Balancing the robustness and predictive performance of biomarkers

Paul Kirk, Aviva Witkover, Charles R. M. Bangham, Sylvia Richardson,

Alexandra M. Lewin* and Michael P. H. Stumpf*

* to whom correspondence should be addressed

Abstract

Recent studies have highlighted the importance of assessing the robustness of putative biomarkers identified from experimental data. This has given rise to the concept of *stable* biomarkers, which are ones that are consistently identified regardless of small perturbations to the data. Since stability is not by itself a useful objective, we present a number of strategies that combine assessments of stability and predictive performance in order to identify biomarkers that are both robust and diagnostically useful. Moreover, by wrapping these strategies around logistic regression classifiers regularised by the elastic net penalty, we are able to assess the effects of correlations between biomarkers upon their perceived stability.

We use a synthetic example to illustrate the properties of our proposed strategies. In this example, we find that: (i) assessments of stability can help to reduce the number of false positive biomarkers, although potentially at the cost of missing some true positives; (ii) combining assessments of stability with assessments of predictive performance can improve the true positive rate; and (iii) correlations between biomarkers can have adverse effects on their stability, and hence must be carefully taken into account when undertaking biomarker discovery. We then apply our strategies in a proteomics context, in order to identify a number of robust candidate biomarkers for the human disease HTLV1-associated myelopathy/tropical spastic paraparesis (HAM/TSP).

1 Introduction

Several recent articles have emphasised the importance of considering the *stability* of gene signatures and biomarkers of disease identified by feature selection algorithms (see, for example, Zucknick *et al.*, 2008; Meinshausen and Bühlmann, 2010; Abeel *et al.*, 2010; Alexander and Lange, 2011; Ahmed *et al.*, 2011). The aim is to establish if the selected predictors are specific to the particular dataset that was observed, or if they are robust to the noise in the data. Although not a new concept (see, for example, Turney, 1995, for an early discussion), selection stability has received a renewed interest in biological contexts due to concerns over the irreproducibility of results (Ein-Dor *et al.*, 2005, 2006). Assessments of stability usually proceed by: (i) subsampling the original dataset; (ii) applying a feature selection algorithm to each subsample; and then (iii) quantifying stability using a method for assessing the agreement among the resulting sets of selections (e.g. Kalousis *et al.*, 2007; Kuncheva, 2007; Jurman *et al.*, 2008). There is an increasing body of literature on this subject, and we refer the reader to He and Yu (2010) for a comprehensive review.

One of the principal difficulties with stability is that it is not by itself a useful objective: a selection strategy that chooses an arbitrary fixed set of covariates regardless of the observed data will achieve perfect stability, but the predictive performance provided by the selected set is likely to be poor (Abeel *et al.*, 2010). Since we ultimately seek biomarkers that are not only robust but which also allow us to discriminate between (for example) different disease states, it is desirable to try to optimise both stability and predictive performance simultaneously. The first contribution of this ar-

ticle is to present a number of strategies for doing this. We follow Meinshausen and Bühlmann (2010) in estimating selection probabilities for different sets of covariates, but diverge from their approach by combining these estimates with assessments of predictive performance. Given that our approach uses subsampling for both model structure estimation and performance assessment, it is somewhat related to double cross validation (see Stone, 1974, and also Smit *et al.*, 2007 for an application similar to the one considered here); however, we do not employ a nested subsampling step.

Our second contribution is to provide a procedure for quantifying the effects of correlation upon selection stability. As discussed in Yu *et al.* (2008), correlations among covariates can have a serious impact upon stability. Since multivariate covariate selection strategies often seek a minimal set of covariates that yield the best predictive performance, a single representative from a group of correlated covariates is often selected in favour of the whole set. This can have a negative impact upon stability (Kirk *et al.*, 2010), as the selected representative is liable to vary from dataset to dataset. We hence consider a covariate selection strategy based upon logistic regression with the elastic net likelihood penalty (see Zou and Hastie, 2005; Friedman *et al.*, 2007, 2010, and Section 2.4), which allows us to control whether we tend to select single representatives or whole sets of correlated covariates. This allows us to investigate systematically how our treatment of correlation affects stability.

2 Methods

Let D be a dataset comprising observations taken on n individuals, $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$. Each $\mathbf{x}_i = [x_{i1}, \ldots, x_{ip}]^\top \in \mathbb{R}^p$ is a vector of measurements taken upon p covariates $v_1, \ldots, v_p$, and $y_i \in \{0, 1\}$ is a corresponding binary class label (e.g. case/control). A *classification rule* is a function, h, such that $h(\mathbf{x}) \in \{0, 1\}$ is the predicted class

label for $\mathbf{x} \in \mathbb{R}^p$. For the time being, we assume only that h was obtained by fitting some predictive model $\mathcal{H}_{\boldsymbol{\theta}}$ to a training dataset (here, $\boldsymbol{\theta}$ denotes the parameters of the model). We write $\mathcal{H}_{\boldsymbol{\theta}}(D)$ to denote the fitted model obtained by training $\mathcal{H}_{\boldsymbol{\theta}}$ on dataset D.

2.1 Assessing predictive performance

Given dataset D and classification rule h, we can calculate the *correct classification rate* when h is applied to D as the proportion of times the predicted and observed class labels are equal,

$$c(D; h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[h(\mathbf{x}_i) = y_i], \tag{1}$$

where $\mathbb{I}(Z)$ is the indicator function, which equals 1 if Z is true and 0 otherwise.

