but their method is only valid for risk score weights obtained by fitting a Cox model. Summary statistics were calculated by fixed effects meta-analysis with the `metafor` R package (Viechtbauer, 2010).

Más-o-menos achieved the highest estimated C-statistics in eight out of the 13 datasets. Lasso had the highest C-statistics in two datasets, as did marginal regression, and ridge had the highest in one dataset. In addition, más-o-menos was about 10 times faster than ridge and 30 times faster than lasso. As expected, lasso, ridge, marginal regression, and más-o-menos outperformed the single best gene in most cases, which in turn outperformed the random score.

5.2 Association with Batch Effects

To help explain the good performance of más-o-menos in experimental data, we looked at the influence of batch effects (date of array hybridization) in the training TCGA dataset. Table 3 provides one-way analysis of variance results for the risk scores of each method with respect to batch in the TCGA training dataset. Each F-statistic, equal to the ratio of the between-batch variance and the within-batch variance, quantifies the amount of variability

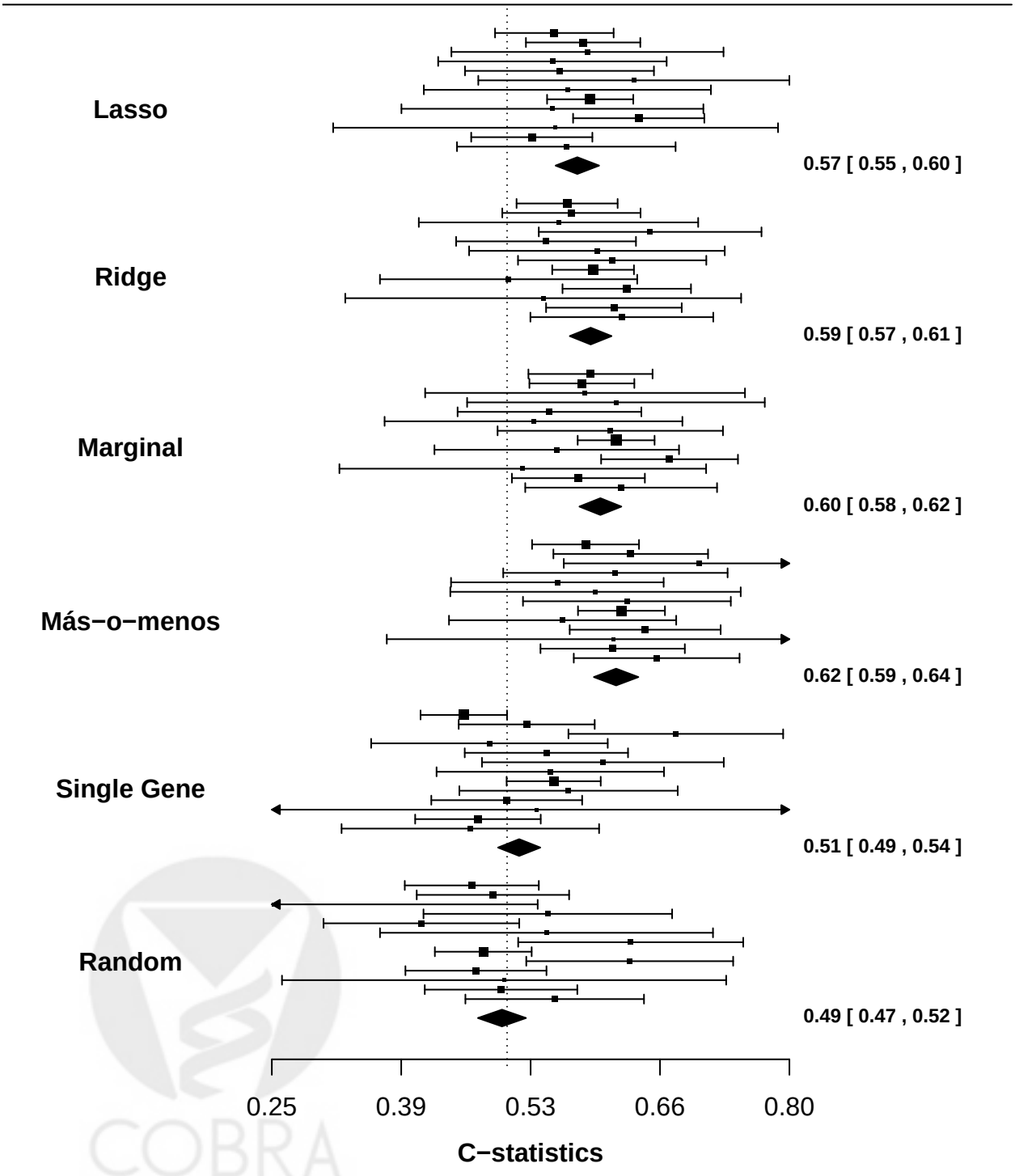

Figure 3: Validation C-statistics at 5 years using different discrimination methods in ovarian cancer datasets.

Table 3: Analysis of Variance of risk scores for the predictions of each model with respect to batches in the TCGA training dataset.

Method	F-statistic ($F_{13,437}$)	p-value
Lasso	3.0	$2.6 \cdot 10^{-4}$
Ridge	3.4	$4.5 \cdot 10^{-5}$
Marginal	2.4	$3.8 \cdot 10^{-3}$
Más-o-menos	2.7	$1.0 \cdot 10^{-3}$
Single	3.9	$5.4 \cdot 10^{-6}$
Random	1.1	$3.5 \cdot 10^{-1}$

due to batch effect. The top single gene was most strongly associated with batch, followed by ridge and lasso. In contrast, marginal regression and más-o-menos were much more weakly affected. This is mostly likely a reflection of the heavily regularized nature of the marginal and más-o-menos estimators, which are more biased than lasso and ridge in the training set but can therefore deliver better results in the testing datasets.

6 Discussion

In this paper we study theoretical and practical properties of más-o-menos for prediction from genomic data. This method has gained popularity in practice for its simplicity and effectiveness, but has little foundation in the statistical literature. We provide theoretical arguments showing that más-o-menos can achieve good discrimination performance and is more robust to overfitting relative to standard methods. These theoretical findings are supported empirically both by simulation studies and by an analysis of microarray data from 14 ovarian cancer microarray studies, where we found that más-o-menos offered prediction performance comparable to or better than lasso and ridge regression, while providing significant advantages in speed and simplicity.

While we focused on microarray data and survival endpoints, más-o-menos can be applied to outcomes of any type, using any regression model, and has precedent for application in diverse settings outside the field of genomics (Wainer, 1976; Laughlin, 1978; Davis-Stober et al., 2010; Lovie and Lovie, 1986). It is fast to implement, simple to understand, comparable in

performance to far more complex methods, and is robust to variation in study-specific features. 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.

Batch effects create study-specific measurement bias, and are widespread and often unidentified in genomic data (Leek et al., 2010). 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 for several endpoints involving numerous research groups and methods, batch correction was seen to provide no overall benefit for validation accuracy (MAQC Consortium, 2010). Rather, we propose that the degrading impact of batch effects on prediction models is best mitigated by methods that are relatively robust to over-fitting and to batch effects. Más-o-menos and marginal regression risk scores were less associated with batch in the TCGA training dataset than lasso and ridge regression, which were in turn less associated than the single gene with strongest survival association (Table 3). The high correlation between the best single gene with training set batch effects, in contrast to the relative insensitivity of más-o-menos to batch effects, provides a likely explanation for the superior prediction performance of más-o-menos across 13 independent validation datasets.

We also compared más-o-menos to marginal regression, which is equally fast and nearly as simple. In our simulations in Section 4, differences between the two methods were not pronounced, but más-o-menos achieved higher C-statistics than marginal regression on nearly all of the ovarian cancer validation datasets. This is likely due to the lower variability of más-o-menos, as discussed in Section 3.2, making it less likely to be affected by study-specific differences between training and test datasets.

We preceded each method with univariate feature selection. We performed additional simulations, not reported here in detail, indicating that feature selection slightly improved the performance of más-o-menos, as suggested by Section 3.2. It is well-known that univari-

ate feature selection has drawbacks when important and unimportant covariates are highly correlated. More complicated techniques such as iterative screening (Fan and Lv, 2008), a type of stepwise variable selection, might improve the performance of the subsequent más-o-menos procedure. On the other hand, these techniques are more difficult to implement, their theoretical properties are not well-understood, and they may not improve the final results by much, as our work indicates that univariate feature selection already performs well in practice. We hope this 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.