One approach for assessing the predictive performance of model $\mathcal{H}_{\boldsymbol{\theta}}$ is random subsampling cross validation (Kohavi, 1995). We train our predictive model on a subsample, D_k, of the training dataset, and then calculate the correct classification rate, c_k, when the resulting classifier is applied to the remaining (left-out) data, $D_{\backslash k} = D \backslash D_k$. Repeating for $k = 1, \ldots, K$, we may calculate the mean correct classification rate and take this as an estimate of the probability that our model classifies correctly,

$$\hat{\mathbb{P}}(\{\text{classify correctly}\}|\mathcal{H}_{\boldsymbol{\theta}}) = \frac{1}{K} \sum_{k=1}^{K} c_k. \tag{2}$$

2.2 Assessing stability

We suppose that – as well as a classification rule – we also obtain a set of selected covariates, s_k, when we train $\mathcal{H}_{\boldsymbol{\theta}}$ on subsample D_k. More precisely, we assume that only the covariates in s_k appear with non-zero coefficients in the fitted predictive model $\mathcal{H}_{\boldsymbol{\theta}}(D_k)$ (for example, this will be the case if we fit logistic regression models with lasso

or elastic net likelihood penalties). For any subset of the covariates, $V \subseteq \{v_1, \ldots, v_p\}$, we may then estimate the probability that the covariates in V are among those selected,

$$\hat{\mathbb{P}}(\{\text{select } V\}|\mathcal{H}_{\boldsymbol{\theta}}) = \frac{1}{K}\sum_{k=1}^{K}\mathbb{I}(V \subseteq s_k). \tag{3}$$

This quantifies the *stability* with which the covariate set V is selected (Meinshausen and Bühlmann, 2010).

2.3 Combining stability and predictive performance

Equation (2) provides an assessment of predictive performance, but gives no information regarding whether or not there is any agreement among the selected sets s_k. On the other hand, Equation (3) allows us to assess the stability of a covariate set V, but does not tell us if the covariates in V are predictive. Since these assessments of stability and predictive performance both require us to subsample the training data, it seems natural to combine them in order to try to resolve their limitations. We here provide a method for doing this.

We shall henceforth assume that the parameters, $\boldsymbol{\theta}$, of $\mathcal{H}_{\boldsymbol{\theta}}$ may be tuned in order to ensure that precisely m covariates are selected. We then write s_{mk} for the selected set of size m obtained when $\mathcal{H}_{\boldsymbol{\theta}}$ is trained on D_k, and h_{mk} for the corresponding classification rule. Similarly, we define $c_{mk} = c(D_{\backslash k}; h_{mk})$. Figure 1 provides a summary of this notation and the way in which we find s_{mk} and c_{mk}. Having made these definitions, we may additionally condition on m in Equations (2) and (3) to obtain,

$$\hat{\mathbb{P}}(\{\text{classify correctly}\}|\mathcal{H}_{\boldsymbol{\theta}}, m) = \frac{1}{K}\sum_{k=1}^{K} c_{mk}, \tag{4}$$

and

$$\hat{\mathbb{P}}(\{\text{select } V\}|\mathcal{H}_{\boldsymbol{\theta}}, m) = \frac{1}{K}\sum_{k=1}^{K}\mathbb{I}(V \subseteq s_{mk}). \tag{5}$$

[Figure 1 about here.]

Instead of estimating the probability of correct classification as in Equation (4), we may wish to restrict our attention to those subsamples for which a particular subset V of the covariates were among the selections. This allows us to quantify the predictive performance associated with a particular set of covariates, rather than averaging the predictive performance over all covariate selections. We therefore calculate the mean correct classification rate over the subsamples D_k for which $V \subseteq s_{mk}$, and identify this as an estimate of the conditional probability that our classifier classifies correctly given that it selects V,

$$\hat{\mathbb{P}}(\{\text{classify correctly}\}|\{\text{select } V\}, \mathcal{H}_{\boldsymbol{\theta}}, m) = \frac{1}{\sum_{k=1}^{K}\mathbb{I}(V \subseteq s_{mk})}\sum_{k=1}^{K} c_{mk}\mathbb{I}(V \subseteq s_{mk}). \tag{6}$$

By multiplying together Equations (5) and (6), we obtain an estimate of the joint probability of our classifier both selecting V and classifying correctly,

$$\hat{\mathbb{P}}(\{\text{select } V \textit{ and } \text{classify correctly}\}|\mathcal{H}_{\boldsymbol{\theta}}, m) = \frac{1}{K}\sum_{k=1}^{K} c_{mk}\mathbb{I}(V \subseteq s_{mk}). \tag{7}$$

Equation (7) provides a simple probabilistic score that combines assessments of predictive performance and stability.

2.3.1 Covariate selection strategies

Adopting the procedure described in Figure 1 provides us with a collection, $\{s_{mk}, c_{mk}\}_{k=1}^{K}$ of K covariate sets and corresponding correct classification rates. In general, the s_{mk}

will not all be the same, so we must apply some strategy in order to decide which to return as our final set of putative biomarkers. We could, for example, return the set that is most frequently selected; i.e. choose the set V whose probability of selection (Equation (5)) is maximal. In Table 1, we present a number of probabilistic and heuristic strategies (S_1 – S_7) that exploit Equations (5) – (7) in order to optimise prediction performance, stability, or combinations of the two. All strategies are defined for a given model $\mathcal{H}_{\boldsymbol{\theta}}$ and set size m.

[Table 1 about here.]

Strategies S_1 – S_4 of Table 1 are *joint* strategies, which consider the joint selection and correct classification probabilities associated with *sets* of covariates. The differences between these strategies are illustrated in Figure 2. In contrast, S_5 and S_6 make use of the *marginal* selection and correct classification probabilities associated with *individual* covariates. S_7 is of a slightly different type, discussed further in Section 2.3.2.

[Figure 2 about here.]

2.3.2 Choosing between different m and $\mathcal{H}_{\boldsymbol{\theta}}$

Each of the strategies in Table 1 returns a final selected set and an associated score (for each pair $\mathcal{H}_{\boldsymbol{\theta}}, m$). If we have a range of predictive models and values for m, then we can consider all of them and return as our final selected set the one that gives the highest score (over all m and $\mathcal{H}_{\boldsymbol{\theta}}$). Adopting this approach, strategy S_7 can be viewed as finding the optimal pair $(\mathcal{H}^*_{\boldsymbol{\theta}}, m^*)$ for which the estimated probability of correct classification (Equation (4)) is largest, and then returning the most predictive set of size m^* selected by $\mathcal{H}^*_{\boldsymbol{\theta}}$. This is analogous to the common practice of using predictive performance to determine an appropriate level of regularisation.

2.4 Implementation

We focus on selection procedures that use logistic regression models with elastic net likelihood penalties (Zou and Hastie, 2005). The standard logistic regression model for the binary classification problem is as follows ,

$$P(y = 1|v_1 = x_1, \ldots, v_p = x_p) = g(\beta_0 + \beta^T \mathbf{x}), \tag{8}$$

where $\beta_0 \in \mathbb{R}$, $\beta = [\beta_1, \ldots, \beta_p]^\top \in \mathbb{R}^p$, $\mathbf{x} = [x_1, \ldots, x_p]^\top \in \mathbb{R}^p$, and g is the logistic function. Estimates for the coefficients $\beta_0, \beta_1, \ldots, \beta_p$ can be found by maximisation of the (log) likelihood function.

The elastic net introduces a penalty term $\lambda Q_\alpha(\beta)$ comprising a mixture of ℓ_1 and ℓ_2 penalties, so that the estimates for the coefficients are given by,

$$\begin{aligned} \widehat{\beta}_0^{(EN)}, \widehat{\beta}^{(EN)} = \underset{\beta_0,\beta}{\operatorname{argmax}} \Bigg[\frac{1}{N} \sum_{i=1}^{N} \Big\{ y_i \log(f(\beta_0 + \beta^T \mathbf{x}_i)) \\ + (1 - y_i) \log(1 - f(\beta_0 + \beta^T \mathbf{x}_i)) \Big\} - \lambda Q_\alpha(\beta) \Bigg], \end{aligned} \tag{9}$$

where

$$Q_\alpha(\beta) = \sum_{j=1}^{p} \left[\frac{1}{2}(1 - \alpha)\beta_j^2 + \alpha|\beta_j| \right]. \tag{10}$$

The estimated coefficients now depend upon the values taken by the parameters α and λ. When $\alpha = 1$, we recover the lasso (ℓ_1) penalty, and when $\alpha = 0$ we recover the ridge (ℓ_2) penalty. As α is decreased from 1 toward 0, the elastic net becomes increasingly tolerant of the selection of groups of correlated covariates. In the following, we consider a grid of α values ($\alpha = 0.1, 0.2, \ldots, 1$), and consider the order in which covariates are selected (acquire a non-zero β coefficient) as λ is decreased from λ_{crit} (the smallest value of λ such that $\widehat{\beta}^{(EN)} = [0, 0, \ldots, 0]^\top$) toward 0. Each different value

of α defines a different classification/selection procedure, $\mathcal{H}_\alpha^{(1)}, \ldots, \mathcal{H}_\alpha^{(10)}$, where $\mathcal{H}_\alpha^{(j)}$ corresponds to $\alpha = j/10$. Throughout, we use the `glmnet` package in `R` (Friedman *et al.*, 2010) to fit our models.

Although we use the elastic net penalty to select covariates, we use an unpenalised logistic regression model when making predictions. This two-step procedure of using the elastic net for variable selection and then obtaining unpenalised estimates of the coefficients in the predictive model is similar to the LARS-OLS hybrid (Efron *et al.*, 2004) or the relaxed lasso (Meinshausen, 2007).

3 Examples

3.1 Simulation example

Following a similar illustration from Meinshausen and Bühlmann (2010), we consider an example in which we have $p = 500$ predictors $v_1, \ldots, v_{500}$ and $n = 200$ observations. The predictors $v_1, \ldots, v_{500}$ are jointly distributed according to a multivariate normal whose mean μ is the zero vector and whose covariance matrix Σ is the identity, except that the elements $\Sigma_{1,2} = \Sigma_{3,4} = \Sigma_{3,5} = \Sigma_{4,5}$ and their symmetric counterparts are equal to 0.9. Thus, there are two strongly correlated sets, $C_1 = \{v_1, v_2\}$ and $C_2 = \{v_3, v_4, v_5\}$, but otherwise the predictors are uncorrelated. Observed class labels y are either 0 or 1, according to the following logistic regression model:

$$P(y = 1 | v_1, \ldots, v_{500}) = \frac{1}{1 + \exp\left(-\sum_{i=1}^{5} v_i\right)}. \tag{11}$$

Due to correlations among the covariates, it is also useful to consider the following

approximation:

$$P(y = 1|v_1, \ldots, v_{500}) \approx \frac{1}{1 + \exp\left(-(2v_{i_1} + 3v_{i_2})\right)}, \tag{12}$$

where $v_{i_1} \in C_1$ and $v_{i_2} \in C_2$.

Since $v_1, \ldots, v_5$ are the only covariates that appear in the generative model given in Equation (11), we refer to these as *relevant* covariates, and to the remainder as *noise* covariates.

We simulate 1,000 datasets — each comprising 200 observations — by first sampling from a multivariate normal in order to obtain realisations of the covariates $v_1, \ldots, v_{500}$, and then generating values for the response y according to Equation (11). We consider a range $m = 1, \ldots, 20$ and use $K = 100$ subsamples.