Funding

This work was funded by the National Cancer Institute at the National Institutes of Health [1RC4CA156551-01 to G.P.] and by the National Science Foundation [CAREER DBI-1053486 to C.H.].

References

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* **12,** 387–415.

Bell, D., Berchuck, A., Birrer, M., Chien, J., Cramer, D., Dao, F., Dhir, R., DiSaia, P., Gabra, H., Glenn, P., et al. (2011). Integrated genomic analyses of ovarian carcinoma**474,** 609–615.

Bentink, S., Haibe-Kains, B., Risch, T., Fan, J., Hirsch, M., Holton, K., Rubio, R., April, C., Chen, J., Wickham-Garcia, E., et al. (2012). Angiogenic mRNA and microRNA Gene Expression Signature Predicts a Novel Subtype of Serous Ovarian Cancer. *PloS One* **7,** e30269.

Berchuck, A., Iversen, E., Luo, J., Clarke, J., Horne, H., Levine, D., Boyd, J., Alonso, M., Secord, A., Bernardini, M., et al. (2009). Microarray analysis of early stage serous ovarian cancers shows profiles predictive of favorable outcome. *Clinical Cancer Research* **15,** 2448–2455.

Bernau, C. Waldron, L. and Riester, M. (2012). *survHD: Synthesis o microarray-based survival analysis*. R package version 0.5.0, https://bitbucket.org/lwaldron/survhd.

Bonome, T., Levine, D., Shih, J., Randonovich, M., Pise-Masison, C., Bogomolniy, F., Ozbun, L., Brady, J., Barrett, J., Boyd, J., et al. (2008). A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer. *Cancer Research* **68,** 5478–5486.

Colman, H., Zhang, L., Sulman, E. P., McDonald, J. M., Shooshtari, N. L., Rivera, A., Popoff, S., Nutt, C. L., Louis, D. N., Cairncross, J. G., Gilbert, M. R., Phillips, H. S., Mehta, M. P., Chakravarti, A., Pelloski, C. E., Bhat, K., Feuerstein, B. G., Jenkins, R. B., and Aldape, K. (2010). A multigene predictor of outcome in glioblastoma. *Neuro-oncology* **12,** 49–57.

Crijns, A., Fehrmann, R., De Jong, S., Gerbens, F., Meersma, G., Klip, H., Hollema, H., Hofstra, R., Te Meerman, G., de Vries, E., et al. (2009). Survival-related profile, pathways, and transcription factors in ovarian cancer. *PLoS Medicine* **6,** e1000024.

Dave, S., Wright, G., and Tan, B. (2004). Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. *New England Journal of Medicine* pages 2159–2169.

Davis-Stober, C., Dana, J., and Budescu, D. (2010). A constrained linear estimator for multiple regression. *Psychometrika* **75,** 521–541.

Denkert, C., Budczies, J., Darb-Esfahani, S., Györffy, B., Sehouli, J., Könsgen, D., Zeillinger, R., Weichert, W., Noske, A., Buckendahl, A., et al. (2009). A prognostic gene expression

index in ovarian cancer-validation across different independent data sets. *The Journal of Pathology* **218,** 273–280.

Dressman, H., Berchuck, A., Chan, G., Zhai, J., Bild, A., Sayer, R., Cragun, J., Clarke, J., Whitaker, R., Li, L., et al. (2007). An integrated genomic-based approach to individualized treatment of patients with advanced-stage ovarian cancer. *Journal of Clinical Oncology* **25,** 517–525.

Emura, T., Chen, Y.-H., and Chen, H.-Y. (2012). Survival prediction based on compound covariate under cox proportional hazard models. *PLoS ONE* **7,** e47627.

Eng, K. H., Wang, S., Bradley, W. H., Rader, J. S., and Kendziorski, C. (2013). Pathway index models for construction of patient-specific risk profiles. *Statistics in medicine* .

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society, Ser. B* **70,** 849–911.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models and NP-dimensionality. *The Annals of Statistics* **38,** 3567–3604.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33,** 1–22.

Ganzfried, B., Riester, M., Haibe-Kains, B., Risch, T., Tyekucheva, S., Jazic, I., Wang, X., Ahmadifar, M., Birrer, M., Parmigiani, G., Huttenhower, C., and Waldron, L. (2013). curatedOvarianData: Clinically Annotated Data for the Ovarian Cancer Transcriptome.