3.2 HTLV1 biomarker discovery

Human T-cell lymphotropic virus type 1 (HTLV1) is a widespread human virus associated with a number of diseases (Bangham, 2000a), including the inflammatory condition HTLV1-associated myelopathy/tropical spastic paraparesis (HAM/TSP). However, the vast majority ($\sim$95%) of individuals infected with HTLV1 remain lifelong asymptomatic carriers (ACs) of the disease (Bangham, 2000b). We seek to identify protein peak biomarkers from SELDI-TOF mass spectral data which allow us to discriminate between ACs and individuals with HAM/TSP.

We have blood plasma samples from a total of 68 HTLV1-seropositive individuals (34 HAM/TSP, 34 AC), processed as in Kirk *et al.* (2011). Here we analyse the combined dataset, D_C, comprising measurements from all 68 patients. We consider $m = 1, \ldots, 12$ and use $K = 250$ subsamples.

4 Results

4.1 Simulation example

We applied our selection strategies (Table 1) to each of our 1000 simulated datasets. For each simulation, each strategy returned a final set, V, containing the selected covariates. Each selected covariate must either be a noise or a relevant covariate. We can hence consider that $V = R \cup N$, where $R \subseteq V$ is a set containing only relevant covariates and $N \subseteq V$ is a set containing only noise covariates. The case $|R| = 5, |N| = 0$ is the ideal, as this corresponds to selecting all 5 relevant covariates, but none of the noise covariates. To assess the quality of our strategies, we therefore calculated for each the proportion of simulated datasets for which this ideal case was achieved. This information is provided in Table 2, along with a summary of the proportion of times that other combinations of the covariates were selected.

[Table 2 about here.]

4.1.1 Fewer false positives for strategies involving stability selection

The selected sets returned by Strategies S_2, S_3 and S_5 always contained at least one relevant covariate, and never any noise covariates. The lack of false positives for these three strategies contrasts with the strategy that uses predictive performance alone (S_1), which returned a selected set containing at least 1 noise covariate for 97.4% of the simulated datasets. Additionally enforcing a stability threshold upon the final selected set (S_4) decreases this percentage (to 15.3% when $\tau = 0.1$ and 4.3% when $\tau = 0.2$). One of the best performing strategies overall is S_6 (the marginal analogue of Strategy S_3), which selects all 5 relevant and 0 noise covariates for about two-thirds of the simulated datasets. In contrast to S_2, S_3 and S_5, however, S_6 does make some

false positive selections, with noise covariates being included among the final selections in 3.3% of cases. S_7 also performs well, selecting all 5 relevant and 0 noise covariates for 38.5% of the simulated datasets and making at least one false positive selection in only 8% of cases.

4.1.2 Smaller values of α yield more stable selections

As well as looking at the final selection made for each dataset (chosen over all classification models), we can also consider the results for each of the classification models $\mathcal{H}_\alpha^{(1)}, \ldots, \mathcal{H}_\alpha^{(10)}$ considered separately. We focus on Strategy S_3. For each simulated dataset and for each $\mathcal{H}_\alpha^{(j)}$, we use S_3 in order to select a final set. Associated with each of these selected sets is a score (the joint probability of selection and correct classification). In Figure 3, we illustrate the distributions of the scores obtained for $\mathcal{H}_\alpha^{(2)}, \mathcal{H}_\alpha^{(4)}, \mathcal{H}_\alpha^{(6)}, \mathcal{H}_\alpha^{(8)}$ and $\mathcal{H}_\alpha^{(10)}$ (i.e. for $\alpha = 0.2, 0.4, 0.6, 0.8$ and 1).

[Figure 3 about here.]

We can see from Figure 3 that smaller values of α tend to yield higher values of the score. Recall that there are two strongly correlated groups of relevant covariates (see Section 3.1), and smaller values of α will tend to allow all of the covariates in these two groups to be selected, while larger values of α will tend to result in a single representative from each of the two groups being selected. Although this does not have a significant impact in terms of predictive performance (since Equation (12) is a good approximation to Equation (11)), it does have a negative effect upon stability (since, for different subsamples of the data, different representatives can be selected).

4.2 HTLV1 biomarker discovery

We applied our selection strategies to the HTLV1 combined dataset, D_C. The selected covariates (protein peaks) are summarised in Table 3.

[Table 3 about here.]

All strategies included the 11.7 and 13.3kDa peaks among their selections. As might be expected from the results of the previous section, Strategy S_1 yields the largest selected set. The strategies that we found to provide the best performance in our simulation example (namely, S_2, S_3, S_6 and S_7) all selected the same 3 covariates. In Figure 4 we further illustrate the selections made using Strategy S_3 by showing how the score returned by this strategy varies as a function of m for each of the classification models $\mathcal{H}_\alpha^{(1)}, \ldots, \mathcal{H}_\alpha^{(10)}$ (i.e. for $\alpha = 0.1, 0.2, \ldots, 1$).

[Figure 4 about here.]

We can see from Figure 4 that the highest joint scores are again achieved for smaller values of α. The second peak in the joint score curve at $m = 6$ (observed for $\alpha = 0.1, 0.2$ and 0.3) is notable, and leads us to propose the proteins corresponding to the 13.3, 11.7 and 14.6kDa peaks as "high confidence" biomarkers, and the proteins corresponding to the 11.9, 17.3 and 17.5kDa peaks as potential biomarkers that might be worthy of further investigation. In Kirk *et al.* (2011) the 11.7 and 13.3kDa peaks were identified as β2-microglobulin and Calgranulin B, and the 17.3kDa peak as apolipoprotein A-II.

5 Discussion

We have considered a number of strategies for covariate selection that employ assessments of stability, predictive performance, and combinations of the two. We have conducted empirical assessments of these strategies using both simulated and real datasets.

Our work indicates that including assessments of stability can help to reduce the number of false positive selections, although this might come at the cost of only making a conservative number of (high confidence) selections. In the context of biomarker discovery, where follow-up work to identify and validate putative biomarkers is likely to be expensive and time-consuming, assessments of stability would seem to provide a useful way in which to focus future study. However, for large-scale hypothesis generation, selection strategies that employ stability assessments might be too conservative. Our simulation results (Section 4.1) suggest that combining assessments of stability and predictive performance can yield selection strategies that have lower false positive rates than strategies based on prediction alone, and lower false negative rates than pure stability selection strategies. We also found that classification/selection models that do not select complete sets of correlated predictive covariates run the risk of appearing to make unstable selections (Section 4.1.2). This will have a detrimental effect on stability selection approaches, further increasing the number of false negatives. It would therefore seem that if we are concerned with the stability with which selections are made (which should always be the case if our main aim is covariate selection/biomarker discovery), then it might be counter-productive just to search for the sparsest classification model that yields the maximal predictive performance. In particular, in order to improve the stability of selections, it would seem sensible to favour mixtures of ℓ_1 and ℓ_2 likelihood penalties (i.e. the elastic net) over lasso (ℓ_1 only) penalties.

Author Disclosure Statement

No competing financial interests exist.

References

Abeel, T., Helleputte, T., de Peer, Y. V., Dupont, P., and Saeys, Y., 2010. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics* 26, 392–8.

Ahmed, I., Hartikainen, A.-L., Järvelin, M.-R., and Richardson, S., 2011. False Discovery Rate Estimation for Stability Selection: Application to Genome-Wide Association Studies. *Statistical applications in genetics and molecular biology* 10.

Alexander, D. H. and Lange, K., 2011. Stability selection for genome-wide association. *Genetic epidemiology* 35, 722–728.

Bangham, C. R. M., 2000a. HTLV-1 infections. *J Clin Pathol* 53, 581–6.

Bangham, C. R. M., 2000b. The immune response to HTLV-I. *Curr Opin Immunol* 12, 397–402.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R., 2004. Least angle regression. *Ann Stat* 32, 407–451.

Ein-Dor, L., Kela, I., Getz, G., Givol, D., and Domany, E., 2005. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* 21, 171–8.

Ein-Dor, L., Zuk, O., and Domany, E., 2006. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci USA* 103, 5923–8.

Friedman, J., Hastie, T., Hoefling, H., and Tibshirani, R., 2007. Pathwise coordinate optimization. *Ann Appl Stat* 1, 302–332.

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

He, Z. and Yu, W., 2010. Stable feature selection for biomarker discovery. *Computational biology and chemistry* 34, 215–225.

Jurman, G., Merler, S., Barla, A., Paoli, S., Galea, A., and Furlanello, C., 2008. Algebraic stability indicators for ranked lists in molecular profiling. *Bioinformatics* 24, 258–64.

Kalousis, A., Prados, J., and Hilario, M., 2007. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and information systems* .

Kirk, P. D. W., Lewin, A. M., and Stumpf, M. P. H., 2010. Discussion of "Stability Selection". *J Roy Stat Soc B* 72.

Kirk, P. D. W., Witkover, A., Courtney, A., Lewin, A. M., Wait, R., Stumpf, M. P. H., Richardson, S., Taylor, G. P., and Bangham, C. R. M., 2011. Plasma proteome analysis in HTLV-1-associated myelopathy/tropical spastic paraparesis. *Retrovirology* 8, 81.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence* .

Kuncheva, L., 2007. A stability index for feature selection. *Proceedings of the 25th International Multi-Conference on Artificial Intelligence and Applications* 390–395.

Meinshausen, N., 2007. Relaxed lasso. *Comput Stat Data An* 52, 374–393.

Meinshausen, N. and Bühlmann, P., 2010. Stability Selection. *J Roy Stat Soc B* 72.

Smit, S., van Breemen, M. J., Hoefsloot, H. C. J., Smilde, A. K., Aerts, J. M. F. G., and de Koster, C. G., 2007. Assessing the statistical validity of proteomics based biomarkers. *Anal Chim Acta* 592, 210–7.

Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)* 36, 111–147.

Turney, P., 1995. Technical note: Bias and the quantification of stability. *Machine Learning* 20, 23–33.

Yu, L., Ding, C., and Loscalzo, S., 2008. Stable feature selection via dense feature groups. *Proceeding of the 14th ACM SIGKDD international conference on knowledge discovery and data mining (KDD'08)* 803–811.

Zou, H. and Hastie, T., 2005. Regularization and variable selection via the elastic net. *J Roy Stat Soc B* 67, 301–320.

Zucknick, M., Richardson, S., and Stronach, E. A., 2008. Comparing the characteristics of gene expression profiles derived by univariate and multivariate classification methods. *Statistical applications in genetics and molecular biology* 7, 7.