Ganzfried, B., Waldron, L., Skates, S., Riester, M., Wang, V., Risch, T., Haibe-Kains, B., Huttenhower, C., Tyekucheva, S., Ding, J., Jazic, I., Birrer, M., and Parmigiani, G. (2012). *curatedOvarianData: Ovarian Cancer Gene Expression Analysis*. R package version 0.3.0, http://bcb.dfci.havard.edu/∼gp/.

Hallett, R. M., Dvorkin, A., Gabardo, C. M., and Hassell, J. a. (2010). An algorithm to discover gene signatures with predictive potential. *Journal of experimental & clinical cancer research : CR* **29,** 120.

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

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer* **27,** 83–85.

Hilden, J. and Gerds, T. A. (2013). A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. *Statistics in medicine* .

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

Kang, J., D'Andrea, A. D., and Kozono, D. (2012). A dna repair pathway-focused score for prediction of outcomes in ovarian cancer treated with platinum-based chemotherapy. *Journal of the National Cancer Institute* **104,** 670–681.

Konstantinopoulos, P., Spentzos, D., Karlan, B., Taniguchi, T., Fountzilas, E., Francoeur, N., Levine, D., and Cannistra, S. (2010). Gene expression profile of BRCAness that correlates with responsiveness to chemotherapy and with outcome in patients with epithelial ovarian cancer. *Journal of Clinical Oncology* **28,** 3555–3561.

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

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics* **11,** 733–739.

Leek, J. T. and Storey, J. D. (2007). Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genet* **3,** e161–e161.

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. *Journal of the American Statistical Association* pages 1074–1078.

Lovie, A. and Lovie, P. (1986). The flat maximum effect and linear scoring models for prediction. *Journal of Forecasting* **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. *Nature Biotechnology* **28,** 827–865.

Mok, S., Bonome, T., Vathipadiekal, V., Bell, A., Johnson, M., Park, D., Hao, K., Yip, D., Donninger, H., Ozbun, L., 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.

Park, M. Y., Hastie, T., and Tibshirani, R. (2007). Averaged gene expressions for regression. *Biostatistics* **8,** 212–227.

Pencina, M., D'Agostino Sr, R., D'Agostino Jr, R., and Vasan, R. (2008). Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine* **27,** 157–172.

Réme, T., Hose, D., Theillet, C., and Klein, B. (2013). Modeling Risk Stratification in Human Cancer. *Bioinformatics (Oxford, England)* pages 1–9.

Shaughnessy, J., Zhan, F., Burington, B. E., Huang, Y., Colla, S., Hanamura, I., Stewart, J. P., Kordsmeier, B., Randolph, C., Williams, D. R., 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.

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

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Ser. B* **58,** 267–288.

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

Tothill, R., Tinker, A., George, J., Brown, R., Fox, S., Lade, S., Johnson, D., Trivett, M., Etemadmoghadam, D., Locandro, B., et al. (2008). Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clinical Cancer Research* **14,** 5198–5208.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L. J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine* **30,** 1105–1117.

Uno, H., Tian, L., Cai, T., Kohane, I., and Wei, L. (2009). Comparing risk scoring systems beyond the ROC paradigm in survival analysis.

van Noort, V., Snel, B., Huynen, M. A., et al. (2003). Predicting gene function by conserved co-expression. *TRENDS in Genetics* **19,** 238–242.

Verhaak, R. and Tamayo, P. (2013). Prognostically relevant gene signatures of high-grade serous ovarian carcinoma. *The Journal of Clinical Investigation* pages 1–9.

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

Viechtbauer, W. (2010). Conducting meta-analyses in r with the metafor package. *Journal of Statistical Software* **36,** 1–48.

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

Yoshihara, K., Tajima, A., Yahata, T., Kodama, S., Fujiwara, H., Suzuki, M., Onishi, Y., Hatae, M., Sueyoshi, K., Fujiwara, H., 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., Tsunoda, T., Shigemizu, D., Fujiwara, H., Hatae, M., Fujiwara, H., Masuzaki, H., Katabuchi, H., Kawakami, Y., Okamoto, A., et al. (2012). High-risk ovarian cancer based on 126-gene expression signature Is uniquely characterized by downregulation of antigen presentation pathway. *Clinical Cancer Research* **18,** 1374–1385.

Zhao, S. D. and Li, Y. (2012). Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis* **105,** 397–411.