List of Figures

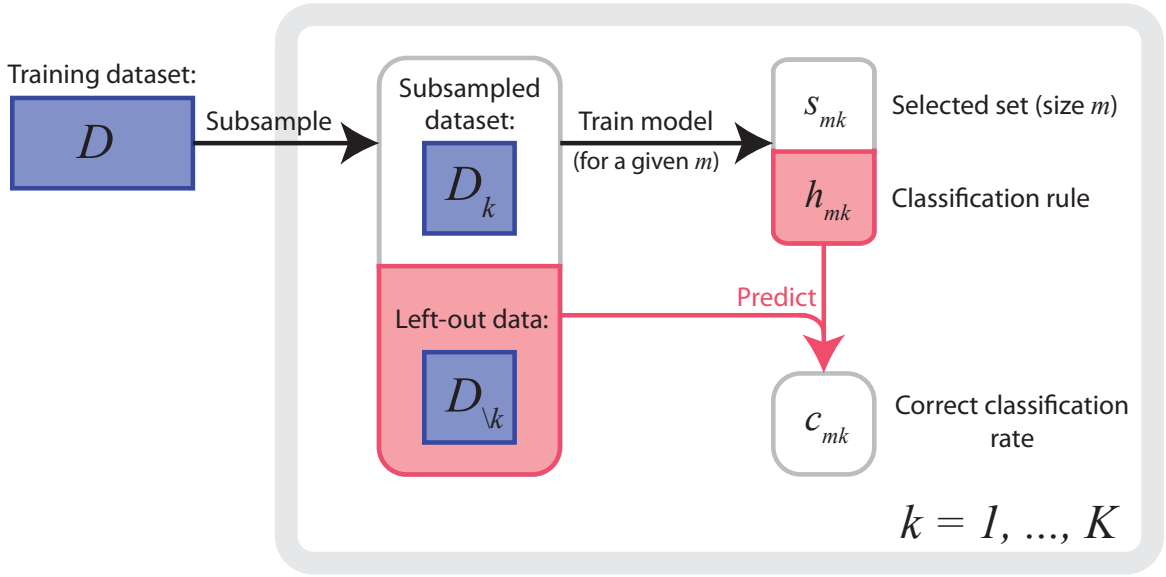

Figure 1: Summary of the notation and basic procedure used throughout this article. The training dataset, D, is repeatedly subsampled to obtain a collection of datasets, $\{D_k\}_{k=1}^K$, and left-out datasets, $\{D_{\setminus k}\}_{k=1}^K$. For $k = 1, \ldots, K$, a predictive model is trained on D_k and then used to predict the class labels of the observations in $D_{\setminus k}$, yielding a selected set of size m, s_{mk}, and a correct classification rate, c_{mk}.

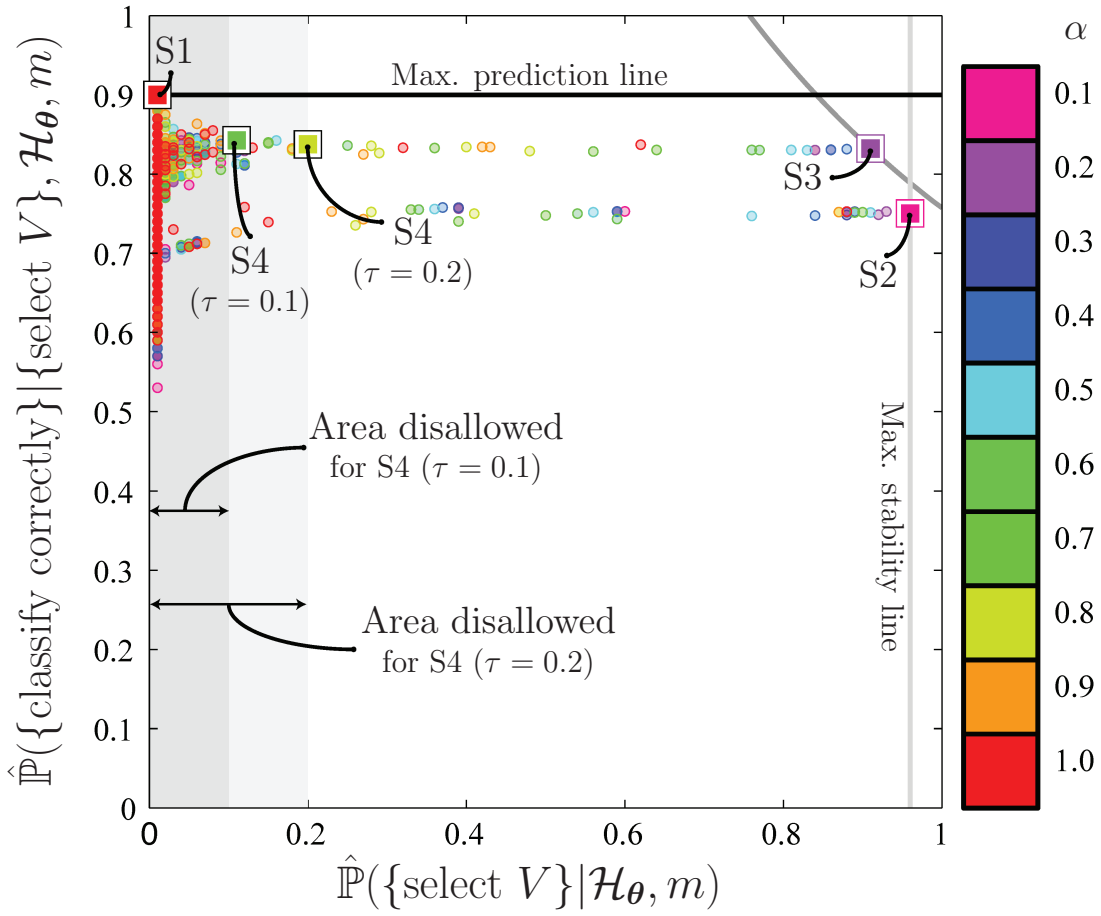

Figure 2: Illustration of the differences between the joint strategies (S_1 – S_4). The round markers correspond to different covariates sets (of various sizes) returned by 10 different models, $\mathcal{H}_\alpha^{(1)}, \ldots, \mathcal{H}_\alpha^{(10)}$, when applied to one of the simulated datasets of Section 3.1. Each model corresponds to a different value of α (see Section 2.4), hence the colours of the markers indicate the model that was used to select each covariate set. The larger, labelled markers correspond to the final sets of selections returned by strategies S_1 – S_4 (as indicated). S_1 returns the set, V, that maximises predictive performance, regardless of how stably it is selected; S_2 returns the most stably selected set, regardless of the predictive performance it offers; S_3 seeks a compromise between stability and predictive performance; and S_4 returns the most predictive covariate set, subject to a stability threshold, τ.

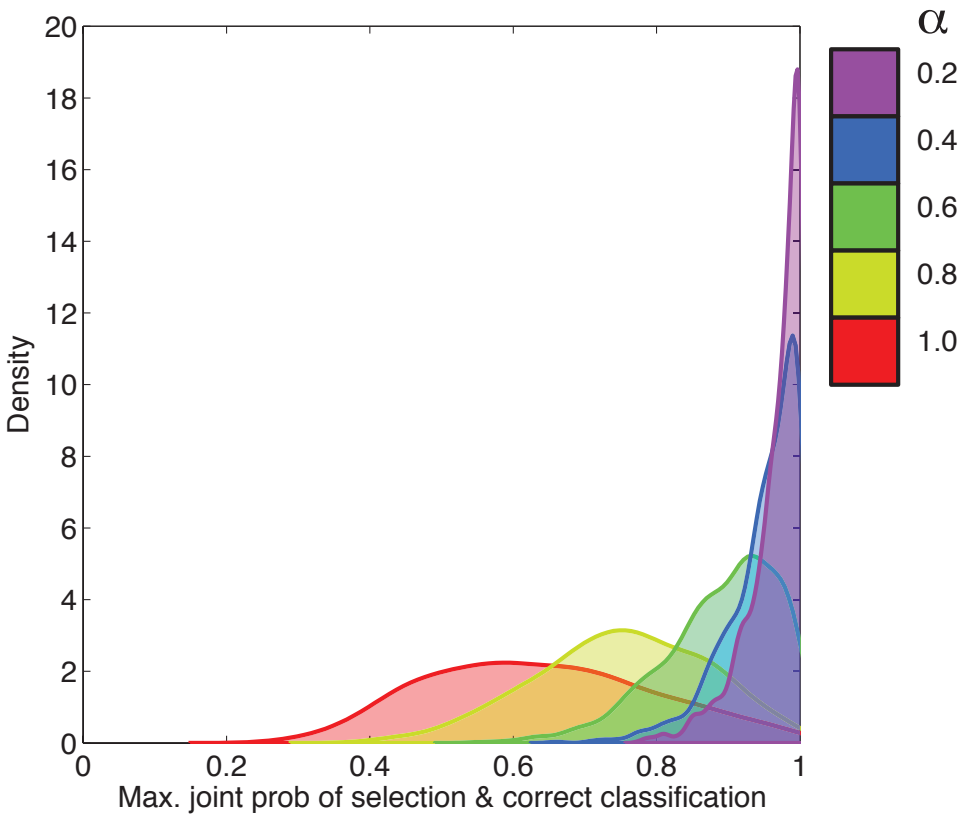

Figure 3: Distributions of the scores returned by S_3 which were obtained in the simulation example for 5 different values of α.

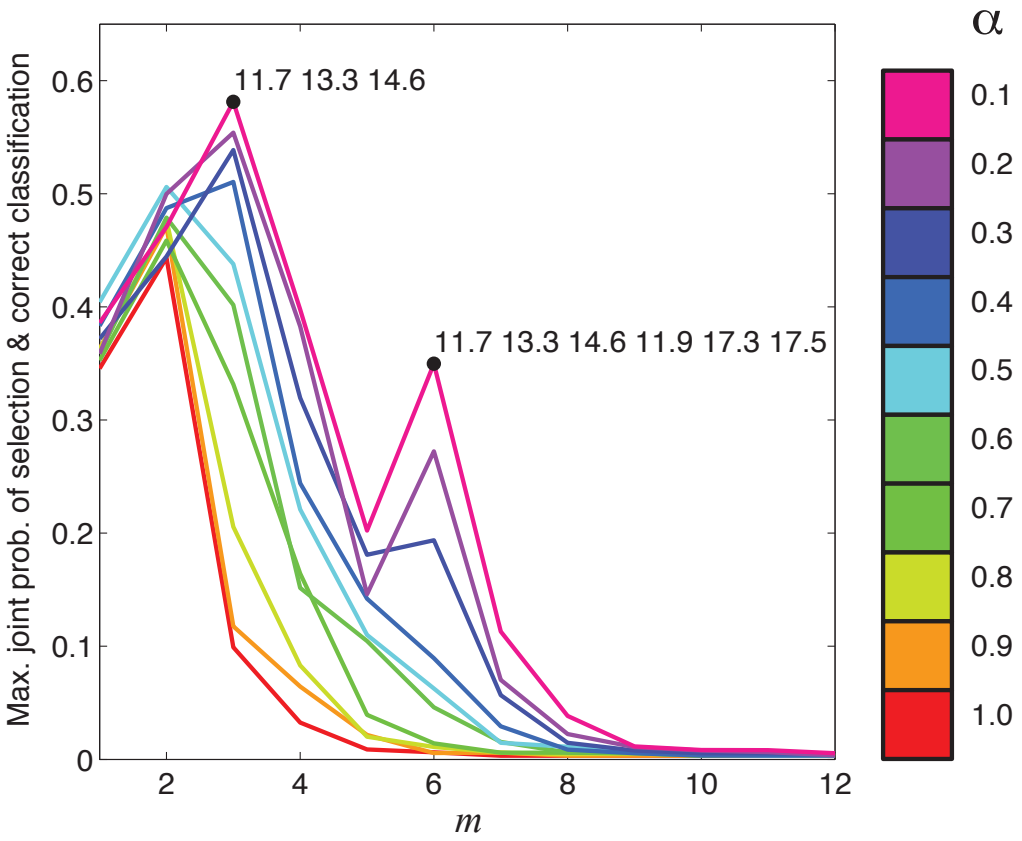

Figure 4: Score returned by S_3 considered as a function of m (when applied to the HTLV1 proteomics dataset).

List of Tables

SELECTION STRATEGIES		
Joint strategies		Select set V to maximise:
S_1	Prediction only	$P_1(V) = \hat{\mathbb{P}}(\{\text{classify correctly}\}\|\{\text{select } V\}, \mathcal{H}_{\boldsymbol{\theta}}, m)$.
S_2	Stability only	$P_2(V) = \hat{\mathbb{P}}(\{\text{select } V\}\|\mathcal{H}_{\boldsymbol{\theta}}, m)$.
S_3	Joint prob. of selection & correct classification	$P_3(V) = \hat{\mathbb{P}}(\{\text{select } V \text{ \& classify correctly}\}\|\mathcal{H}_{\boldsymbol{\theta}}, m)$. $= P_1(V)P_2(V)$
S_4	Prediction with stability threshold, τ	$P_4(V) = \hat{\mathbb{P}}(\{\text{classify correctly}\}\|\{\text{select } V\}, \mathcal{H}_{\boldsymbol{\theta}}, m)$, subject to the constraint $\hat{\mathbb{P}}(\{\text{select } V\}\|\mathcal{H}_{\boldsymbol{\theta}}, m) \geq \tau$.
Marginal strategies		Select set V to maximise:
S_5	Stability only (marginal case)	$P_5(V) = \frac{1}{m}\sum_{v_i \in V} \hat{\mathbb{P}}(\{\text{select } \{v_i\}\}\|\mathcal{H}_{\boldsymbol{\theta}}, m)$ $= \frac{1}{m}\sum_{v_i \in V} P_2(\{v_i\})$.
S_6	Joint prob. of selection & correct classification (marginal case)	$P_6(V) = \frac{1}{m}\sum_{v_i \in V} \hat{\mathbb{P}}(\{\text{select } v_i \text{ \& classify correctly}\}\|\mathcal{H}_{\boldsymbol{\theta}}, m)$ $= \frac{1}{m}\sum_{v_i \in V} P_3(\{v_i\})$.
Other		Select set V to maximise:
S_7	Average prediction	$P_1(V) = \hat{\mathbb{P}}(\{\text{classify correctly}\}\|\{\text{select } V\}, \mathcal{H}_{\boldsymbol{\theta}}, m)$. Also calculate $P_7 = \hat{\mathbb{P}}(\{\text{classify correctly}\}\|\mathcal{H}_{\boldsymbol{\theta}}, m)$.

Table 1: Selection strategies considered in this article. In each case, we assume that we have a predictive model, $\mathcal{H}_{\boldsymbol{\theta}}$, and that we specify the number, m, of covariates that we wish to select. For $j = 1, \ldots, 6$, strategy S_j returns selected set, V^*, together with maximised score $P_j(V^*)$. S_7 returns selected set V^* together with the score P_7.

	Strategy:	S_1	S_2	S_3	S_4 ($\tau = 0.1$)	S_4 ($\tau = 0.2$)	S_5	S_6	S_7
$\lvert R\rvert$	$\lvert N\rvert$	*Percentage of selections*							
5	0	0	36.3	50.5	5.2	10.6	42.5	66.9	38.5
4	0	0.6	2.4	8.5	32.4	40.3	3.4	16	50.2
3	0	1.8	57.4	39.1	39.6	36.9	52.9	13.6	2.9
2	0	0.2	1.4	0.5	7.5	7.8	0.6	0.1	0.4
1	0	0	2.5	1.4	0	0.1	0.6	0.1	0
5	1	10.5	0	0	11.6	4.1	0	0	0.5
4	1	9.2	0	0	2.8	0.2	0	1.3	4.8
3	1	2.8	0	0	0.6	0	0	0.4	2
5	2	19.7	0	0	0.2	0	0	0	0
4	2	5.1	0	0	0	0	0	0	0.1
3	2	1.8	0	0	0	0	0	0.5	0.2
5	3	12.5	0	0	0	0	0	0	0
4	3	3.5	0	0	0	0	0	0	0
5	4	7	0	0	0	0	0	0	0
4	4	1.9	0	0	0	0	0	0	0
5	5	4.7	0	0	0	0	0	0	0
5	6	2.6	0	0	0	0	0	0	0
4	6	1.1	0	0	0	0	0	0	0
5	7	1.5	0	0	0	0	0	0	0
5	8	1.8	0	0	0	0	0	0	0
5	9	1.8	0	0	0	0	0	0	0
5	10	1	0	0	0	0	0	0	0

Table 2: Summary of the final selections made using the strategies described in Table 1. The first two columns summarise the final selections in terms of the number of relevant covariates, $|R|$, and the number of noise covariates, $|N|$, that appear in the final selected set. The entries in the table indicate the percentage of simulated datasets for which each of the combinations of relevant and noise covariates was obtained. Any rows for which the percentage is $< 1\%$ for all strategies are omitted (hence columns need not sum to 100%).

Covariate selections	Strategies
11.7 13.3	S_4 ($\tau = 0.2$)
11.7 13.3 17.5	S_4 ($\tau = 0.1$)
11.7 13.3 14.6	S_2, S_3, S_5, S_6, S_7
10.8 11.7 11.9 13.3 14.6 25.1	S_1

Table 3: Covariates selected by strategies S_1–S_7. Covariates correspond to protein peaks in the mass-spectrum, and are labelled according to the m/z value at which the peak was located (units: kDa).